Федеральное государственное автономное
образовательное учреждение высшего образования
«Московский физико-технический институт
(национальный исследовательский университет)»

Физтех-школа прикладной математики и информатики

ЛЕКЦИИ И СЕМИНАРЫ ПО МЕТОДАМ ОПТИМИЗАЦИИ

Авторы:
А. Н. Безносиков, Е. Д. Бородич, Д. М. Двинских, Г. В. Кормаков, Н. М. Корнилов, И. А. Курузов, А. В. Лобанов, Д. С. Метелев, А. А. Моложавенко, С. А. Чежегов, Ю. И. Шароватова, Н. Е. Юдин, Д. В. Ярмошик, А. В. Андреев, А. И. Богданов, Д. А. Былинкин, Е. В. Рябинин, С. С. Ткаченко, А. А. Шестаков, Р. Е. Воронов

МОСКВА
ФПМИ МФТИ
2025

Учебное пособие содержит конспект лекций и материалы семинарских занятий по курсу «Методы оптимизации». В таком виде курс был прочитан осенью 2025–2026 учебного года студентам 3-го курса физтех-школы прикладной математики и информатики МФТИ. Важно отметить, что материалы курса опираются на различные курсы методов оптимизации, прочитанные в России и за рубежом. А именно, авторы данного пособия хотели бы поблагодарить А. Бен-Тала, А. Г. Бирюкова, А. В. Гасникова, В. Г. Жадана, А. М. Катруцу, Д. А. Кропотова, А. С. Немировского, Ю. Е. Нестерова, Б. Т. Поляка, Ф. С. Стонякина, М. Такача, Р. Хильдебранда, М. Шмидта, М. Ягги за неоценимый вклад в преподавание и популяризацию численных методов оптимизации во всем мире.

РАБОТА НАД ПОСОБИЕМ ВЕДЕТСЯ В ДАННЫЙ МОМЕНТ! АВТОРЫ ПРИЗНАТЕЛЬНЫ ЗА ЛЮБЫЕ КОММЕНТАРИИ ПО ЕГО УЛУЧШЕНИЮ!

Содержание

Обозначения и классические факты

Числовые множества

$\mathbb{N}$	натуральные числа
$\mathbb{Z}$	целые числа
$\mathbb{R}$	вещественные числа
$\mathbb{R}_+$	неотрицательные вещественные числа
$\mathbb{R}_{++}$	положительные вещественные числа

Кванторы и логические символы

$\forall$	квантор всеобщности
$\exists$	квантор существования
$\exists!$	квантор существования и единственности

Векторные обозначения

$x \succ y$	покомпонентное сравнение векторов	$x_i > y_i$ для всех i
$\langle x, y \rangle$	скалярное произведение в $\mathbb{R}^d$	$\langle x, y \rangle = \sum\limits_{i=1}^{d} x_i y_i$
$\|x\|_2$	евклидова норма	$\|x\|_2 = \sqrt{\langle x, x \rangle}$
$\|x\|_p$	p-норма для $1 \leq p < \infty$	$\|x\|_p = \left(\sum\limits_{i=1}^{d} \|x_i\|^p \right)^{\frac{1}{p}}$
$\|x\|_\infty$	∞-норма	$\|x\|_\infty = \max\limits_{i=1,d} \|x_i\|$
$\nabla f(x)$	градиент функции f в точке x	$\left(\nabla f(x)\right)_i = \dfrac{\partial f}{\partial x_i}$
$\nabla^2 f(x)$	гессиан функции f в точке x	$\left(\nabla^2 f(x)\right)_{ij} = \dfrac{\partial^2 f}{\partial x_i \partial x_j}$

Матричные обозначения и нормы

$\langle X, Y \rangle$	скалярное произведение матриц	$\langle X, Y \rangle = \operatorname{Tr}\left(X^\top Y\right) = \sum\limits_{i=1}^{d} \sum\limits_{j=1}^{n} X_{ij} Y_{ij}$
$A \odot B$	поэлементное умножение матриц	$\left(A \odot B\right)_{ij} = A_{ij} \cdot B_{ij}$
$\|A\|$	индуцированная матричная норма	$\|A\| = \sup\limits_{\|x\|=1} \|Ax\|$
$\|A\|_\infty$	бесконечная норма	$\|A\|_\infty = \max\limits_{i=1,d} \sum\limits_{j=1}^{n} \|a_{ij}\|$
$\|A\|_1$	первая норма	$\|A\|_1 = \max\limits_{j=1,n} \sum\limits_{i=1}^{d} \|a_{ij}\|$
$\|A\|_2$	вторая (евклидова) норма матрицы	$\|A\|_2 = \sqrt{\lambda_{\max}(A^\top A)}$
$\|A\|_F$	норма Фробениуса	$\|A\|_F = \sqrt{\sum\limits_{i=1}^{d} \sum\limits_{j=1}^{n} A_{ij}^2} = \sqrt{\operatorname{Tr}\left(A^\top A\right)}$

симметричные матрицы и сравнения

$A \succ B$	$A - B$ положительно определена
$A \in \mathbb{S}^d$	$A = A^\top$ (симметричная матрица)
$A \in \mathbb{S}^d_+$	A симметрична и $A \succeq 0$
$A \in \mathbb{S}^d_{++}$	A симметрична и $A \succ 0$

Полезные неравенства

Неравенство Йенсена

$$f\left(\sum_{i=1}^{n} \alpha_i x_i\right) \leq \sum_{i=1}^{n} \alpha_i f(x_i), \tag{0.1}$$

где f — выпуклая, $\alpha_i \geq 0$, $\sum\limits_{i=1}^{n} \alpha_i = 1$.

Неравенство Гельдера

$$\left|\langle x, y\rangle\right| \leq \|x\|_p \|y\|_q, \tag{0.2}$$

где $1 \leq p, q \leq \infty$, $\frac{1}{p} + \frac{1}{q} = 1$.

Неравенство Коши–Буняковского–Шварца

$$\left|\langle x, y\rangle\right| \leq \|x\|_2 \|y\|_2. \tag{0.3}$$

Неравенства о средних

$$\frac{n}{\sum\limits_{i=1}^{n} \frac{1}{x_i}} \leq \left(\prod_{i=1}^{n} x_i\right)^{\frac{1}{n}} \leq \frac{1}{n}\sum_{i=1}^{n} x_i \leq \sqrt{\frac{1}{n}\sum_{i=1}^{n} x_i^2},$$

для $x_i > 0$.

Неравенство между степенной функцией и экспонентой

$$(1-\alpha)^k \leq \exp(-\alpha k), \quad \alpha \in [0, 1]. \tag{0.4}$$

Асимптотические обозначения

$f(n) \in \mathcal{O}(g(n))$	$\exists\, C > 0$: $f(n) \leq C \cdot g(n)$ при всех $n \in \mathbb{N}$
$f(n) \in \Omega(g(n))$	$\exists\, C > 0$: $f(n) \geq C \cdot g(n)$ при всех $n \in \mathbb{N}$
$f(n) \in \Theta(g(n))$	$\exists\, C_1, C_2 > 0$: $C_1 g(n) \leq f(n) \leq C_2 g(n)$ при всех $n \in \mathbb{N}$
$f(n) \in \tilde{\mathcal{O}}(g(n))$	$\exists$ полином p: $f(n) \in \mathcal{O}(g(n) \cdot p(\log n))$ при всех $n \in \mathbb{N}$

Обозначения для распределений

$U(\overline{1,n})$	равномерное распределение на множестве $\overline{1,n}$

C1 Пререквизиты из линейной алгебры

Линейная алгебра играет фундаментальную роль во многих областях математики. Её методы и концепции лежат в основе таких дисциплин, как численные методы, теория управления, анализ данных и, в том числе, теория оптимизации — предмет, которому посвящено пособие. В рамках этого параграфа обсуждаются пререквизиты из линейной алгебры, необходимые для успешного усвоения материала по оптимизации. В частности, обсуждаются векторные и матричные нормы, а также разложения матриц.

C1.1 Векторные нормы

Определение C1.1. Рассмотрим функцию $\|\cdot\| : \mathbb{R}^d \to \mathbb{R}$. Она называется *векторной нормой*, если удовлетворяет следующим условиям:

- $\|x\| \geq 0$, $\|x\| = 0 \iff x = 0$;
- $\|\lambda x\| = |\lambda|\|x\|$, $\lambda \in \mathbb{R}$;
- $\|x + y\| \leq \|x\| + \|y\|$.

Пример C1.1. Функция $\|\cdot\| : \mathbb{R}^d \to \mathbb{R}$, определённая как

$$\|x\|_p = \sqrt[p]{\sum_{i=1}^{d} |x_i|^p}, \ p \geq 1,$$

является векторной нормой.

Доказательство. Первые два условия из Определения C1.1 не нуждаются в доказательстве. Покажем, что для произвольных $x, y \in \mathbb{R}^d$ выполнено третье условие. Рассмотрим $p \neq 1$, поскольку случай $p = 1$ тривиален. Нетрудно видеть, что

$$|x_i + y_i|^p \leq |x_i||x_i + y_i|^{p-1} + |y_i||x_i + y_i|^{p-1}.$$

Обозначив $q = \frac{p}{p-1}$, применим неравенство Гельдера (0.2) к каждому из слагаемых. Получим

$$\begin{aligned}\|x + y\|_p^p &= \sum_{i=1}^{d} |x_i + y_i|^p \leq \sum_{i=1}^{d} |x_i||x_i + y_i|^{p-1} + \sum_{i=1}^{d} |y_i||x_i + y_i|^{p-1} \\ &\leq \left(\sum_{i=1}^{d} |x_i|^p\right)^{\frac{1}{p}} \left(\sum_{i=1}^{d} |x_i + y_i|^p\right)^{\frac{1}{q}} + \left(\sum_{i=1}^{d} |y_i|^p\right)^{\frac{1}{p}} \left(\sum_{i=1}^{d} |x_i + y_i|^p\right)^{\frac{1}{q}}.\end{aligned}$$

Сократив обе части на $\left(\sum_{i=1}^{d} |x_i + y_i|^p\right)^{\frac{1}{q}}$, учитывая, что $1 - \frac{1}{q} = \frac{1}{p}$, имеем

$$\|x + y\|_p \leq \|x\|_p + \|y\|_p.$$

■

Замечание C1.1. Нормы, введённые в Примере C1.1, будем называть *гёльдеровыми нормами* или просто *p-нормами*.

В анализе некоторых алгоритмов иногда возникает так называемая ∞-норма.

Определение С1.2. Функцию $\|\cdot\|_\infty : \mathbb{R}^d \to \mathbb{R}$, определённую как

$$\|x\|_\infty = \max_{i=\overline{1,d}} |x_i|,$$

будем называть ∞-*нормой*.

Пример С1.2. ∞-норма является предельным случаем гёльдеровых норм при $p \to \infty$.

Доказательство. Во-первых, известно, что $|x_j| \le \max_{i=\overline{1,d}} |x_i|$. Это означает, что

$$\|x\|_p \le \left(d \max_{i=\overline{1,d}} |x_i|^p \right)^{\frac{1}{p}} = d^{\frac{1}{p}} \max_{i=\overline{1,d}} |x_i| \underset{p\to\infty}{\longrightarrow} \|x\|_\infty.$$

С другой стороны, имеем

$$\|x\|_p \ge \left(\max_{i=\overline{1,d}} |x_i|^p \right)^{\frac{1}{p}} = \max_{i=\overline{1,d}} |x_i| = \|x\|_\infty.$$

По теореме о трех последовательностях (см. Теорему 2 Параграфа 5 в [27]), имеем

$$\lim_{p\to\infty} \|x\|_p = \|x\|_\infty.$$

■

Теперь, когда мы ввели в рассмотрение множество примеров векторных норм, закономерно возникает вопрос о том, как они соотносятся между собой. Действительно, показав сходимость численного метода в некоторой норме, мы хотели бы иметь уверенность, что он сходится и в других нормах тоже. Из курса математического анализа известно утверждение.

Утверждение С1.1 (Теорема 2 Части 14 в [9])**.** Рассмотрим $\|\cdot\|_A, \|\cdot\|_B : \mathbb{R}^d \to \mathbb{R}$. Существуют такие $c_1, c_2 > 0$, что для любого вектора x из $\mathbb{R}^d$ выполняется соотношение:

$$c_1\|x\|_A \le \|x\|_B \le c_2\|x\|_A.$$

Таким образом, в конечномерном вещественном пространстве все нормы эквивалентны. Доказательство утверждения не конструктивно, однако в ряде задач мы хотели бы конкретизировать значения c_1 и c_2.

Пример С1.3. Для $\|\cdot\|_1$ и $\|\cdot\|_2$ выполнено соотношение:

$$\frac{1}{\sqrt{d}}\|x\|_1 \le \|x\|_2 \le \|x\|_1.$$

Доказательство. Начнём с верхней оценки. Запишем

$$\begin{aligned}\|x\|_2^2 &= |x_1|^2 + \ldots + |x_d|^2 \le |x_1|^2 + \ldots + |x_d|^2 + 2|x_1||x_2| + \ldots + 2|x_{d-1}||x_d| \\ &= (|x_1| + \ldots + |x_d|)^2 = \|x_1\|_1^2.\end{aligned}$$

Таким образом,

$$\|x\|_2 \le \|x\|_1.$$

Заметим, что равенство переходит в равенство на векторе $x = e_1$. Чтобы оценить $\|x\|_2$ снизу, воспользуемся неравенством Гельдера (0.2), положив один из векторов в скалярном произведении равным единице:

$$\|x\|_1^2 = \left(\sum_{i=1}^{d} 1|x_i|\right)^2 \le \left(\sum_{i=1}^{d} 1^2\right)\left(\sum_{i=1}^{d} |x_i|^2\right) = d\|x\|_2^2.$$

Заметим, что неравенство переходит в равенство на векторе $x = \mathbf{1}$, где $\mathbf{1}$ — вектор из всех единиц. Комбинируя результаты, получаем

$$\frac{1}{\sqrt{d}}\|x\|_1 \le \|x\|_2 \le \|x\|_1.$$

■

C1.2 Матричные нормы

В анализе численных методов помимо векторов также приходится работать с матрицами. Существует естественное обобщение определения нормы.

Определение C1.3. Рассмотрим функцию $\|\cdot\| : \mathbb{R}^{n\times d} \to \mathbb{R}$. Она называется *матричной нормой*, если удовлетворяет следующим условиям:

- $\|A\| \ge 0, \quad \|A\| = 0 \iff A = 0;$
- $\|\lambda A\| = |\lambda|\|A\|, \ \lambda \in \mathbb{R};$
- $\|A + B\| \le \|A\| + \|B\|.$

Замечание C1.2. Для матриц помимо сложения также определена операция умножения. Ключевым для анализа численных методов является свойство

$$\|AB\| \le \|A\|\|B\|.$$

Тем не менее, оно выполнено не для всех матричных норм. Рассмотрим норму:

$$\|A\| = \max_{\substack{i=\overline{1,n}\\ j=\overline{1,d}}} |a_{ij}|$$

и матрицы

$$A = \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}, \quad B = \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}.$$

Получим, что $\|AB\| = 2$, $\|A\| = 1$, $\|B\| = 1$, то есть свойство не выполняется.

Замечание C1.3. Если для $\|\cdot\| : \mathbb{R}^{n\times d} \to \mathbb{R}$ выполнено

$$\|AB\| \le \|A\|\|B\|,$$

то её называют *субмультипликативной* матричной нормой.

Может показаться, что мы вынуждены вручную проверять субмультипликативность, сталкиваясь с новой матричной нормой. Эту проблему можно решить, рассматривая более узкий класс норм, для которых это свойство выполняется автоматически. В дальнейшем окажется, что наиболее часто используемые матричные нормы принадлежат этому классу.

Определение C1.4. Рассмотрим $A \in \mathbb{R}^{n\times d}$, норму $\|\cdot\|_\alpha : \mathbb{R}^d \to \mathbb{R}$ и норму $\|\cdot\|_\beta : \mathbb{R}^n \to \mathbb{R}$. Матричной нормой, *подчинённой* $\|\cdot\|_\alpha$ *и* $\|\cdot\|_\beta$, называется функция $\|\cdot\|_{\alpha,\beta} : \mathbb{R}^{n\times d} \to \mathbb{R}$, заданная как

$$\|A\|_{\alpha,\beta} = \sup_{\|x\|_\alpha=1} \|Ax\|_\beta.$$

Замечание C1.4. Мы будем рассматривать только случай, когда $\alpha = \beta$, тогда определение можно переписать следующим образом:

$$\|A\| = \sup_{\|x\|=1} \|Ax\|.$$

Замечание C1.5. Определение C1.4 можно переформулировать, представив единичный x как некоторый вектор после нормировки на единичную сферу. Тогда имеем

$$\|A\| = \sup_{\|x\|=1} \|Ax\| = \sup_{x\neq 0} \frac{\|Ax\|}{\|x\|}.$$

Замечание C1.6. Из Замечания C1.5 следует неравенство

$$\frac{\|Ax\|}{\|x\|} \leq \|A\| \implies \|Ax\| \leq \|A\|\|x\|.$$

Это очень полезное свойство, которое будет неоднократно использовано в дальнейшем изложении.

Пример C1.4. Подчинённая матричная норма всегда субмультипликативна.

Доказательство. Рассмотрим матрицы A и B, имеющие правильные размерности. Записав определение подчинённой нормы, получим

$$\|AB\| = \sup_{\|x\|=1} \|ABx\| \leq \sup_{\|x\|=1} \|A\|\|Bx\| \leq \sup_{\|x\|=1} \|A\|\|B\|\|x\| = \|A\|\|B\|.$$

■

Использование подчинённой нормы также позволяет оценивать спектральный радиус матрицы.

Утверждение C1.2. Рассмотрим $A \in \mathbb{S}^d$, где $\mathbb{S}^d$ — множество симметричных матриц. Обозначим $\rho(A) = \max\limits_{i=\overline{1,d}} |\lambda_i(A)|$ — *спектральный радиус*, где $\lambda_i(A)$ — собственные числа матрицы A. Тогда

$$\rho(A) \leq \|A\|,$$

где $\|\cdot\|$ — подчинённая норма.

Доказательство. Запишем уравнение на собственные векторы матрицы A:

$$Ax = \lambda x \implies \|Ax\| = |\lambda|\|x\|.$$

В Замечании C1.6 было получено неравенство

$$\|Ax\| \leq \|A\|\|x\|.$$

Тогда можем утверждать, что

$$|\lambda| \leq \|A\|.$$

Отсюда очевидно следует, что

$$\rho(A) = \max_{i=\overline{1,d}} |\lambda_i(A)| \leq \|A\|.$$

Заметим, что на I_d достигается равенство. ■

Помимо собственных чисел часто оказывается удобным использовать сингулярные. Например, если работаем с прямоугольной матрицей.

Определение C1.5. *Сингулярным числом* матрицы $A \in \mathbb{R}^{n\times d}$ называется

$$\sigma_i(A) = \sqrt{\lambda_i(A^\top A)}.$$

На практике часто приходится работать с квадратными симметричными матрицами. В этом случае сингулярные числа имеют более удобный вид.

Утверждение C1.3. Пусть $A \in \mathbb{S}^d$. Тогда сингулярные числа имеют вид

$$\sigma_i(A) = |\lambda_i(A)|.$$

Доказательство. Собственные векторы вещественной симметричной матрицы A ортогональны (см. Теорему 1 Параграфа 3 Главы 6 в [26]). Рассмотрев их в качестве базиса, можно привести A к диагональному виду. Запишем её преобразование при переходе от стандартного базиса в базис собственных векторов:

$$A = V\Lambda V^\top,$$

где $V \in \mathbb{R}^{d\times d}$ — ортогональная матрица, составленная из собственных векторов A, $\Lambda = \operatorname{diag}(\lambda_1(A), \ldots, \lambda_d(A))$. Теперь запишем

$$A^\top A = A^2 = (V\Lambda V^\top)(V\Lambda V^\top) = V\Lambda^2 V^\top.$$

Таким образом, доказали, что $\lambda_i(A^\top A) = \lambda_i(A)^2$, откуда следует, что $\sigma_i(A) = |\lambda_i(A)|$. ■

Таким образом, подчинённая матричная норма определена достаточно удачно и имеет ряд хороших свойств. Более того, она даёт связь с векторными нормами. Оказывается, что Определение C1.4 позволяет давать аналитические выражения для подсчета матричных норм.

Пример C1.5. Рассмотрим $A \in \mathbb{R}^{n\times d}$. Для подчинённых норм $\|\cdot\|_\infty, \|\cdot\|_1, \|\cdot\|_2 : \mathbb{R}^{n\times d} \to \mathbb{R}$, имеем следующие выражения:

- $\|A\|_\infty = \max\limits_{i=\overline{1,n}} \sum\limits_{j=1}^{d} |a_{ij}|$ — максимум сумм по строкам,

- $\|A\|_1 = \max\limits_{j=\overline{1,d}} \sum\limits_{i=1}^{n} |a_{ij}|$ — максимум сумм по столбцам.

- $\|A\|_2 = \max\limits_{i=\overline{1,d}} \sqrt{\lambda_i(A^\top A)} = \max\limits_{i=\overline{1,d}} \sigma_i(A)$ — максимальное сингулярное число.

Доказательство. Будем доказывать в том же порядке, в котором утверждения приведены в формулировке примера.

- Распишем:

$$\|Ax\|_\infty = \max_{i=\overline{1,d}} \left| \sum_{j=1}^{d} a_{ij} x_j \right| \le \max_{i=\overline{1,d}} \sum_{j=1}^{d} |a_{ij} x_j| \le \max_{k=\overline{1,d}} |x_k| \max_{i=\overline{1,d}} \sum_{j=1}^{d} |a_{ij}|$$
$$= \|x\|_\infty \max_{i=\overline{1,d}} \sum_{j=1}^{d} |a_{ij}|.$$

  Пусть максимум правой части достигается при $i = i_0$. Рассмотрим вектор $x = \left(\frac{a_{i_0 1}}{|a_{i_0 1}|}, \dots, \frac{a_{i_0 d}}{|a_{i_0 d}|} \right)^\top$. Тогда, подставляя его в выражение, выше получаем равенство. Таким образом, существует вектор, на котором верхняя грань достигается.

- Будем расписывать почти как в предыдущем примере:

$$\|Ax\|_1 = \sum_{i=1}^{n} \left| \sum_{j=1}^{d} a_{ij} x_j \right| \le \sum_{j=1}^{d} |x_j| \sum_{i=1}^{n} |a_{ij}| \le \|x\|_1 \max_{j=\overline{1,d}} \sum_{i=1}^{d} |a_{ij}|.$$

  Пусть максимум правой части достигается при $j = j_0$. Рассмотрим вектор $x = e_{j_0}$. Тогда, подставляя его в выражение, выше получаем равенство. Таким образом, существует вектор, на котором верхняя грань достигается.

- Раскроем определение:

$$\|A\|_2 = \sup_{x \ne 0} \frac{\|Ax\|_2}{\|x\|_2} = \sup_{x \ne 0} \sqrt{\frac{\langle A^\top Ax, x \rangle}{\langle x, x \rangle}}.$$

  Разложим x по ортонормированному базису $\nu_1, \dots, \nu_d$ собственных векторов, отвечающих собственным числам $\lambda_1, \dots, \lambda_d$ матрицы $A^\top A$. Напомним, что собственные числа симметричной положительно полуопределённой матрицы неотрицательны. Тогда:

$$\|A\|_2 = \sup_{\|x\|_2 = 1} \sqrt{\frac{\sum_{j=1}^{d} \lambda_j c_j^2}{\sum_{j=1}^{d} c_j^2}} \le \max_{i=\overline{1,d}} \sqrt{\lambda_i} = \max_{i=\overline{1,d}} \sigma_i(A),$$

где c_i — коэффициенты разложения x по базису $\nu_1, \ldots, \nu_d$:

$$x = \sum_{i=1}^{d} c_i \nu_i.$$

Рассмотрим нормированный вектор, соответствующий максимальному собственному значению матрицы $A^\top A$. Тогда, подставляя его в выражение, выше получаем равенство. Таким образом, существует вектор, на котором верхняя грань достигается.

■

Пример С1.6. Матричные нормы $\|\cdot\|_\infty$, $\|\cdot\|_1$ и $\|\cdot\|_2$ связаны соотношением:

$$\|A\|_2^2 \leq \|A\|_1 \|A\|_\infty.$$

Доказательство. Поскольку при транспонировании строки и столбцы меняем местами, верно равенство $\|A\|_1 = \|A^\top\|_\infty$. Запишем определение второй нормы матрицы и для того, чтобы ограничить собственное значение матрицы воспользуемся Утверждением С1.2:

$$\|A\|_2^2 = \max_{i=\overline{1,d}} \lambda_i(A^\top A) \leq \|A^\top A\|_1 \leq \|A^\top\|_1 \|A\|_1 = \|A\|_1 \|A\|_\infty.$$

■

Пример С1.7. Матричная норма $\|A\|_2$ может быть определена иначе:

$$\|A\|_2 = \sup_{\|x\|_2=1, \|y\|_2=1} \left| y^\top A x \right|.$$

Доказательство. Ранее уже отмечалось, что для матричных норм верно неравенство $\|Ax\| \leq \|A\|\|x\|$. Пользуясь этим свойством и неравенством Коши-Буняковского-Шварца (0.3) для нормы $p = 2$, запишем:

$$\left| y^\top A x \right| = |\langle y, Ax \rangle| \leq \|y\|_2 \|Ax\|_2 \leq \|y\|_2 \|A\|_2 \|x\|_2 = \|A\|_2.$$

Выберем единичный вектор x_*, на котором достигается $\|A\|_2$ и определим $y_* = \frac{Ax_*}{\|Ax_*\|_2}$. Тогда:

$$\left| y_*^\top A x_* \right| = \left| \frac{x_*^\top A^\top A x_*}{\|Ax_*\|_2} \right| = \frac{\|Ax_*\|_2^2}{\|Ax_*\|_2} = \|Ax_*\|_2 = \|A\|_2.$$

■

Тем не менее, не все нормы, используемые на практике являются подчинёнными. Чтобы перейти к их рассмотрению, требуется ввести понятие ранга матрицы.

Определение С1.6. *Столбцовым рангом* матрицы $A \in \mathbb{R}^{n \times d}$ называется число $\mathrm{rank}_c(A)$, такое, что в A существует линейно независимая система из $\mathrm{rank}_c(A)$ столбцов и нет линейно независимой системы из большего числа столбцов.

Замечание C1.7. Существует множество способов определения ранга матрицы. Мы считаем известным из курса линейной алгебры утверждение об их эквивалентности (см. Теорему 1 Параграфа 3 Главы 5 в [25]). В связи с этим, будем пользоваться обозначением $\operatorname{rank}(A) = \operatorname{rank}_c(A)$.

Пример C1.8. Пусть $A \in \mathbb{R}^{n\times d}$. Матричная норма

$$\|A\|_F = \sqrt{\sum_{i=1}^{\operatorname{rank}(A)} \sigma_i^2(A)}$$

не подчинена никакой векторной норме.

Доказательство. Если бы норма была подчинённой, то для единичной матрицы I_d выполнялось бы

$$\|I_d\|_F = \sup_{\|x\|=1} \|I_d x\| = \sup_{\|x\|=1} \|x\| = 1.$$

Однако

$$\|I_d\|_F = \sqrt{\sum_{i=1}^{d} 1} = \sqrt{d}.$$

Таким образом, предложенная норма не подчинена ни одной из векторных норм. ■

Определение C1.7. Пусть $A \in \mathbb{R}^{n\times d}$. Функцию $\|\cdot\|_F : \mathbb{R}^{n\times d} \to \mathbb{R}$, определённую как

$$\|A\|_F = \sqrt{\sum_{i=1}^{\operatorname{rank}(A)} \sigma_i^2(A)},$$

будем называть *фробениусовой нормой*.

Замечание C1.8. Обратим внимание, что пространство матриц является конечномерным вещественным. Это означает, что матричные нормы эквивалентны как и векторные.

Пример C1.9. Для $\|\cdot\|_2$ и $\|\cdot\|_F$ выполнено соотношение:

$$\|A\|_2 \le \|A\|_F \le \sqrt{d}\|A\|_2.$$

Доказательство. Без ограничения общности расположим сингулярные числа в порядке убывания и положим $\operatorname{rank}(A) = r$. Поскольку $\|A\|_2 = \sigma_1$, то по определению фробениусовой нормы имеем

$$\|A\|_2 \le \|A\|_F.$$

С другой стороны, $\|A\|_F \le \sqrt{r}\sigma_1 \le \sqrt{d}\sigma_1 = \sqrt{d}\|A\|_2$. Объединяя неравенства, запишем

$$\|A\|_2 \le \|A\|_F \le \sqrt{d}\|A\|_2.$$

■

Фробениусова норма может показаться сложной для вычисления по сравнению с рассмотренными ранее. В дальнейшем выяснится, что можно дать простое эквивалентное определение. Для этого нам потребуется ряд вспомогательных утверждений.

Теорема C1.1. Матрица $A \in \mathbb{R}^{n\times d}$ представима в виде SVD-разложения:

$$A = U\Sigma V^\top,$$

где $U \in \mathbb{R}^{n\times n}, V \in \mathbb{R}^{d\times d}$ — ортогональные, $\Sigma \in \mathbb{R}^{n\times d}$ — диагональная матрица, составленная из сингулярных чисел A, расположенных в порядке убывания.

Доказательство. $A^\top A$ неотрицательно определена и симметрична, поэтому её собственные числа неотрицательны и соответствующие им собственные векторы ортогональны. Отсюда из Утверждения C1.3 следует существование ортогональной матрицы V, такой что

$$V^\top A^\top AV = \operatorname{diag}\left(\sigma_1^2, \ldots, \sigma_d^2\right).$$

Без ограничения общности, будем считать $\sigma_1^2 \geq \ldots \geq \sigma_d^2 \geq 0$. Поскольку $\operatorname{rank}(A) = r$, имеем

$$\sigma_i = 0, \ \forall i > r.$$

Обозначим $V_r = (v_1, \ldots, v_r)$, $\Sigma_r = \operatorname{diag}(\sigma_1, \ldots, \sigma_r)$. Тогда имеем

$$V_r^\top A^\top AV_r = \Sigma_r^2.$$

Умножив равенство слева и справа на Σ_r^{-1}, получим

$$\left(\Sigma_r^{-1}V_r^\top A^\top\right)\left(AV_r\Sigma_r^{-1}\right) = I.$$

Обозначим $U_r = AV_r\Sigma_r^{-1}$. Из написанного выше следует $U_r^\top U_r = I$, то есть U_r матрица с ортонормированными столбцами. Поскольку систему линейно независимых векторов можно дополнить до базиса, присоединим к U_r произвольные ортонормированные столбцы и получим новую матрицу U. Тогда

$$A = U\Sigma V^\top.$$

■

Теперь мы готовы дать эквивалентное определение фробениусовой нормы.

Утверждение C1.4. Для матрицы $A \in \mathbb{R}^{n\times d}$ фробениусова норма может быть эквивалентно определена как

$$\|A\|_F^2 = \sum_{i=1}^{n}\sum_{j=1}^{d}|a_{ij}|^2.$$

Доказательство. Для простоты изложения введем обозначение:

$$\|A\|_S^2 = \sum_{i=1}^{n}\sum_{j=1}^{d}|a_{ij}|^2.$$

Будем работать с квадратными матрицами $A \in \mathbb{R}^{d\times d}$. Во-первых, заметим, что для ортогональной матрицы Q выполнено

$$\|QA\|_S = \|A\|_S.$$

Действительно, рассмотрим A как совокупность вектор-столбцов: $A = (a_1, \ldots, a_d)$. Тогда имеем

$$\|QA\|_S^2 = \|(Qa_1, \ldots, Qa_d)\|_S^2 = \sum_{i=1}^{d}\sum_{j=1}^{d} |(Qa_i)_j|^2 = \sum_{i=1}^{d} \|Qa_i\|_2^2 = \sum_{i=1}^{d} \|a_i\|_2^2 = \|A\|_S^2.$$

Аналогично проверяется инвариантность относительно умножения на Q справа. Воспользуемся SVD-разложением С1.1 и запишем $\|A\|_S$, применив эту идею. Получим

$$\|A\|_S^2 = \|U\Sigma V^\top\|_S^2 = \|\Sigma\|_S^2 = \sum_{i=1}^{r} \sigma_i^2 = \|A\|_F^2.$$

■

Теперь мы получили простой способ подсчёта фробениусовой нормы. Поскольку она не подчинена ни одной из векторных норм, выполнение субмультипликативного свойства надо проверять вручную.

Пример С1.10. Для матриц $A \in \mathbb{R}^{d\times n}$ и $B \in \mathbb{R}^{n\times k}$ фробениусова норма субмультипликативна:

$$\|AB\|_F \leq \|A\|_F \|B\|_F.$$

Доказательство. Пользуясь новым определением и неравенством Коши-Буняковского-Шварца (0.3), запишем

$$\begin{aligned}\|AB\|_F^2 &= \sum_{i=1}^{d}\sum_{j=1}^{k} \left|\sum_{t=1}^{n} a_{it}b_{tj}\right|^2 \leq \sum_{i=1}^{d}\sum_{j=1}^{k} \left(\sum_{t=1}^{n} |a_{it}|^2\right)\left(\sum_{m=1}^{n} |b_{mj}|^2\right) \\ &= \left(\sum_{i=1}^{d}\sum_{t=1}^{n} |a_{it}|^2\right)\left(\sum_{m=1}^{n}\sum_{j=1}^{k} |b_{mj}|^2\right) = \|A\|_F^2 \|B\|_F^2.\end{aligned}$$

■

Фробениусова норма широко используется в анализе методов наряду с подчиненными. Действительно, её вычисление значительно проще, чем, например, второй нормы, при этом она субмультипликативна.

Замечание С1.9. Используя эквивалентное определение, можно заметить, что для квадратных матриц выполнено

$$\|A\|_F^2 = \mathrm{Tr}\left(A^\top A\right),$$

где $\mathrm{Tr}(A) = \sum_{i=1}^{d} a_{ii}$ — *след матрицы*. Таким образом, мы говорим, что фробениусова норма порождена скалярным произведением матриц:

$$\langle A, B\rangle = \mathrm{Tr}\left(A^\top B\right).$$

Это замечание окажется особенно полезным для дифференцирования функций, принимающих на вход матрицу. Напомним без доказательства ряд важных свойств следа.

Утверждение С1.5. Для $\mathrm{Tr} : \mathbb{R}^{d\times d} \to \mathbb{R}$ выполнены следующие равенства:

- $\mathrm{Tr}(A^\top) = \mathrm{Tr}(A)$,

- $\operatorname{Tr}(A+B) = \operatorname{Tr}(A) + \operatorname{Tr}(B)$,
- $\operatorname{Tr}(cA) = c\operatorname{Tr}(A)$,
- $\operatorname{Tr}(A_1 \dots A_n) = \operatorname{Tr}(A_n A_1 \dots A_{n-1})$.

C1.3 Сопряжённые нормы

Если на исходном пространстве задана норма, то на сопряжённом пространстве задана сопряжённая норма.

Определение C1.8. Пусть в $\mathbb{R}^d$ задана норма $\|\cdot\|$. Тогда *сопряжённая норма* $\|\cdot\|_* : \mathbb{R}^d \to \mathbb{R}$ определяется как

$$\|y\|_* = \sup_{\|x\| \le 1} x^\top y.$$

Введённая функция еще не обязана удовлетворять свойствам нормы, докажем их напрямую.

Утверждение C1.6. $\|\cdot\|_*$ является нормой в $\mathbb{R}^d$.

Доказательство. Рассмотрим свойства нормы. Однородность очевидна, покажем положительную определенность: для произвольного $y \neq 0$ можно взять $x = \frac{y}{\|y\|}$, для которого выполняется

$$x^\top y = \frac{y^\top y}{\|y\|} > 0.$$

А для $y = 0$ имеем

$$\|0\|_* = \sup_{\|x\| \le 1} x^\top 0 = 0.$$

Остается показать неравенство треугольника: $\forall y_1, y_2 \in \mathbb{R}^d$:

$$\|y_1 + y_2\|_* = \sup_{\|x\| \le 1} x^\top (y_1 + y_2) \le \sup_{\|x\| \le 1} x^\top y_1 + \sup_{\|x\| \le 1} x^\top y_2 = \|y_1\|_* + \|y_2\|_*.$$

Следовательно, сопряжённая норма является нормой. ■

Пример C1.11. Пусть $\frac{1}{p} + \frac{1}{q} = 1$, тогда сопряжённая норма к $\|\cdot\|_p$ имеет вид $\|\cdot\|_* = \|\cdot\|_q$.

Доказательство. Для начала покажем, что $\forall y \in \mathbb{R}^d \ \|y\|_* \le \|y\|_q$. Из неравенства Гёльдера (0.2) $\forall x, y \in \mathbb{R}^d : \|x\|_p \le 1$:

$$x^\top y \le \|x\|_p \|y\|_q \le \|y\|_q.$$

Покажем, что равенство достигается. Пусть $y \neq 0$ и $x \in \mathbb{R}^d$, тогда:

$$x_i = \frac{|y_i|^{q-1} \cdot \operatorname{sign}(y_i)}{\|y\|_q^{q-1}}$$

Нетрудно проверить, что $\|x\|_p = 1$, кроме того имеем

$$x^\top y = \frac{\sum\limits_{i=1}^{n} |y_i|^q}{\|y\|_q^{q-1}} = \|y\|_q.$$

Таким образом, $\forall y \in \mathbb{R}^d \ \|y\|_* = \|y\|_q$. ■

C1.4 Квадратичные формы

Часто специальные свойства матриц играют ключевую роль в эффективности методов оптимизации и гарантируют некоторые теоретические оценки. Разберёмся с ними подробнее.

Определение C1.9. Пусть на $\mathbb{R}^d$ задана симметричная билинейная (линейная по обоим аргументам) функция $\alpha : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$, тогда соответствующая ей *квадратичная форма* $Q : \mathbb{R}^d \to \mathbb{R}$ определяется следующим образом $\forall x \in \mathbb{R}^d$:

$$Q(x) = \sum_{i=1}^{d} \sum_{j=1}^{d} \alpha_{ij} x_i x_j, \ \alpha_{ij} = \alpha_{ji}.$$

Замечание C1.10. Заметим, что любой квадратичной форме в соответствие можно поставить матрицу $A \in \mathbb{S}^d$. То есть $A = \{\alpha\}_{ij}$, и $Q(x) = x^\top Ax$.

Далее мы будем отождествлять квадратичную форму и соответствующую ей симметричную матрицу.

Определение C1.10. Квадратичная форма $A \in \mathbb{S}^d$ называется *положительно определённой (полуоопределённой)*, если $\forall x \in \mathbb{R}^d \setminus \{0\}$:

$$x^\top Ax \underset{(\geq)}{>} 0.$$

Аналогично определяются *отрицательная определённость (полуопределённость)*.

Замечание C1.11. Множество положительно полуопределённых матриц обозначается $\mathbb{S}^d_+$. Если $A \in \mathbb{S}^d_+$, то $A \succeq 0$. Множество положительно определённых матриц обозначается $\mathbb{S}^d_{++}$. Если $A \in \mathbb{S}^d_{++}$, то $A \succ 0$.

Полезными фактами о квадратичных формах, которые пригодятся нам в дальнейшем, являются критерии Сильвестра, которые позволяют проверять матрицы на положительную/отрицательную определённость и полуопределённость.

Теорема C1.2. (Критерий Сильвестра)

- Квадратичная форма положительно определена $\iff$ все угловые миноры соответствующей ей матрицы положительны.
- Квадратичная форма положительно полуопределена $\iff$ все главные миноры соответствующей ей матрицы неотрицательны. *Главным минором* называется определитель подматрицы, симметричной относительно главной диагонали.

Пример С1.12. Покажите, что матрица

$$A = \begin{pmatrix} 2 & -1 & 0 \\ -1 & 2 & -1 \\ 0 & -1 & 2 \end{pmatrix}.$$

положительно определённая.

Решение. Угловой минор порядка 1: $2 > 0$. Угловой минор порядка 2:

$$\det \begin{pmatrix} 2 & -1 \\ -1 & 2 \end{pmatrix} = 5 > 0.$$

Угловой минор порядка 3:

$$\det \begin{pmatrix} 2 & -1 & 0 \\ -1 & 2 & -1 \\ 0 & -1 & 2 \end{pmatrix} = 4 > 0.$$

По критерию Сильвестра, матрица A положительно определена. ■

Пример С1.13. Покажите, что матрица

$$A = \begin{pmatrix} 2 & -1 & 1 \\ -1 & 1 & 0 \\ 1 & 0 & 1 \end{pmatrix}$$

положительно полуопределённая.

Решение. Заметим, что все диагональные элементы неотрицательны. Рассмотрим теперь миноры порядка 2:

$$\det \begin{pmatrix} 2 & -1 \\ -1 & 1 \end{pmatrix} = 3 \geq 0,$$

$$\det \begin{pmatrix} 2 & 1 \\ 1 & 1 \end{pmatrix} = 1 \geq 0,$$

$$\det \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} = 1 \geq 0.$$

Теперь рассмотрим единственный главный минор порядка 3:

$$\det \begin{pmatrix} 2 & -1 & 1 \\ -1 & 1 & 0 \\ 1 & 0 & 1 \end{pmatrix} = 0.$$

По критерию Сильвестра матрица A положительно полуопределена. Заметим, что она не является положительно определённой. ■

Замечание С1.12. Отметим, что неотрицательности только лишь угловых миноров недостаточно для положительной полуопределённости. Действительно, рассмот-

рим матрицу

$$A = \begin{pmatrix} 0 & 0 \\ 0 & -1 \end{pmatrix}.$$

Нетрудно видеть, что все угловые миноры матрицы равны нулю. Тем не менее, она не является положительно полуопределённой. Действительно,

$$(0,1)A\,(0,1)^\top = (0,1)\,(0,-1)^\top = -1.$$

C2 Матрично-векторное дифференцирование

Классические курсы анализа начинаются с введения в дифференциальное исчисление. Аналогично, теория дифференцирования строится в рамках выпуклого анализа. Чтобы поддерживать наибольшую общность, мы формулируем её для полных линейных нормированных пространств. Это окажется полезным во второй половине курса, когда разговор пойдет о сопряженных конусах. Кроме этого, мы также подробно поговорим о частных случаях — пространствах векторов и матриц.

Определение C3.1. Линейное нормированное пространство $\big(X, \|\cdot\|\big)$ называется *банаховым*, если оно полно по метрике, порождённой нормой $\|\cdot\|$.

C2.1 Первый дифференциал

Определение C3.2. Пусть U, V — множества в банаховых пространствах. Линейный непрерывный оператор $\mathrm{d}f\big(x\big)$ называется *производной отображения* $f : U \to V$ *в точке* $x \in \operatorname{int} U$, если выполнено

$$f\big(x+h\big) = f\big(x\big) + \mathrm{d}f\big(x\big)\,[h] + o\big(\|h\|\big),\ \|h\| \to 0.$$

Сама функция f в случае существования такого оператора называется *дифференцируемой по Фреше в точке*.

Замечание C3.1. Если точка не является внутренней, то понятие дифференцируемости не определено.

Определение C3.3. Действие оператора производной $\mathrm{d}f\big(x\big)$ на приращение h будем называть *дифференциалом отображения* f.

Дифференциал $\mathrm{d}f\big(x\big)\,[h]$ показывает, как значение функции меняется вдоль направления $h \in U$.

Определение C3.4. Пусть U, V — множества в банаховых пространствах. Отображение $f : U \to V$ называется *дифференцируемым по Гато в точке* $x \in \operatorname{int} U$, если сушествует предел

$$\mathrm{d}f\big(x\big)\,[h] = \lim_{t\to 0} \frac{f\big(x+th\big) - f\big(x\big)}{t},\ \forall h \in U.$$

Из дифференцируемости по Фреше следует дифференцируемость по Гато, однако обратное неверно.

Определение C3.5. В конечномерном случае $\mathrm{d}f\big(x\big)\,[e_i]$ из Определения C3.4 будем называть *i-ой частной производной* f и обозначать $\frac{\partial f}{\partial x_i}$. Здесь e_i — i-й элемент стандартного базиса.

Пример C3.1. Из дифференцируемости по Гато не следует дифференцируемость по Фреше.

Доказательство. Пусть $f(x) = \|x\|_2$. Тогда её дифференциал по Гато в нуле равен

$$\mathrm{d}f(0)\,[h] = \|h\|_2.$$

Но вторая норма нелинейная, что противоречит Определению С3.2. Таким образом, вторая норма не дифференцируема по Фреше в нуле. ■

Замечание С3.2. В дальнейшем, если функция дифференцируема, то будем воспринимать, что она дифференцируема по Фреше.

Далее сфокусируемся на *гильбертовых пространствах*, то есть таких полных линейных нормированных пространствах, в которых норма порождается скалярным произведением. Без доказательства рассмотрим теорему представлений Рисса.

Теорема С3.1. Пусть $\mathrm{d}f(x) : H \to \mathbb{R}$ — линейный непрерывный функционал, действующий на гильбертовом пространстве H. Тогда существует единственный $g_f \in H$, такой, что

$$\mathrm{d}f(x)\,[h] = \langle g_f, h\rangle.$$

g_f будем называть *градиентом отображения f в точке x* и обозначать $\nabla f(x)$.

Замечание С3.3. Поговорим немного о частных случаях. Если отображение принимало вектор и отображало его в число, градиентом будет вектор частных производных. Если на вход подавалась матрица, то градиент — также матрица. Более интересно рассмотреть $f : \mathbb{R}^d \to \mathbb{R}^n$. Представим

$$f(x) = \big(f_1(x), \ldots, f_n(x)\big)^\top.$$

Каждая координатная функция — понятное с точки зрения производной отображение $\mathbb{R}^d \to \mathbb{R}$. Получается, производная отображения $\mathbb{R}^d \to \mathbb{R}^n$ — есть транспонированная матрица Якоби (J) этого отображения. В общем же случае отображения матрицы в матрицу для построения градиента достаточно найти все частные производные в виде тензора

$$\left\{\frac{\partial f_{ij}}{\partial x_{kl}}\right\}_{ijkl}.$$

Следует помнить, что из существования частных производных ещё не следует дифференцируемость. Однако на практике чаще всего все эти частные производные непрерывны, и, как следствие, функция дифференцируема. Рассуждения, приведённые в Замечании С3.3, для наглядности сформулируем в виде таблицы.

Вход/Выход	Скаляр $\mathbb{R}$	Вектор $\mathbb{R}^d$
Скаляр $\mathbb{R}$	$\nabla f(x)$ скаляр, h скаляр	$\nabla f(x)$ вектор, h скаляр
Вектор $\mathbb{R}^d$	$\nabla f(x)$ вектор, h вектор	$\nabla f(x)$ матрица, h вектор
Матрица $\mathbb{R}^{n\times d}$	$\nabla f(x)$ матрица, h матрица	$\nabla f(x)$ тензор, h матрица

Стоит отметить, что данная таблица верна для произвольных скалярных произведений, а не только для стандартного.

Утверждение С3.1. Из формулы градиента верно, что

$$\mathrm{d}f(x) = \big(\nabla f(x)\big)^\top \mathrm{d}x \iff \nabla f(x) = J_f^\top.$$

Доказательство. Утверждение легко проверяется покомпонентной подстановкой. ■

С2.2 Второй дифференциал

Пусть U, V — множества в банаховых пространствах. Пусть отображение $f : U \to V$, дифференцируемое в каждой точке $x \in \operatorname{int} U$. Второй дифференциал будем определять, фиксируя приращение и рассматривая первый дифференциал как функцию от x:

$$g(x) = \mathrm{d}f(x)\,[h_1].$$

Определение С3.6. Если в некоторой точке x функция $g = \mathrm{d}f(x)\,[h_1]$ дифференцируема, то билинейный оператор $\mathrm{d}g(x)\,[h_2] = \mathrm{d}(\mathrm{d}f\,[h_1])\,(x)[h_2]$ называется *оператором второй производной.*

По аналогии определяется третий дифференциал $\mathrm{d}^3 f(x)\,[h_1, h_2, h_3]$, четвёртый и так далее.

Утверждение С3.2. Пусть H — гильбертово пространство. Рассмотрим дважды дифференцируемое отображение $f : H \to \mathbb{R}$. Тогда существует единственный $G_f \in H \times H$, такой что

$$\mathrm{d}^2 f(x)\,[h_1, h_2] = \langle G_f h_1, h_2 \rangle.$$

G_f будем называть *гессианом отображения* f в точке x и обозначать $\nabla^2 f(x)$.

Доказательство. Утверждение очевидно следует из Теоремы С3.1, поскольку оператор второй производной является билинейным непрерывным функционалом. ■

Замечание С3.4. В стандартном базисе гессиан имеет вид

$$\big(\nabla^2 f(x)\big)_{ij} = \frac{\partial^2 f}{\partial x_i \partial x_j}(x).$$

Утверждение С3.3. Из формулы гессиана верно, что

$$\mathrm{d}\big(\nabla f(x)\big) = \big(\nabla^2 f(x)\big)^\top \mathrm{d}x \iff \nabla^2 f(x) = J_{\nabla f}^\top.$$

Доказательство. Утверждение легко проверяется покомпонентной подстановкой. ■

С2.3 Примеры и задачи

В рамках этого пособия вектора обозначаются маленькими латинскими буквами, матрицы/тензоры — заглавными латинскими.
Известные из курса математического анализа правила дифференцирования (производная произведения, частного, композиции) сохраняются. Мы напомним их, не обращаясь к доказательствам.

Утверждение С3.4. Пусть U — множество в банаховом пространстве. Для векторных дифференциалов справедливы утверждения:

- $\mathrm{d}(\alpha x) = \alpha\,\mathrm{d}x$;
- $\mathrm{d}(x+y) = \mathrm{d}x + \mathrm{d}y$;
- $\mathrm{d}\langle x, y\rangle = \langle \mathrm{d}x\,, y\rangle + \langle x, \mathrm{d}y\rangle$;
- $\mathrm{d}(f(x)g(x)) = \mathrm{d}(f(x))\,g(x) + f(x)\,\mathrm{d}(g(x))$ $(f : U \to \mathbb{R},\ g : U \to \mathbb{R})$;
- $\mathrm{d}(g(f(x))) = g'(f)\,\mathrm{d}f(x)$ $(f : U \to f(U),\ g : f(U) \to \mathbb{R})$;
- $J_{g(f)} = J_g J_f$ $(f : U \to f(U),\ g : f(U) \to \mathbb{R})$.

Утверждение С3.5. Пусть U — множество в банаховом пространстве. Для матричных дифференциалов справедливы утверждения:

- $\mathrm{d}(\alpha X) = \alpha\,\mathrm{d}X$;
- $\mathrm{d}(AXB) = A\,\mathrm{d}X\,B$;
- $\mathrm{d}(X+Y) = \mathrm{d}X + \mathrm{d}Y$;
- $\mathrm{d}(XY) = (\mathrm{d}X)Y + X(\mathrm{d}Y)$;
- $\mathrm{d}(X^\top) = (\mathrm{d}X)^\top$;
- $\mathrm{d}\langle X, Y\rangle = \langle \mathrm{d}X\,, Y\rangle + \langle X, \mathrm{d}Y\rangle$;
- $\mathrm{d}(f(X)g(X)) = \mathrm{d}(f(X))\,g(X) + f(X)\,\mathrm{d}(g(X))$ $(f : U \to \mathbb{R},\ g : U \to \mathbb{R})$;
- $\mathrm{d}(g(f(X))) = g'(f)\,\mathrm{d}f(X)$ $(f : U \to f(U),\ g : f(U) \to \mathbb{R})$;
- $J_{g(f)} = J_g J_f$ $(f : U \to f(U),\ g : f(U) \to \mathbb{R})$.

Перед тем, как перейти к вычислению нетривиальных примеров, придётся воспользоваться определением и получить выражения для базовых градиентов.

Пример С3.2. Справедливы следующие выражения:

- $\mathrm{d}\langle c, x\rangle\,[h] = \langle c, h\rangle$;
- $\mathrm{d}\langle Ax, x\rangle\,[h] = \langle (A + A^\top)x, h\rangle$;
- $\mathrm{d}(X^{-1})\,[H] = -X^{-1}HX^{-1}$;
- $\mathrm{d}\det(X)\,[H] = \langle \det(X)X^{-\top}, H\rangle$.

Доказательство.

- Для $f(x) = \langle c, x\rangle$ запишем

$$f(x+h) - f(x) = \langle c, x+h\rangle - \langle c, x\rangle = \langle c, h\rangle.$$

При этом $h \to \langle c, h\rangle$ — линейный оператор. То есть по определению

$$\mathrm{d}f(x)\,[h] = \langle c, h\rangle.$$

- Для $f(x) = \langle Ax, x\rangle$ запишем

$$f(x+h) - f(x) = \langle Ax + Ah, x + h\rangle - \langle Ax, x\rangle = \langle (A + A^\top)x, h\rangle + \langle Ah, h\rangle.$$

Заметим, что

$$\langle Ah, h\rangle \leq \|Ah\|\|h\| \leq \|A\|\|h\|^2 = o(\|h\|),$$

где первое неравенство следует из Коши-Буняковского-Шварца (0.3), а второе — из подчинённости матричной нормы (Замечание С1.6). При этом $h \to \langle (A + A^\top)x, h\rangle$ — линейный оператор. То есть по определению

$$\mathrm{d}f(x)\,[h] = \langle (A + A^\top)x, h\rangle.$$

- Для $f(X) = X^{-1}$ запишем

$$\begin{aligned} f(X+H) - f(X) &= (X+H)^{-1} - X^{-1} = \big(X\big(I_d + X^{-1}H\big)\big)^{-1} - X^{-1} \\ &= \Big(\big(I_d + X^{-1}H\big)^{-1} - I_d\Big)X^{-1}. \end{aligned}$$

Поскольку H мало, запишем разложение в ряд Тейлора:

$$\big(I_d + X^{-1}H\big)^{-1} = I_d - X^{-1}H + \sum_{k=2}^{\infty} \big(-X^{-1}H\big)^k.$$

Оценим норму последнего слагаемого:

$$\left\|\sum_{k=2}^{\infty} \big(-X^{-1}H\big)^k\right\| \leq \sum_{k=2}^{\infty} \left\|\big(-X^{-1}H\big)^k\right\| \leq \sum_{k=2}^{\infty} \|-X^{-1}\|^k \|H\|^k = \frac{\|X^{-1}\|^2\|H\|^2}{1 - \|X^{-1}\|\|H\|}$$

$$= o(\|H\|),$$

где первое неравенство получается из неравенства треугольника в пределе, второе — из свойств матричной нормы, а третье равенство — это сумма геометрической прогрессии. В итоге получаем разность

$$f(X+H) - f(X) = -X^{-1}HX^{-1} + o(\|H\|),$$

при этом $H \to -X^{-1}HX^{-1}$ — линейный оператор. То есть по определению

$$\mathrm{d}f(X)\,[H] = -X^{-1}HX^{-1}.$$

- Для $f(X) = \det(X)$ запишем

$$\begin{aligned} f(X+H) - f(X) &= \det(X+H) - \det(X) = \det\big(X\big(I_d + X^{-1}H\big)\big) - \det(X) \\ &= \det(X)(\det\big(I_d + X^{-1}H\big) - 1). \end{aligned}$$

Отдельно оценим $\det\big(I_d + X^{-1}H\big)$. Пусть $\lambda_i\big(X^{-1}H\big)$ — собственные числа матрицы $X^{-1}H$ (в произвольном порядке и, возможно, комплексные). Тогда собственными числами матрицы $I_d + X^{-1}H$ будут $1 + \lambda_i(X^{-1}H)$. Поэтому

$$\begin{aligned}\det\big(I_d + X^{-1}H\big) &= \prod_{i=1}^{d}[1 + \lambda_i(X^{-1}H)] \\ &= 1 + \sum_{i=1}^{d}\lambda_i\big(X^{-1}H\big) + \left(\sum_{1\leq i\leq j\leq d}\lambda_i\big(X^{-1}H\big)\lambda_j\big(X^{-1}H\big) + \ldots\right),\end{aligned}$$

где многоточие обозначает всевозможные тройки, четвёрки и т.д. из $\lambda_i\big(X^{-1}H\big)$. Для произвольной матрицы $A \in \mathbb{R}^{d\times d}$ все её собственные числа по модулю не превосходят её нормы (Утверждение С1.2). Следовательно, всё выражение в скобках — есть $o\big(\|H\|\big)$, поскольку в каждом слагаемом больше одного собственного числа. Таким образом,

$$\det\big(I_d + X^{-1}H\big) = 1 + \sum_{i=1}^{d}\lambda_i\big(X^{-1}H\big) + o\big(\|H\|\big) = 1 + \operatorname{Tr}\big(X^{-1}H\big) + o\big(\|H\|\big).$$

При этом $H \to \det(X)\operatorname{Tr}\big(X^{-1}H\big)$ — линейный оператор. То есть по определению

$$\mathrm{d}f\big(X\big)\,[H] = \det(X)\operatorname{Tr}\big(X^{-1}H\big).$$

■

Замечание С3.5. Мы доказали формулу $\mathrm{d}(\det(X)) = \det(X)\big\langle X^{-\top}, \mathrm{d}X\big\rangle$ только для обратимых X. Однако выражение для $\mathrm{d}(\det(X))$ можно получить и для произвольной матрицы из $\mathbb{R}^{d\times d}$. Эта формула называется *формулой Якоби* и записывается как

$$\mathrm{d}(\det(X)) = \big\langle \operatorname{Adj}(X)^{\top}, \mathrm{d}X\big\rangle,$$

где $\operatorname{Adj}(X)$ — присоединённая матрица к X. Она определяется как

$$\operatorname{Adj}(X)_{ji} = (-1)^{(i+j)}M_{ij},$$

где M_{ij} — дополнительный минор, определитель матрицы, получившийся вычеркиванием i-ой строки и j-ого столбца из X. В случае невырожденной X выполнено $\operatorname{Adj}(X) = \det(X)X^{-1}$ и формулы переходят друг в друга. Вспомним формулу вычисления определителя через дополнительные миноры по i-ой строке

$$\det(X) = \sum_{j=1}^{d} x_{ij}(-1)^{(i+j)}M_{ij}.$$

Тогда градиент равен

$$\frac{\partial\det(X)}{\partial x_{ij}} = (-1)^{(i+j)}M_{ij} = \operatorname{Adj}(X)_{ji}.$$

C2.3.1 Дифференцирование по вектору

В Машинном обучении широко известен Метод Наименьших Квадратов (МНК). По матрице признаков $\tilde{A}$ и вектору параметров x требуется линейно приблизить целевой вектор $\tilde{b}$. Для решения этой задачи требуется вычислять градиент невязки

$$f(x) = \frac{1}{2}\|\tilde{A}x - \tilde{b}\|^2 = \frac{1}{2}\langle \tilde{A}x, \tilde{A}x\rangle - \langle \tilde{A}x, \tilde{b}\rangle + \frac{1}{2}\|\tilde{b}\|_2^2$$

Если мы введём обозначения $A = \tilde{A}^\top \tilde{A}$, $b = -\tilde{b}^\top \tilde{A}$, $c = \frac{1}{2}\|\tilde{b}\|_2^2$, то вычисление градиента невязки эквивалентно вычислению градиента функции

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

Пример C3.3. Найдите $\nabla f(x)$ и $\nabla^2 f(x)$ функции $f : \mathbb{R}^d \to \mathbb{R}$:

$$f(x) = \frac{1}{2}\langle Ax, x\rangle + \langle b, x\rangle + c,\ A \in \mathbb{R}^{d\times d},\ b \in \mathbb{R}^d,\ c \in \mathbb{R}.$$

Решение. Найдём первый дифференциал

$$\begin{aligned} \mathrm{d}f(x) &= \mathrm{d}\left(\frac{1}{2}\langle Ax, x\rangle + \langle b, x\rangle + c\right) = \frac{1}{2}\langle (A + A^\top)x, \mathrm{d}x\rangle + \langle b, \mathrm{d}x\rangle + 0 \\ &= \left\langle \frac{1}{2}(A + A^\top)x + b, \mathrm{d}x \right\rangle. \end{aligned}$$

Приведя к стандартному виду $\mathrm{d}f(x) = \langle \nabla f(x), \mathrm{d}x\rangle$, получаем

$$\nabla f(x) = \frac{1}{2}(A + A^\top)x + b.$$

Для поиска гессиана фиксируем первое приращение $\mathrm{d}x_1$ у первого дифференциала и берём уже от него ещё один дифференциал

$$\mathrm{d}^2 f(x) = \mathrm{d}\left\langle \frac{1}{2}(A + A^\top)x + b, \mathrm{d}x_1 \right\rangle = \left\langle \frac{1}{2}(A + A^\top)\,\mathrm{d}x\,, \mathrm{d}x_1 \right\rangle.$$

Переносим и транспонируем матрицу в скалярном произведении, но поскольку $A + A^\top$ симметричная, то она не меняется:

$$\mathrm{d}^2 f(x) = \left\langle \mathrm{d}x\,, \frac{1}{2}(A + A^\top)^\top\,\mathrm{d}x_1 \right\rangle = \left\langle \frac{1}{2}(A + A^\top)\,\mathrm{d}x_1\,, \mathrm{d}x \right\rangle.$$

Следовательно, приведя к стандартному виду $\mathrm{d}^2 f(x) = \langle \nabla^2 f(x)\,\mathrm{d}x_1\,, \mathrm{d}x\rangle$, получаем гессиан

$$\nabla^2 f(x) = \frac{1}{2}(A + A^\top).$$

■

Замечание C3.6. Заметим, что в случае, если A симметричная, то

$$\nabla f(x) = Ax + b, \quad \nabla^2 f(x) = A.$$

Пример С3.4. Найдите $\nabla f(x)$ и $\nabla^2 f(x)$ функции $f : \mathbb{R}^d \to \mathbb{R}$:

$$f(x) = \ln \langle Ax, x\rangle,\ A \in \mathbb{S}^d_{++}.$$

Решение. Найдём первый дифференциал

$$\mathrm{d}f(x) = \mathrm{d}\ln\langle Ax, x\rangle = \frac{1}{\langle Ax, x\rangle}\,\mathrm{d}\langle Ax, x\rangle = \frac{2\langle Ax, \mathrm{d}x\rangle}{\langle Ax, x\rangle} = \left\langle \frac{2Ax}{\langle Ax, x\rangle}, \mathrm{d}x \right\rangle.$$

Приведя к стандартному виду $\mathrm{d}f(x) = \langle \nabla f(x), \mathrm{d}x\rangle$, получаем

$$\nabla f(x) = \frac{2Ax}{\langle Ax, x\rangle}.$$

Теперь найдём дифференциал градиента

$$\mathrm{d}\left(\frac{2Ax}{\langle Ax, x\rangle}\right) = \frac{\mathrm{d}(2Ax)\,\langle Ax, x\rangle - (2Ax)\,\mathrm{d}\langle Ax, x\rangle}{\langle Ax, x\rangle^2} = \frac{2\langle Ax, x\rangle A\,\mathrm{d}x - 4Ax\langle Ax, \mathrm{d}x\rangle}{\langle Ax, x\rangle^2}$$
$$= \left(\frac{2A}{\langle Ax, x\rangle} - \frac{4Axx^\top A}{\langle Ax, x\rangle^2}\right)\mathrm{d}x = J_{\nabla f}\,\mathrm{d}x\,.$$

Поскольку $\nabla^2 f(x) = J^\top_{\nabla f}$, а гессиан симметричен из-за непрерывности, то

$$\nabla^2 f(x) = \frac{2A}{\langle Ax, x\rangle} - \frac{4Axx^\top A}{\langle Ax, x\rangle^2}.$$

■

Пример С3.5. Найдите $\nabla f(x)$ и $\nabla^2 f(x)$ функции $f : \mathbb{R}^d \setminus \mathbf{0} \to \mathbb{R}$:

$$f(x) = \|x\|_2.$$

Решение. Найдём первый дифференциал

$$\mathrm{d}f(x) = \mathrm{d}\langle x, x\rangle^{\frac{1}{2}} = \frac{\mathrm{d}(\langle x, x\rangle)}{2\langle x, x\rangle^{\frac{1}{2}}} = \left\langle \frac{2x}{2\langle x, x\rangle^{\frac{1}{2}}}, \mathrm{d}x \right\rangle = \left\langle \frac{x}{\|x\|}, \mathrm{d}x \right\rangle.$$

Приведя к стандартному виду $\mathrm{d}f(x) = \langle \nabla f(x), \mathrm{d}x\rangle$, получаем

$$\nabla f(x) = \frac{x}{\|x\|}.$$

Теперь посчитаем второй дифференциал, зафиксировав приращение $\mathrm{d}x_1$ первого

$$\mathrm{d}^2 f\,(x) = \mathrm{d}\left\langle \frac{x}{\|x\|}, \mathrm{d}x_1 \right\rangle = \left\langle \frac{\mathrm{d}x\,\|x\| - x\,\mathrm{d}(\|x\|)}{\|x\|^2}, \mathrm{d}x_1 \right\rangle$$
$$= \left\langle \frac{\mathrm{d}x}{\|x\|} - x\left\langle \frac{x}{\|x\|^3}, \mathrm{d}x \right\rangle, \mathrm{d}x_1 \right\rangle = \left\langle \left(\frac{I_d}{\|x\|} - \frac{xx^\top}{\|x\|^3}\right)\mathrm{d}x\,, \mathrm{d}x_1 \right\rangle.$$

Поскольку $\nabla^2 f(x) = J^\top_{\nabla f}$, а гессиан симметричен из-за непрерывности, то

$$\nabla^2 f(x) = \frac{I_d}{\|x\|} - \frac{xx^\top}{\|x\|^3}.$$

■

Пример С3.6. Найдите $\nabla f(x)$ и $\nabla^2 f(x)$ функции $f : \mathbb{R}^d \setminus \mathbf{0} \to \mathbb{R}$:

$$f(x) = \frac{1}{3}\|x\|_2^3.$$

Решение. Найдём первый дифференциал

$$\mathrm{d}f(x) = \frac{1}{3}\,\mathrm{d}\langle x, x\rangle^{3/2} = \frac{1}{2}\langle x, x\rangle^{1/2}\,\mathrm{d}\langle x, x\rangle = \langle x, x\rangle^{1/2}\langle x, \mathrm{d}x\rangle = \langle x\|x\|, \mathrm{d}x\rangle.$$

Приведя к стандартному виду $\mathrm{d}f = \langle \nabla f(x), \mathrm{d}x\rangle$, получаем

$$\nabla f(x) = \|x\|x.$$

Найдём второй дифференциал

$$\begin{aligned}\mathrm{d}^2 f(x) = \mathrm{d}\big(\|x\|\langle x, \mathrm{d}x_1\rangle\big) &= \left\langle \frac{x}{\|x\|}, \mathrm{d}x \right\rangle \langle x, \mathrm{d}x_1\rangle + \|x\|\langle \mathrm{d}x, \mathrm{d}x_1\rangle \\ &= \left\langle \mathrm{d}x\,, \left(\frac{xx^\top}{\|x\|} + I_d\|x\|\right)\mathrm{d}x_1 \right\rangle.\end{aligned}$$

Получаем гессиан

$$\nabla^2 f(x) = \frac{xx^\top}{\|x\|} + I_d\|x\|.$$

■

Замечание С3.7. Заметим, что гессиан не определен в точке $x = 0$, поскольку мы дифференцировали норму, а её градиент не определен в 0. Тем не менее, можно найти производную отображения из условия в $x = 0$ по определению. Зафиксируем приращение h_1 и рассмотрим

$$\begin{aligned}\lim_{h_2\to 0} \frac{\|\mathrm{d}f(h_2)\,[h_1] - \mathrm{d}f(0)\,[h_1]\|}{\|h_2\|} = \lim_{h_2\to 0} \frac{|\langle \|h_2\|h_2, h_1\rangle|}{\|h_2\|} &= \lim_{h_2\to 0} \frac{\|h_2\|\,|\langle h_2, h_1\rangle|}{\|h_2\|} \\ &= \lim_{h_2\to 0} |\langle h_2, h_1\rangle| = 0.\end{aligned}$$

Следовательно, по определению вторая производная в точке $x = 0$ равна 0. Можно даже сказать, что отображение дважды непрерывно дифференцируемо, потому что

$$\lim_{x\to 0}\left(\frac{xx^\top}{\|x\|} + I_d\|x\|\right) = 0.$$

Логистическая регрессия — модель машинного обучения для задачи разделения двух классов. Её обучение может сводится к оптимизации функции

$$f(x) = \ln\big(1 + \exp\big(\langle x, a\rangle\big)\big).$$

Пример С3.7. Найдите $\nabla f(x)$ и $\nabla^2 f(x)$ функции $f : \mathbb{R}^d \to \mathbb{R}$:

$$f(x) = \ln\bigl(1 + \exp\bigl(\langle x, a\rangle\bigr)\bigr),\ a \in \mathbb{R}^d.$$

Решение. Найдём первый дифференциал

$$\begin{aligned}\mathrm{d}f(x) &= \mathrm{d}\ln\bigl(1 + \exp\bigl(\langle x, a\rangle\bigr)\bigr) = \frac{1}{1 + \exp\bigl(\langle x, a\rangle\bigr)}\,\mathrm{d}\bigl(1 + \exp\bigl(\langle x, a\rangle\bigr)\bigr)\\ &= \frac{\exp\bigl(\langle x, a\rangle\bigr)}{1 + \exp\bigl(\langle x, a\rangle\bigr)}\,\mathrm{d}\langle x, a\rangle = \left\langle \frac{\exp\bigl(\langle x, a\rangle\bigr)}{1 + \exp\bigl(\langle x, a\rangle\bigr)} a, \mathrm{d}x \right\rangle.\end{aligned}$$

Для удобства введём сигмоиду

$$\sigma(x) := \frac{1}{1 + \exp(-x)}.$$

При этом заметим, что $\sigma(-x) = 1 - \sigma(x)$ и $\sigma'(x) = \sigma(x)(1 - \sigma(x))$. После этого приведём $\mathrm{d}f(x)$ к стандартному виду и получим

$$\nabla f(x) = \sigma\bigl(\langle x, a\rangle\bigr) a.$$

Таким образом, градиент $\nabla f(x)$ — вектор коллинеарный вектору a с коэффициентом $\sigma(\langle a, x\rangle) \in (0, 1)$. От точки x зависит лишь длина градиента, но не направление.
Теперь посчитаем второй дифференциал, зафиксировав приращение $\mathrm{d}x_1$ первого. Имеем

$$\begin{aligned}\mathrm{d}^2 f(x) = \mathrm{d}\langle \sigma\bigl(\langle x, a\rangle\bigr) a, \mathrm{d}x_1\rangle &= \bigl\langle \mathrm{d}\sigma\bigl(\langle x, a\rangle\bigr)\, a, \mathrm{d}x_1\bigr\rangle = \bigl\langle \sigma'\bigl(\langle x, a\rangle\bigr)\,\mathrm{d}\langle x, a\rangle\, a, \mathrm{d}x_1\bigr\rangle\\ &= \bigl\langle \sigma\bigl(\langle x, a\rangle\bigr)\bigl(1 - \sigma\bigl(\langle x, a\rangle\bigr)\bigr)\langle \mathrm{d}x\,, a\rangle a, \mathrm{d}x_1\bigr\rangle\\ &= \sigma\bigl(\langle x, a\rangle\bigr)\bigl(1 - \sigma\bigl(\langle x, a\rangle\bigr)\bigr)\bigl\langle \langle \mathrm{d}x\,, a\rangle a, \mathrm{d}x_1\bigr\rangle\\ &= \sigma\bigl(\langle x, a\rangle\bigr)\bigl(1 - \sigma\bigl(\langle x, a\rangle\bigr)\bigr)(\mathrm{d}x^\top a a^\top \mathrm{d}x_1)\\ &= \sigma\bigl(\langle x, a\rangle\bigr)\bigl(1 - \sigma\bigl(\langle x, a\rangle\bigr)\bigr)\bigl\langle a a^\top \mathrm{d}x_1\,, \mathrm{d}x\bigr\rangle.\end{aligned}$$

Получим

$$\nabla^2 f(x) = \sigma\bigl(\langle x, a\rangle\bigr)\bigl(1 - \sigma\bigl(\langle x, a\rangle\bigr)\bigr) a a^\top.$$

Заметим, что $\nabla^2 f(x)$ — одноранговая матрица, пропорциональная $aa^\top$ с коэффициентом $\sigma\bigl(\langle x, a\rangle\bigr)\bigl(1 - \sigma\bigl(\langle x, a\rangle\bigr)\bigr) \in (0, 0.25)$. Точка x влияет лишь на коэффициент. ■

С2.3.2 Дифференцирование по матрице

Для подсчёта градиентов по матрицам активно применяются производные таких функций, как $\det, \mathrm{Tr}, X^{-1}$. Именно поэтому мы посвятили параграф их вычислению. Более того, очень полезным окажется след матрицы (Утверждение С1.5).

Пример С3.8. Найдите $\nabla f(X)$ функции $f : \mathbb{R}^{n \times d} \to \mathbb{R}$:

$$f(X) = \|AX - B\|_F^2,\ A \in \mathbb{R}^{m \times n},\ B \in \mathbb{R}^{m \times d}.$$

Решение. Сначала отдельно вычислим $\mathrm{d}\|X\|$:

$$\mathrm{d}\|X\| = \mathrm{d}\langle X, X\rangle^{1/2} = \frac{\mathrm{d}\langle X, X\rangle}{2\langle X, X\rangle^{1/2}} = \left\langle \frac{2X}{2\langle X, X\rangle^{1/2}}, \mathrm{d}X \right\rangle = \left\langle \frac{X}{\|X\|}, \mathrm{d}X \right\rangle.$$

Тогда первый дифференциал

$$\begin{aligned}\mathrm{d}f(X) = \mathrm{d}\|AX - B\|_F &= \left\langle \frac{AX - B}{\|AX - B\|}, \mathrm{d}(AX - B) \right\rangle = \left\langle \frac{AX - B}{\|AX - B\|}, A\,\mathrm{d}X \right\rangle \\ &= \mathrm{Tr}\left(\frac{(AX - B)^\top}{\|AX - B\|} A\,\mathrm{d}X \right) = \mathrm{Tr}\left(\left(\frac{A^\top (AX - B)}{\|AX - B\|} \right)^\top \mathrm{d}X \right) \\ &= \left\langle \frac{A^\top (AX - B)}{\|AX - B\|}, \mathrm{d}X \right\rangle.\end{aligned}$$

Тогда градиент

$$\nabla f(X) = \frac{A^\top (AX - B)}{\|AX - B\|}.$$

■

Пример С3.9. Найдите $\nabla f(X)$ функции $f : \mathbb{R}^{d\times d} \to \mathbb{R}$:

$$f(X) = \mathrm{Tr}(AXBX^{-1}),\ A, B \in \mathbb{R}^{d\times d},$$

при $\det(X) \neq 0$.

Решение. Перепишем след через скалярное произведение:

$$f(X) = \langle I_d, AXBX^{-1}\rangle.$$

Найдём первый дифференциал

$$\begin{aligned}\mathrm{d}f(X) &= \langle I_d, \mathrm{d}(AXBX^{-1})\rangle = \langle I_d, A(\mathrm{d}X)BX^{-1} + AXB\,\mathrm{d}(X^{-1})\rangle \\ &= \langle I_d, A(\mathrm{d}X)BX^{-1} - AXBX^{-1}(\mathrm{d}X)X^{-1}\rangle \\ &= \mathrm{Tr}(A(\mathrm{d}X)BX^{-1}) - \mathrm{Tr}(AXBX^{-1}(\mathrm{d}X)X^{-1}) \\ &= \mathrm{Tr}(BX^{-1}A(\mathrm{d}X)) - \mathrm{Tr}(X^{-1}AXBX^{-1}(\mathrm{d}X)) \\ &= \langle A^\top X^{-\top} B^\top - X^{-\top} B^\top X^\top A^\top X^{-\top}, \mathrm{d}X\rangle.\end{aligned}$$

При работе со скалярными произведениями можно переходить к следу и обратно, используя таким образом его полезные свойства. Главное, не забывать о транспонировании.
Тогда градиент

$$\nabla f(X) = A^\top X^{-\top} B^\top - X^{-\top} B^\top X^\top A^\top X^{-\top}.$$

■

Пример С3.10. Найдите $\nabla f(X)$ функции $f : \mathbb{R}^{n\times d} \to \mathbb{R}$:

$$f(X) = \operatorname{Tr}\Big(AX^\top X\Big),\ A \in \mathbb{R}^{d\times d}.$$

Решение. Перепишем след через скалярное произведение

$$f(X) = \langle I_d, AX^\top X\rangle.$$

Найдём первый дифференциал

$$\begin{aligned}\mathrm{d}f(X) &= \mathrm{d}\langle I_d, AX^\top X\rangle = \Big\langle I_d, A\,\mathrm{d}\Big(X^\top X\Big)\Big\rangle = \langle I_d, A\,\mathrm{d}X^\top\, X\rangle + \langle I_d, AX^\top\,\mathrm{d}X\rangle\\ &= \operatorname{Tr}\Big(A\,\mathrm{d}X^\top\, X\Big) + \langle XA^\top, \mathrm{d}X\rangle = \operatorname{Tr}\Big(XA\,\mathrm{d}X^\top\Big) + \langle XA^\top, \mathrm{d}X\rangle\\ &= \langle XA, \mathrm{d}X\rangle + \langle XA^\top, \mathrm{d}X\rangle = \langle XA + XA^\top, \mathrm{d}X\rangle.\end{aligned}$$

Тогда градиент

$$\nabla f(X) = XA + XA^\top.$$

■

Пример С3.11. Найдите $\nabla f(X)$ и $\mathrm{d}^2 f(X)$ функции $f : \mathbb{S}^d_{++} \to \mathbb{R}$:

$$f(X) = \ln(\det(X)).$$

Решение. Найдём первый дифференциал

$$\mathrm{d}f(X) = \mathrm{d}\ln(\det(X)) = \frac{\mathrm{d}(\det(X))}{\det(X)} = \frac{\det(X)\langle X^{-\top}, \mathrm{d}X\rangle}{\det(X)} = \langle X^{-1}, \mathrm{d}X\rangle.$$

Тогда градиент

$$\nabla f(X) = X^{-1}.$$

Теперь посчитаем второй дифференциал, зафиксировав приращение $\mathrm{d}X_1$ первого

$$\mathrm{d}^2 f(X) = \langle \mathrm{d}X^{-1}, \mathrm{d}X_1\rangle = -\langle X^{-1}\,\mathrm{d}X\, X^{-1}, \mathrm{d}X_1\rangle.$$

Мы получили билинейную форму от $\mathrm{d}X, \mathrm{d}X_1$, выписывать тензор производных в явном виде мы не будем. Вместо этого, посмотрим, является ли эта форма отрицательно полуопределённой. По мере прохождения курса станет понятно, зачем мы проверяем это условие. Возьмём $H \in \mathbb{S}^d$ из исходного пространства. Поскольку $X \in \mathbb{S}^d_{++}$, то $X^{-1} \in \mathbb{S}^d_{++}$. Матрицу X^{-1} можно разложить на произведение двух одинаковых матриц, обозначим их $X^{-1/2}$, т.е. $X^{-1} = X^{-1/2}X^{-1/2}$. Такое разложение можно получить, перейдя в ортонормированный базис из собственных векторов V, который всегда существует для симметричных матриц, при этом все собственные значения будут положительными:

$$X^{-1} = V\Lambda V^\top \quad \Longrightarrow\ X^{-1/2} = V\sqrt{\Lambda}V^\top.$$

Тогда

$$\begin{aligned}\mathrm{d}^2 f(X)\,[H,H] &= -\langle X^{-1}HX^{-1}, H\rangle = -\operatorname{Tr}(X^{-1}HX^{-1}H)\\ &= -\operatorname{Tr}\Big(X^{-1/2}X^{-1/2}HX^{-1/2}X^{-1/2}H\Big)\\ &= -\operatorname{Tr}\Big(X^{-1/2}HX^{-1/2}X^{-1/2}HX^{-1/2}\Big)\\ &= -\langle X^{-1/2}HX^{-1/2}, X^{-1/2}HX^{-1/2}\rangle\\ &= -\big\|X^{-1/2}HX^{-1/2}\big\|_F^2 \le 0.\end{aligned}$$

■

Пример C3.12. Найдите $\nabla f(X)$ функции $f : \mathbb{S}^d_{++} \to \mathbb{R}$:

$$f(X) = \det(AX^{-1}B),\ A \in \mathbb{R}^{n\times d}, B \in \mathbb{R}^{d\times n},$$

при $\det(AX^{-1}B) \neq 0$.

Решение. Найдём первый дифференциал

$$\begin{aligned}\mathrm{d}f(X) &= \mathrm{d}(\det(AX^{-1}B)) = \det(AX^{-1}B)\Big\langle (AX^{-1}B)^{-\top}, \mathrm{d}(AX^{-1}B)\Big\rangle\\ &= \det(AX^{-1}B)\Big\langle (AX^{-1}B)^{-\top}, A\,\mathrm{d}X^{-1}\,B\Big\rangle\\ &= -\det(AX^{-1}B)\Big\langle (AX^{-1}B)^{-\top}, AX^{-1}\,\mathrm{d}X\,X^{-1}B\Big\rangle\\ &= -\det(AX^{-1}B)\operatorname{Tr}\Big((AX^{-1}B)^{-1}AX^{-1}\,\mathrm{d}X\,X^{-1}B\Big)\\ &= -\det(AX^{-1}B)\operatorname{Tr}\Big(X^{-1}B(AX^{-1}B)^{-1}AX^{-1}\,\mathrm{d}X\Big)\\ &= -\det(AX^{-1}B)\Big\langle \Big(X^{-1}B(AX^{-1}B)^{-1}AX^{-1}\Big)^{\top}, \mathrm{d}X\Big\rangle.\end{aligned}$$

Тогда градиент

$$\nabla f(X) = -\det(AX^{-1}B)X^{-\top}A^{\top}(AX^{-1}B)^{-\top}B^{\top}X^{-\top}.$$

■

C2.3.3 Другие примеры

Помимо классических примеров, когда нужно продифференцировать функцию по вектору/матрице, проводя алгоритмические действия, бывают и менее тривиальные. В этом параграфе мы посвятим их рассмотрению немного времени.
В нейронных сетях часто встречается функция softmax, которая позволяет отобразить вектор из d координат в распределение вероятностей на d исходах:

$$\operatorname{softmax}(x) := \left(\frac{\exp(x_1)}{\sum_{i=1}^d \exp(x_i)}, \dots, \frac{\exp(x_d)}{\sum_{i=1}^d \exp(x_i)}\right)^{\top}.$$

Пример С3.13. Найдите матрицу Якоби J функции $s : \mathbb{R}^d \to \mathbb{R}^d$:

$$s(x) = \left(\frac{\exp(x_1)}{\sum_{i=1}^d \exp(x_i)}, \ldots, \frac{\exp(x_d)}{\sum_{i=1}^d \exp(x_i)} \right)^\top.$$

Решение. Считаем частные производные по определению

- При $k \neq j$

$$\frac{\partial s_k}{\partial x_j} = \frac{\partial}{\partial x_j} \frac{\exp(x_k)}{\sum_{i=1}^d \exp(x_i)} = -\frac{\exp(x_k)\exp(x_j)}{\left(\sum_{i=1}^d \exp(x_i)\right)^2} = -s_k \cdot s_j,$$

- при $k = j$

$$\frac{\partial s_j}{\partial x_j} = \frac{\partial}{\partial x_j} \frac{\exp(x_j)}{\sum_{i=1}^d \exp(x_i)} = \frac{\exp(x_j)(\sum_{i=1}^d \exp(x_i)) - \exp(x_j)\frac{\partial}{\partial x_j}(\sum_{i=1}^d \exp(x_i))}{(\sum_{i=1}^d \exp(x_i))^2}$$

$$= \frac{\exp(x_j)}{\sum_{i=1}^d \exp(x_i)} - \frac{\exp(x_j)\exp(x_j)}{(\sum_{i=1}^d \exp(x_i))^2} = s_j(1 - s_j).$$

Таким образом, получаем

$$J_{k,j} = \begin{cases} -s_k \cdot s_j, & k \neq j; \\ s_j(1 - s_j), & k = j. \end{cases}$$

■

Не менее часто на практике приходится применять какую-либо функцию к выходу линейного слоя покоординатно, то есть для каждого отдельно взятого нейрона.

Пример С3.14. Найдите $\nabla f(x)$ и $\nabla^2 f(x)$ функции $f : \mathbb{R}^d \to \mathbb{R}$:

$$f(x) = h(g(x)),$$

где функция $g : \mathbb{R} \to \mathbb{R}$ действует поэлементно:

$$g(x) = \sin(x),$$

а функция $h : \mathbb{R}^d \to \mathbb{R}$:

$$h(u) = \sum_{i=1}^d u_i.$$

Решение. Вспомним правило матрицы Якоби сложной функции:

$$J_f = J_h J_g \implies \nabla f = J_g^\top \nabla h.$$

Далее посчитаем матрицу Якоби функции вида $g(x) = \begin{pmatrix} g(x_1) \\ \vdots \\ g(x_d) \end{pmatrix}$. Получим

$$J_g = \operatorname{diag}\left(g'(x_1), \ldots, g'(x_d)\right) = \operatorname{diag}\left(g'(x)\right) = J_g^\top.$$

При умножении J_g на вектор удобно пользоваться поэлементным умножением матриц:

$$(A \odot B)_{ij} = A_{ij} \cdot B_{ij}.$$

Результат умножения J_g на вектор y равен

$$J_g y = \begin{pmatrix} g'(x_1) \\ \vdots \\ g'(x_d) \end{pmatrix} \odot y = g'(x) \odot y.$$

Заметим, что эта операция является довольно быстро вычислимой и легко поддаётся параллелизации. Теперь приступим непосредственно к примеру. Найдем J_g:

$$J_g = \operatorname{diag}\big(\cos(x_1), \ldots, \cos(x_d)\big) = \operatorname{diag}(\cos(x)).$$

Теперь найдем $\nabla h\big(u\big)$:

$$\nabla h\big(u\big) = \mathbf{1}.$$

Тогда $\nabla f\big(x\big)$:

$$\nabla f\big(x\big) = J_g^\top \nabla h\big(u\big) = \cos(x) \odot \mathbf{1} = \cos(x).$$

Выпишем гессиан функции:

$$\nabla^2 f\big(x\big) = J_{\nabla f}^\top = \operatorname{diag}\big(-\sin(x)\big).$$

■

Пример С3.15. Выразите $\phi'(\alpha)$ и $\phi''(\alpha)$ функции $\phi : \mathbb{R} \to \mathbb{R}$:

$$\phi(\alpha) = f\big(x + \alpha p\big),\ x, p \in \mathbb{R}^d$$

через $\nabla f, \nabla^2 f$ дважды непрерывно дифференцируемой функции $f : \mathbb{R}^d \to \mathbb{R}$.

Решение. Важно помнить, что дифференцирование происходит не по стандартному вектору x, а по скаляру α, со всеми вытекающими свойствами. Первый дифференциал

$$\mathrm{d}\phi = \big\langle \nabla f\big(x + \alpha p\big), \mathrm{d}(x + \alpha p)\big\rangle = \big\langle \nabla f\big(x + \alpha p\big), p\big\rangle \,\mathrm{d}\alpha\,.$$

Заметим, что мы привели дифференциал к стандартному виду $\mathrm{d}\phi = \phi'(\alpha)\,\mathrm{d}\alpha$, то есть множитель перед $\mathrm{d}\alpha$ — это производная:

$$\phi'(\alpha) = \big\langle \nabla f\big(x + \alpha p\big), p\big\rangle.$$

Дифференциал от первой производной:

$$\begin{aligned}
\mathrm{d}\big(\phi'(\alpha)\big) &= \mathrm{d}\big\langle \nabla f\big(x + \alpha p\big), p\big\rangle = \big\langle \big\langle \nabla^2 f\big(x + \alpha p\big), \mathrm{d}(x + \alpha p)\big\rangle, p\big\rangle \\
&= \Big\langle \big(\nabla^2 f\big(x + \alpha p\big)\big)^\top \mathrm{d}(x + p\alpha)\,, p\Big\rangle = \big\langle \nabla^2 f\big(x + \alpha p\big) p\,\mathrm{d}\alpha\,, p\big\rangle \\
&= \big\langle \nabla^2 f\big(x + \alpha p\big) p, p\big\rangle\,\mathrm{d}\alpha\,.
\end{aligned}$$

Тогда вторая производная:

$$\phi''(\alpha) = \big\langle \nabla^2 f\big(x + \alpha p\big) p, p\big\rangle.$$

■

Теперь посмотрим на случай, когда функция переводит матрицу в матрицу, и записать производные в компактном виде через матрицу или вектор не получается.

Пример С3.16. Найдите $\nabla f(X)$ функции $f : \mathbb{R}^{n\times d} \to \mathbb{R}^n$:

$$f(X) = Xa,\ a \in \mathbb{R}^d.$$

Решение. Дифференциал найти достаточно просто:

$$\mathrm{d}f(X) = \mathrm{d}X\, a.$$

Но вот записать производную в виде вектора или матрицы уже не выйдет — нужен тензор третьего порядка, имеющий вид

$$\frac{\partial f_k}{\partial X_{ij}}(X)$$

размерности $n \times n \times d$. Для этого выразим скалярную зависимость функции $f(X)$:

$$f_k(X) = \sum_{l=1}^{d} X_{kl} a_l.$$

Теперь возьмём производную $\frac{\partial}{\partial X_{ij}}$. Получим

$$\frac{\partial f_k}{\partial X_{ij}}(X) = \sum_{l=1}^{d} \frac{\partial (X_{kl} a_l)}{\partial X_{ij}} = \sum_{l=1}^{d} \delta(i=k, j=l) a_l = \delta(i=k) \sum_{l=1}^{d} \delta(j=l) a_l = \delta(i=k) a_j.$$

■

Пример С3.17. Рассмотрим прикладную задачу. Пусть имеется модель, которая для некоторого значения x возвращает y по правилу: $y = Ax + \varepsilon$, где $A \in \mathbb{R}^{n\times d}$ — некоторая матрица, а ε — некоторый шум. Чтобы не загромождать повествование техническими деталями, будем предполагать, что $\det(A^\top A) \neq 0$. Имея дело с такой моделью, мы хотим решать задачу:

$$\hat{x} = \operatorname*{argmin}_{x:\hat{y}=Ax+\varepsilon} \|\varepsilon\|_2^2.$$

Мы хотим определять, какой вход модели был наиболее вероятен, если она выдала значение $\hat{y}$. Тогда, выразив $\xi = \hat{y} - Ax$, будем искать минимум функции $\|\hat{y} - Ax\|_2^2$. Для этого нужно посчитать градиент и приравнять его к 0. Запишем

$$\begin{aligned}\mathrm{d}\|\hat{y} - Ax\|_2^2 &= \mathrm{d}\langle \hat{y} - Ax, \hat{y} - Ax\rangle = -\langle A\,\mathrm{d}x\,, \hat{y} - Ax\rangle - \langle \hat{y} - Ax, A\,\mathrm{d}x\rangle \\ &= -2\langle \hat{y} - Ax, A\,\mathrm{d}x\rangle = -2\langle A^\top(\hat{y} - Ax), \mathrm{d}x\rangle.\end{aligned}$$

Отсюда получаем, что градиент равен $\nabla f(x) = -2A^\top(\hat{y} - Ax)$. Приравнивая его к 0, имеем

$$\hat{x} = (A^\top A)^{-1} A^\top \hat{y}.$$

C2.4 Автоматическое дифференцирование

C2.4.1 Граф вычислений

Настало время поговорить о том, как происходит подсчёт градиентов в реальной жизни. Чаще всего функции, с которыми приходится иметь дело на практике представляют собой последовательность (дифференцируемых) параметрических преобразований. Таким образом, их можно представить в виде *вычислительного графа*, где промежуточным вершинам соответствуют преобразования, входящим стрелкам — входные переменные, а выходным стрелкам — результат преобразования. Этот граф должен быть ациклическим.

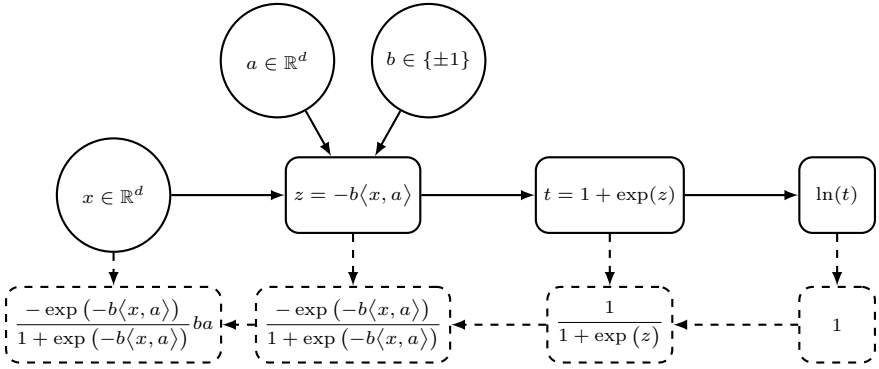


Рис. C3.1: Граф вычислений для логистической регрессии

На Рисунке C3.1 сплошными линиями приведён вычислительный граф для логистической регрессии

$$f(x) = \ln\Big(1 + \exp\Big(-bx^\top a\Big)\Big),$$

где $a \in \mathbb{R}^d$, $b \in \{-1, 1\}$ — вход, а $x \in \mathbb{R}^d$ — параметр.
Вычисление значения функции по заданному входу называется *forward pass (прямым проходом)*. На этом этапе происходит преобразование исходного вектора в целевой. Последовательно строятся промежуточные значения — результаты применения преобразований к предыдущим значениям слева направо. В случае линейного графа можно записать его формулой

$$f(x) = u_n(u_{n-1}(\dots(u_1(x))). \tag{C3.1}$$

Но как же считать производные в таких графах? Ответ может дать правило вычисления дифференциала сложной функции. Рассмотрим одну конкретную вершину в графе вида $u(x_1, \dots, x_d)$ и её дифференциал

$$\mathrm{d}u = J_u \,\mathrm{d}x\,,$$

причём каждый следующий дифференциал $\mathrm{d}x_i$ можно расписать через предыдущую вершину (если, конечно, x_i не являются искомыми переменными), двигаясь по рекурсии в графе от детей к родителям. В случае линейного графа (C3.1) получим итоговую формулу

$$\frac{\partial f}{\partial x}(x) = \frac{\partial u_1}{\partial x}(x) \cdot \frac{\partial u_2}{\partial u_1}(u_1) \cdot \ldots \cdot \underbrace{\frac{\partial u_n}{\partial u_{n-1}}}_{\frac{\partial f}{\partial u_{n-1}}}(u_{n-1}). \tag{C3.2}$$

C2.4.2 Backward propagation

Основная идея *backward propagation (обратного распространения ошибки)* заключается в подсчёте формулы (C3.2) **слева направо**.
В общем случае, пусть $u_1, \ldots, u_n$ — вершины графа вычислений в топологическом порядке (т.е. родители идут перед потомками). Тогда общий алгоритм действий выглядит так

1. Произвести forward pass и сохранить все значения u_i как функции от их родителей;

2. Воспользуемся производной сложной функции:

$$J_{f(u)} = J_f J_u.$$

   Или, что тоже самое, что для всех $i = \overline{n-1, 1}$ посчитать

$$\frac{\partial f}{\partial u_i} = \sum_{j \in \text{потомки}(u_i)} \frac{\partial u_j}{\partial u_i} \frac{\partial f}{\partial u_j}.$$

Вычисление backward propagation на Рисунке C3.1 показано пунктирными линиями.

Пример C3.18. Посчитайте backward propagation для графа, изображенного на Рисунке C3.2, где вход $x \in \mathbb{R}^d$, параметры $\omega \in \mathbb{R}^d$, $b \in \mathbb{R}$, а $\lambda, t \in \mathbb{R}$ — фиксированные константы.

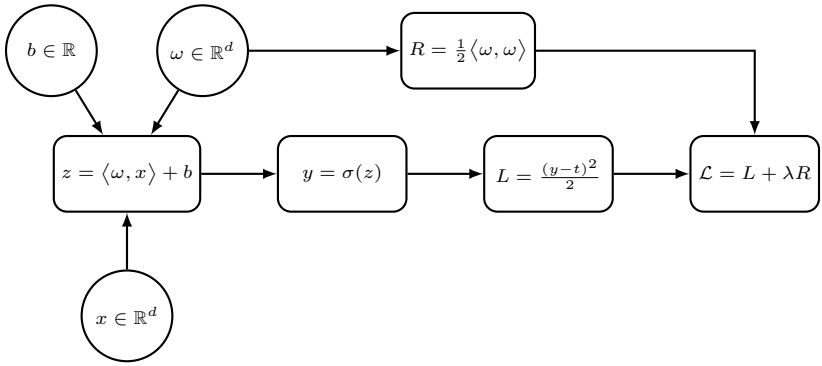


Рис. C3.2: Граф вычислений: логистическая регрессия с MSE и L_2-регуляризацией.

Решение.

$$\begin{aligned}
\frac{\partial \mathcal{L}}{\partial R} &= \frac{\partial \mathcal{L}}{\partial R}\frac{\partial \mathcal{L}}{\partial \mathcal{L}} = \lambda,\\
\frac{\partial \mathcal{L}}{\partial L} &= \frac{\partial \mathcal{L}}{\partial L}\frac{\partial \mathcal{L}}{\partial \mathcal{L}} = 1,\\
\frac{\partial \mathcal{L}}{\partial y} &= \frac{\partial L}{\partial y}\frac{\partial \mathcal{L}}{\partial L} = (y-t)\frac{\partial \mathcal{L}}{\partial L},\\
\frac{\partial \mathcal{L}}{\partial z} &= \frac{\partial y}{\partial z}\frac{\partial \mathcal{L}}{\partial y} = \sigma'(z)\frac{\partial \mathcal{L}}{\partial y},\\
\frac{\partial \mathcal{L}}{\partial \omega} &= \frac{\partial z}{\partial \omega}\frac{\partial \mathcal{L}}{\partial z} + \frac{\partial R}{\partial \omega}\frac{\partial \mathcal{L}}{\partial R} = x\frac{\partial \mathcal{L}}{\partial z} + \omega\frac{\partial \mathcal{L}}{\partial R},\\
\frac{\partial \mathcal{L}}{\partial b} &= \frac{\partial z}{\partial b}\frac{\partial \mathcal{L}}{\partial z} = \frac{\partial \mathcal{L}}{\partial z}.
\end{aligned}$$

■

Это достаточно простой пример, в то время как на практике в основном встречаются отображения, действующие в пространствах матриц и векторов. Для лучшего понимания вычислим backward propagation в полносвязной нейронной сети.

Пример С3.19. Рассмотрим полносвязную нейронную сеть для задачи бинарной классификации с кросс-энтропийным лоссом:

$$\begin{aligned}
\hat{y} &= \sigma(\sigma(XW_1)W_2),\\
\mathcal{L}(\hat{y}) &= -\sum_{i=1}^{n}\big(y_i \ln \hat{y}_i + (1-y_i)\ln(1-\hat{y}_i)\big),
\end{aligned}$$

где вход $X \in \mathbb{R}^{n\times d}$, $y \in \{0,1\}^n$, параметры $W_1 \in \mathbb{R}^{d\times k}$, $W_2 \in \mathbb{R}^k$.

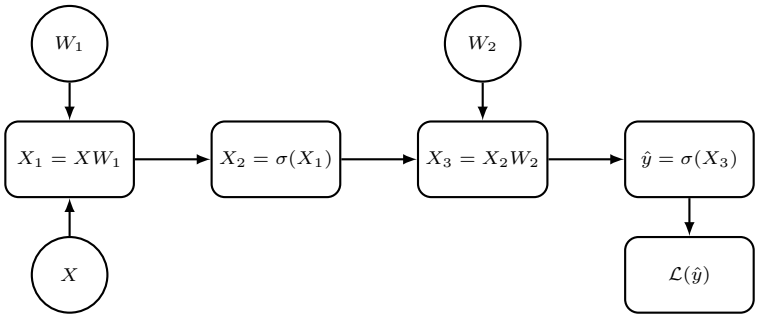


Рис. С3.3: Прямой проход полносвязной нейросети с кросс-энтропийным лоссом

Найдём производные по параметрам W_1, W_2. Для удобства записи введём обозначения выходов промежуточных слоёв:

$$X_1 = XW_1,\ X_2 = \sigma(X_1),\ X_3 = X_2W_2.$$

1. Для начала найдём градиент лосса по $\hat{y}$ (все произведения покомпонентные):

$$\mathrm{d}\mathcal{L}(\hat{y}) = -\sum_{i=1}^{n} \mathrm{d}(y_i \ln \hat{y}_i + (1-y_i)\ln(1-\hat{y}_i)) = \sum_{i=1}^{n} \left(-\frac{y_i}{\hat{y}_i}\,\mathrm{d}\hat{y}_i + \frac{1-y_i}{1-\hat{y}_i}\,\mathrm{d}\hat{y}_i \right)$$
$$= \sum_{i=1}^{n} \frac{\hat{y}_i - y_i}{\hat{y}_i(1-\hat{y}_i)}\,\mathrm{d}\hat{y}_i\,.$$

   Тогда получаем:

$$\frac{\partial \mathcal{L}}{\partial \hat{y}_i} = \frac{\hat{y}_i - y_i}{\hat{y}_i(1-\hat{y}_i)}.$$

2. Воспользуемся Примером С3.14 для нахождения производной по X_3, положив g — кросс-энтропией, а h — сигмоидой:

$$\frac{\partial \mathcal{L}}{\partial X_3} = \sigma'(X_3) \odot \frac{\partial \mathcal{L}}{\partial \hat{y}} = \sigma(X_3)(1-\sigma(X_3)) \odot \frac{\partial \mathcal{L}}{\partial \hat{y}}.$$

3. Отсюда уже можно найти $\frac{\partial \mathcal{L}}{\partial W_2}$:

$$\frac{\partial \mathcal{L}}{\partial W_2} = \frac{\partial X_3}{\partial W_2}\frac{\partial \mathcal{L}}{\partial X_3} = X_2^\top \frac{\partial \mathcal{L}}{\partial X_3}.$$

4. Аналогично можно найти $\frac{\partial \mathcal{L}}{\partial X_2}$:

$$\frac{\partial \mathcal{L}}{\partial X_2} = \frac{\partial X_3}{\partial X_2}\frac{\partial \mathcal{L}}{\partial X_3} = \frac{\partial \mathcal{L}}{\partial X_3} W_2^\top.$$

5. Снова ищем градиент для композиции с сигмоидой:

$$\frac{\partial \mathcal{L}}{\partial X_1} = \sigma'(X_1) \odot \frac{\partial \mathcal{L}}{\partial X_2} = \sigma(X_1)(1-\sigma(X_1)) \odot \frac{\partial \mathcal{L}}{\partial X_2}.$$

6. Последний шаг проделываем по аналогии с 3:

$$\frac{\partial \mathcal{L}}{\partial W_1} = \frac{\partial X_1}{\partial W_1}\frac{\partial \mathcal{L}}{\partial X_1} = X^\top \frac{\partial \mathcal{L}}{\partial X_1}.$$

- Для подсчёта backward propagation необходимо хранить все промежуточные значения u_i на всех итерациях алгоритма. Это может быть существенным требованием к памяти, например, в больших нейронных сетях.

- Необязательно полностью считать производную $\frac{\partial u_j}{\partial u_i}$, важно уметь быстро превращать градиент по выходу в градиент по входу. Об этом будет рассказано в секции С2.4.4.

- Отдельно стоит отметить нейронные сети, которые состоят из блоков. Одному блоку совершенно не надо знать, что происходит вокруг. То есть он может быть запрограммирован как отдельная сущность, умеющая внутри себя делать forward pass и backward propagation, после чего блоки механически, как кубики в конструкторе, собираются в большую сеть, которая работает как одно целое.

- Вычисление градиента можно представить через ещё один граф вычислений, тем самым, сделав уже backward propagation по графу градиента. Так можно считать гессианы и производные высших порядков.

C2.4.3 Forward propagation

Может возникнуть желание посчитать формулу (C3.2) не слева направо, а **справа налево**, распространяя производную в графе в направлении forward pass: от родителей к детям. Так производная по параметрам x будет считаться как

$$\frac{\partial u_i}{\partial x} = \sum_{u_j \in \text{родители}(u_i)} \frac{\partial u_j}{\partial x} \frac{\partial u_i}{\partial u_j}.$$

При этом можно совершать проход вместе с подсчётом u_i и нужно будет хранить только один слой графа.

Пример C3.20. Посчитайте forward propagation для графа из Примера C3.18.

Решение.

$$\begin{aligned}
\frac{\partial R}{\partial \omega} &= \omega, & \frac{\partial R}{\partial b} &= 0,\\
\frac{\partial z}{\partial \omega} &= x, & \frac{\partial z}{\partial b} &= 1,\\
\frac{\partial y}{\partial \omega} &= \frac{\partial z}{\partial \omega}\sigma'(z), & \frac{\partial y}{\partial b} &= \frac{\partial z}{\partial b}\sigma'(z),\\
\frac{\partial L}{\partial \omega} &= \frac{\partial y}{\partial \omega}\cdot(y-t) + \lambda\frac{\partial R}{\partial \omega}. & \frac{\partial L}{\partial b} &= \frac{\partial y}{\partial \omega}\cdot(y-t).
\end{aligned}$$

■

- У forward propagation есть нюанс — нужно хранить производные $\frac{\partial u_i}{\partial x}$, а в backward propagation — $\frac{\partial f}{\partial u_i}$.

- Если размерность входа x намного больше, чем размерность выхода $f\big(x\big)$, то на каждом шаге forward propagation нужно будет хранить и обсчитывать больше данных, чем в backward propagation.

- Если, наоборот, размерность выхода функции $f\big(x\big)$ намного больше, чем размерность входа x, то выгоднее использовать forward propagation.

C2.4.4 Умножение гессиана на вектор и якобиана на строку

Пусть дана дважды дифференцируемая функция $f : \mathbb{R}^d \to \mathbb{R}$. Покажем, как можно эффективно считать

$$\nabla^2 f\big(x\big) \cdot u$$

для любого вектора $u \in \mathbb{R}^d$.
Считать полный гессиан и умножать его на вектор неэффективно особенно при большой размерности d. Но заметим, что

$$\mathrm{d}\big\langle u, \nabla f\big(x\big)\big\rangle = \big\langle u, J_{\nabla f}\,\mathrm{d}x\big\rangle = \big\langle \nabla^2 f\big(x\big)\cdot u, \mathrm{d}x\big\rangle = \big\langle \nabla g\big(x\big), \mathrm{d}x\big\rangle,$$

где $g\big(x\big) = \big\langle \nabla f\big(x\big), u\big\rangle$. Таким образом вместо полного гессиана можно найти градиент функции вида $g\big(x\big)$, с которым могут справиться библиотеки автоматического дифференцирования jax/autograd/pytorch/tensorflow.

Аналогично для функции вида $f : \mathbb{R}^d \to \mathbb{R}^n$ и вектора $u \in \mathbb{R}^n$ считается значение $\nabla f\big(x\big) \cdot u$:

$$\mathrm{d}\big\langle u, f\big(x\big)\big\rangle = \big\langle u, J_f\,\mathrm{d}x\big\rangle = \big\langle \nabla f\big(x\big) \cdot u, \mathrm{d}x\big\rangle = \big\langle \nabla g\big(x\big), \mathrm{d}x\big\rangle,$$

где $g\big(x\big) = \big\langle f\big(x\big), u\big\rangle$.
Именно поэтому backward propagation можно эффективно реализовать на практике.

C4 Классы множеств

Теоретический анализ широкого множества методов машинного обучения задействует предположения, которые можно ввести только на множествах, обладающих «хорошими» геометрическими свойствами. Неформально говоря, требуется, чтобы рассматриваемое множество, из которого выбираются параметры, не имело «впадин» или разрывов. Формальное описание такого множества дает понятие выпуклости.

C4.1 Выпуклые множества

Пусть X — линейное пространство.

Определение C4.1. Множество $S \subseteq X$ называется *выпуклым*, если для любых двух точек $x_1,\ x_2 \in S$ и любого числа $\alpha \in [0, 1]$, точка $\alpha x_1 + (1 - \alpha)x_2$ также принадлежит S.

Интуитивно, это означает, что отрезок, соединяющий любые две точки множества, полностью лежит во множестве.

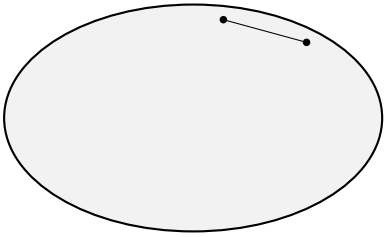


Рис. C4.1: Выпуклое множество.

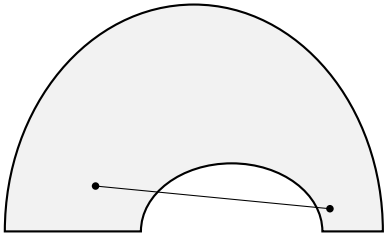


Рис. C4.2: Невыпуклое множество.

Пример C4.1. Пусть $a \in \mathbb{R}^d \setminus \{0\}$, $b \in \mathbb{R}$. Покажите, что полуплоскость

$$S(a, b) = \left\{ x \;\middle|\; a^\top x \geq b \right\}$$

является выпуклым множеством.

Решение. Пусть $x_1, x_2 \in S(a, b)$, тогда $a^\top x_1 \geq b$ и $a^\top x_2 \geq b$. Покажем, что для любого $\alpha \in [0, 1]$ выполняется $a^\top(\alpha x_1 + (1 - \alpha)x_2) \geq b$:

$$a^\top(\alpha x_1 + (1 - \alpha)x_2) = \alpha a^\top x_1 + (1 - \alpha)a^\top x_2 \geq \alpha b + (1 - \alpha)b = b.$$

Следовательно, всякий отрезок между x_1 и x_2 также принадлежит полуплоскости. ■

Пример C4.2. Пусть $a, b \in \mathbb{R}^d$ — различные точки. Покажите, что полуплоскость Вороного, то есть множество всех точек, которые ближе к a, чем к b в евклидовой норме:

$$V(a, b) = \left\{ x \in \mathbb{R}^d \;\middle|\; \|x - a\|_2 \leq \|x - b\|_2 \right\}$$

является выпуклым множеством.

Решение. Перепишем множество $V(a,b)$:

$$\begin{aligned}\|x-a\|_2^2 \le \|x-b\|_2^2 &\iff (x-a)^\top (x-a) \le (x-b)^\top (x-b)\\ &\iff x^\top x - 2a^\top x + a^\top a \le x^\top x - 2b^\top x + b^\top b\\ &\iff 2(b-a)^\top x \le b^\top b - a^\top a.\end{aligned}$$

Таким образом, множество $V(a,b)$ является полуплоскостью, а то, что полуплоскость — выпуклое множество было показано в Примере С4.1. ■

Пример С4.3. Пусть $\|\cdot\|$ — норма в $\mathbb{R}^d$, $r>0$ и $c\in\mathbb{R}^d$. Покажите, что шар

$$\overline{B}(c,r) = \left\{ x\in\mathbb{R}^d \;\middle|\; \|x-c\| \le r \right\}$$

является выпуклым множеством.

Решение. Пусть $x_1, x_2 \in \overline{B}(c,r)$. Тогда

$$\begin{aligned}\|\alpha x_1 + (1-\alpha)x_2 - c\| &= \|\alpha(x_1-c) + (1-\alpha)(x_2-c)\|\\ &\le \alpha\|x_1-c\| + (1-\alpha)\|x_2-c\|\\ &\le \alpha r + (1-\alpha)r = r.\end{aligned}$$

Следовательно, $\alpha x_1 + (1-\alpha)x_2 \in \overline{B}(c,r)$. ■

Обратим внимание, что, хоть этот пример и был рассмотрен для случая векторных пространств, те же рассуждения можно провести и для случая матричных норм, поскольку для них выполнены те же неравенства.

Пример С4.4. Пусть $\|\cdot\|$ — норма в $\mathbb{R}^d$, $r>0$ и $c\in\mathbb{R}^d$. Покажите, что сфера

$$S(c,r) = \left\{ x\in\mathbb{R}^d \;\middle|\; \|x-c\| = r \right\}$$

не является выпуклым множеством.

Решение. Возьмем две диаметрально противоположные точки $x+c$ и $-x+c$ такие, что $\|x\| = r$. Тогда c является их средним арифметическим, но он не принадлежит S. Получили противоречие. ■

Пример С4.5. Покажите, что множество всех положительно полуопределенных матриц $\mathbb{S}^d_+$ является выпуклым.

Решение. Пусть $X_1, X_2 \in \mathbb{S}^d_+$ и $\alpha\in[0,1]$, тогда матрица $\alpha X_1 + (1-\alpha)X_2$ принадлежит $\mathbb{S}^d_+$. Покажем это, возьмем произвольный $z\in\mathbb{R}^d$:

$$z^\top(\alpha X_1 + (1-\alpha)X_2)z = \alpha z^\top X_1 z + (1-\alpha) z^\top X_2 z \ge 0.$$

■

Пример С4.6. Покажите, что множество матриц 100×100 ранга хотя бы 2, у которых все элементы положительны является не выпуклым.

Решение. Пусть A — верхнетреугольная матрица, у которой все элементы на диаго-

нали и выше равны 1, а $B = 11^\top - A$ — дополняющая ее нижнетреугольная матрица. Легко заметить, что у матриц A и B ранг хотя бы 2, но у матрицы $\frac{1}{2}A + \frac{1}{2}B = \frac{1}{2}11^\top$ ранг 1. ■

Пример C4.7. Покажите, что множество

$$M = \left\{ x \in \mathbb{R}^2_{++} \;\middle|\; x_1 x_2 \geq 1 \right\}$$

является выпуклым.

Решение. Пусть $x, y \in M$. Рассмотрим $\alpha x + (1-\alpha)y$, где $\alpha \in [0,1]$:

$$\begin{aligned}(\alpha x_1 + (1-\alpha)y_1)(\alpha x_2 + (1-\alpha)y_2) &= \alpha^2 x_1 x_2 + \alpha(1-\alpha)(x_1 y_2 + x_2 y_1) + (1-\alpha)^2 y_1 y_2 \\ &\geq \alpha^2 + 2\alpha(1-\alpha)\sqrt{x_1 x_2 y_1 y_2} + (1-\alpha)^2 \geq 1.\end{aligned}$$

Таким образом, точка $\alpha x + (1-\alpha)y \in M$, значит M выпукло. ■

Пример C4.8. Покажите, что множество

$$M = \left\{ x \in \mathbb{R}^2_{++} \;\middle|\; \sqrt{x_1} + x_2 \geq 1 \right\}$$

является выпуклым.

Решение. Мы будем использовать следующий несложный факт: если $z_1, z_2 > 0$ и $\alpha \in [0,1]$, то

$$\sqrt{\alpha z_1 + (1-\alpha)z_2} \geq \alpha\sqrt{z_1} + (1-\alpha)\sqrt{z_2}.$$

Вообще говоря, это неравенство Йенсена для вогнутой функции, но оно будет пройдено позже, поэтому его можно просто проверить возведением обеих частей неравенства в квадрат.
Пусть $x, y \in M$. Рассмотрим $\alpha x + (1-\alpha)y$, где $\alpha \in [0,1]$:

$$\begin{aligned}\sqrt{\alpha x_1 + (1-\alpha)y_1} + \alpha x_2 + (1-\alpha)y_2 &\geq \alpha\sqrt{x_1} + (1-\alpha)\sqrt{y_1} + \alpha x_2 + (1-\alpha)y_2 \\ &\geq \alpha + (1-\alpha) = 1.\end{aligned}$$

Таким образом, точка $\alpha x + (1-\alpha)y \in M$, значит M выпукло. ■

Определение позволяет проверять выпуклость некоторых наиболее простых множеств. В случае, когда множество имеет более сложную структуру, пользоваться определением может быть затруднительно. В связи с этим, требуется расширить математический аппарат анализа выпуклых множеств.

Утверждение C4.1. Пусть $S \subseteq X$ — выпуклое множество. Тогда $\operatorname{int} S$ и $\operatorname{cl} S$ также являются выпуклыми множествами.

Доказательство.

- Внутренность S.

  Пусть $x, y \in \operatorname{int} S$. По определению внутренней точки существуют радиусы r_x и r_y такие, что открытые шары $B(x, r_x)$ и $B(y, r_y)$ полностью содержатся в S. Рассмотрим отрезок, соединяющий x и y. Для любой точки $z \in [x, y]$, можно построить такой шар $B(z, r_z)$, который полностью лежит в S. Действительно, поскольку

шары $B(x, r_x)$, $B(y, r_y)$ содержатся в выпуклом множестве S, то отрезки, соединяющие их точки, также лежат в S. То есть можно выбрать $r_z = \min\{r_x, r_y\}$ и гарантировать принадлежность S шара с таким радиусом. Таким образом, все точки отрезка $[x, y]$ являются внутренними точками S, поэтому $\operatorname{int} S$ является выпуклым множеством.

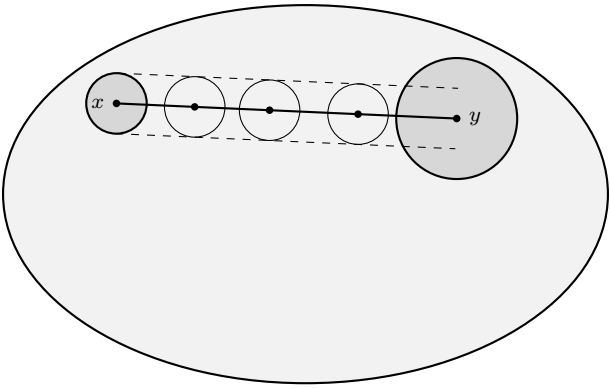


Рис. С4.3: Иллюстрация выпуклости внутренности множества.

- Замыкание S.

  Пусть $x, y \in \operatorname{cl} S$. Тогда найдутся последовательности $\{x_k\}_{k=1}^{\infty} \subseteq S$ и $\{y_k\}_{k=1}^{\infty} \subseteq S$ такие, что $x_k \to x$ и $y_k \to y$. Таким образом, для любого $\alpha \in [0, 1]$ последовательность $\{\alpha x_k + (1-\alpha) y_k\}_{k=1}^{\infty} \subseteq S$ и $\alpha x_k + (1-\alpha) y_k \to \alpha x + (1-\alpha) y \in \operatorname{cl} S$.

■

Приведем пример, иллюстрирующий это утверждение.

Пример С4.9. Множество положительно определенных матриц $\mathbb{S}^d_{++}$ является выпуклым.

Доказательство. Ранее мы показали выпуклость $\mathbb{S}^d_{+}$. Теперь рассмотрим $A \in \mathbb{S}^d_{++}$. Известно, что собственные числа матрицы A строго положительны:

$$\lambda_i(A) > 0, \ \forall i \in [1, d].$$

Зафиксируем возмущение $E \in \mathbb{R}^{d \times d}$. Рассмотрим действие квадратичной формы $A + tE$ на нормированный вектор $x \in \mathbb{R}^d$:

$$x^\top (A + tE) x = x^\top A x + t x^\top E x \geq \lambda_{min}(A) + t\lambda_{min}(E).$$

Поскольку действительные числа отделимы, можно подобрать такую t, что правая часть остается положительной. Таким образом, $\mathbb{S}_{++} \subseteq \operatorname{int} \mathbb{S}_{+}$. Рассмотрим теперь $B \in \mathbb{S}_{+}$, такую, что $\lambda_{min}(B) = 0$. Любое сколь угодно малое возмущение против направления соответствующего собственного вектора делает минимальное собственное число отрицательным. Это доказывает обратное включение. Тогда $\mathbb{S}^d_{+}$ выпукло как внутренность выпуклого множества. ■

Отметим, что выпуклость $\mathbb{S}^d_{++}$ может быть проверена значительно проще по определению и пример выше следует воспринимать, как искусственный.

C4.2 Размерность выпуклых множеств

Возникает вопрос, как определить размерность выпуклого множества. Сделать это формально помогает понятие аффинной оболочки множества.

Определение C4.2. *Аффинной оболочкой* множества $S \subseteq X$ будем называть множество вида

$$\operatorname{aff} S = \bigcup_{k\in\mathbb{N}} \left\{ \sum_{i=1}^{k} \alpha_i x_i \;\middle|\; \sum_{i=1}^{k} \alpha_i = 1,\ x_i \in S \right\}.$$

Афинная оболочка имеет «хорошую» геометрию. Сформулируем это в виде теоремы.

Теорема C4.1. Аффинная оболочка $\operatorname{aff} S$ множества $S \subseteq X$ имеет вид

$$\operatorname{aff} S = s + L_S,$$

где $s \in \operatorname{aff} S$ — произвольная точка из аффинной оболочки; L_S — линейное подпространство, причем L_S — единственное.

Доказательство. Зафиксируем $s \in \operatorname{aff} S$. Положим

$$L_S = \left\{ x \in \mathbb{R}^d \;\middle|\; x = y - s,\ y \in \operatorname{aff} S \right\}.$$

Докажем, что L_S — действительно линейное подпространство. Пусть $x \in L_S$, тогда $x = y - s$, где $y \in \operatorname{aff} S$. По определению аффинной оболочки, имеем $\lambda y + (1-\lambda)s \in \operatorname{aff} S$. Тогда

$$\lambda x = \lambda(y - s) = \lambda y + (1 - \lambda)s - s \in L_S.$$

Это доказывает замкнутость относительно операции умножения. Рассмотрим теперь $x_1 = y_1 - s,\ x_2 = y_2 - s \in L_S$, где $y_1, y_2 \in \operatorname{aff} S$. Так как

$$\frac{y_1 + y_2}{2} \in \operatorname{aff} S,$$

то

$$\frac{x_1 + x_2}{2} = \frac{y_1 + y_2}{2} - s \in L_S.$$

Домножив равенство на 2, получаем замкнутость L_S относительно операции сложения. Остается показать единственность. Выберем другую точку $\widetilde{s} \in \operatorname{aff} S$ и определим линейное пространство M_S аналогично L_S. Имеем

$$\widetilde{s} + L_S = s + (\widetilde{s} - s) + L_S = s + L_S = \operatorname{aff} S.$$

Таким образом, $L_S = M_S$. ■

Итак, мы определили аффинную оболочку точек, которая является сдвинутым линейным подпространством. Естественно определить $\dim S = \dim \operatorname{aff} S = \dim L_S$.

C4.3 Выпуклая оболочка

Определение С4.3. *Выпуклой комбинацией* точек $x_1, \ldots, x_k \in X$ называется любая точка вида

$$\sum_{i=1}^{k} \alpha_i x_i,$$

где $\sum_{i=1}^{k} \alpha_i = 1$, $\alpha_i \geq 0$.

Утверждение С4.2. Если $S \subseteq X$ является выпуклым множеством и $x_1, \ldots, x_k \in S$, то любая выпуклая комбинация точек $x_1, \ldots, x_k$ также принадлежит S.

Решение. Доказательство проведем индукцией по k.

- База при $k = 2$ верна, поскольку тогда это определение выпуклого множества.
- Пусть утверждение верно для $k-1$. Рассмотрим $x_1, \ldots, x_k \in S$, и пусть $\alpha_1, \ldots, \alpha_k$ таковы, что $\alpha_1 + \ldots + \alpha_k = 1$ и $0 \leq \alpha_i \leq 1$. Если $\alpha_k = 1$, то утверждение очевидно, поскольку тогда имеем комбинацию из одной точки. В противном случае, перепишем комбинацию:

  $$\alpha_1 x_1 + \ldots + \alpha_k x_k = (1 - \alpha_k)\left(\frac{\alpha_1}{1 - \alpha_k} x_1 + \ldots + \frac{\alpha_{k-1}}{1 - \alpha_k} x_{k-1}\right) + \alpha_k x_k,$$

  где каждое слагаемое $\frac{\alpha_i}{1-\alpha_k}$ лежит в интервале $[0, 1]$, а их сумма равна 1. По предположению индукции,

  $$\frac{\alpha_1}{1 - \alpha_k} x_1 + \ldots + \frac{\alpha_{k-1}}{1 - \alpha_k} x_{k-1}$$

  принадлежит S. Комбинация из двух точек принадлежит S по определению выпуклого множества.

■

Определение С4.4. Выпуклой оболочкой множества $S \subseteq X$ будем называть множество всех выпуклых комбинаций элементов S:

$$\operatorname{conv} S = \bigcup_{k \in \mathbb{N}} \left\{ \sum_{i=1}^{k} \alpha_i x_i \;\middle|\; \sum_{i=1}^{k} \alpha_i = 1,\ \alpha_i \geq 0,\ x_i \in S \right\}.$$

Заметим, что так, как $\operatorname{conv} S \subseteq \operatorname{aff} S$, то $\dim S = \dim \operatorname{conv} S = \dim \operatorname{aff} S = \dim L_S$.

Теорема С4.2. Выпуклая оболочка $S \subseteq X$ является пересечением всех выпуклых множеств, содержащих S.

Доказательство. Пусть T — минимальное по включению выпуклое множество, содержащее S.

$\circledcirc$ Произвольная выпуклая комбинация элементов S, также является выпуклой комбинацией выпуклого множества T. Поэтому $\operatorname{conv} S \subseteq \operatorname{conv} T \subseteq T$.

$\circledcirc$ Заметим, что $\operatorname{conv} S$ является выпуклым множеством, так как если x и y выпуклые комбинации S, то $\alpha x + (1-\alpha)y$ является выпуклой комбинацией большей размерности для $\alpha \in [0, 1]$. Отсюда следует, что $T \subseteq \operatorname{conv} S$. ■

Замечание C4.1. Несложно проверить, что пересечение любого семейства выпуклых множеств также является выпуклым. Действительно, пусть $x, y \in \bigcap_{i \in I} S_i$, тогда $\alpha x + (1-\alpha)y \in S_i \ \forall i \in I$, значит эта точка лежит в пересечении.

Утверждение C4.3. Пусть $S, T \subseteq X$. Тогда верны следующие утверждения:

1. $S \subseteq \operatorname{conv} S$;
2. $S \subseteq T \to \operatorname{conv} S \subseteq \operatorname{conv} T$;
3. S является выпуклым тогда и только тогда, когда $S = \operatorname{conv} S$.

Доказательство. Выше мы выяснили, что выпуклая оболочка S — пересечение всех выпуклых множеств, содержащих S. Очевидно, что пересечение также содержит S. Это рассуждение доказывает первый пункт. Второй пункт следует из того, что $\operatorname{conv} T$ является пересечением меньшего количества множеств, чем $\operatorname{conv} S$. Третий пункт следует из первого пункта и того, что $\operatorname{conv} S$ — наименьшее выпуклое множество, содержащее S, но $S \subseteq S$ и S — выпукло. ■

Пример C4.10. Докажите, что

$$\operatorname{conv}\left\{ xx^\top,\ x \in \mathbb{R}^d \;\middle|\; \|x\|_2 = 1 \right\} = \left\{ A \in \mathbb{S}^d_+ \;\middle|\; \operatorname{Tr}(A) = 1 \right\}.$$

Доказательство.

$\circledast$ Рассмотрим $x \in \mathbb{R}^d : \|x\|_2 = 1$. Покажем, что след матрицы $xx^\top$ равен 1:

$$\operatorname{Tr}\left(xx^\top\right) = \operatorname{Tr}\left(x^\top x\right) = \|x\|_2^2 = 1.$$

Теперь рассмотрим матрицу $A \in \operatorname{conv}\left\{ xx^\top \mid x \in \mathbb{R}^d, \|x\|_2 = 1 \right\}$. Поскольку выпуклая оболочка — есть множество всевозможных выпуклых комбинаций, имеем

$$A = \alpha_1 x_1 x_1^\top + \alpha_2 x_2 x_2^\top + \ldots + \alpha_k x_k x_k^\top,$$

где $\alpha_1 + \alpha_2 + \ldots + \alpha_k = 1$, $0 \leq \alpha_i \leq 1$ и $\|x_i\|_2 = 1$. Поэтому, используя линейность следа, получим $\operatorname{Tr}(A) = 1$. Более того, она положительно полуопределённая в силу положительной полуопределенности каждого из слагаемых и неотрицательности α_i.

$\circledast$ Пусть $A \in \mathbb{S}^d_+$ и $\operatorname{Tr}(A) = 1$. Матрица A симметричная, значит у нее есть базис из собственных векторов. Применив спектральное разложение, получим

$$A = S^\top \operatorname{diag}(\lambda_1, \ldots, \lambda_d) S,$$

где S — ортогональная матрица. Заметим, что

$$\operatorname{Tr}\left(S^\top \operatorname{diag}(\lambda_1, \ldots, \lambda_d) S\right) = \operatorname{Tr}(\operatorname{diag}(\lambda_1, \ldots, \lambda_d)) = \lambda_1 + \lambda_2 + \ldots + \lambda_d = 1.$$

Из спектрального разложения можно сделать вывод, что

$$A = \lambda_1 s_1 s_1^\top + \lambda_2 s_2 s_2^\top + \ldots + \lambda_d s_d s_d^\top,$$

где s_i — соответствующие нормированные собственные вектора. Это завершает доказательство. ■

Теорема С4.3 (Каратеодори). Пусть $S \subseteq X$ и $\dim \operatorname{conv} S = d$. Тогда любой элемент $\operatorname{conv} S$ представляется как выпуклая комбинация не более чем $d+1$ точки множества S.

Доказательство. Будем доказывать от противного. Пусть имеется вектор $x \in \operatorname{conv} S$, такой что

$$x = \sum_{i=1}^{n} \alpha_i x_i,\ \sum_{i=1}^{n} \alpha_i = 1,\ \alpha_i \geq 0,\ n \geq d+2.$$

$x_2 - x_1, \ldots, x_n - x_1$ принадлежат линейному пространству L_S из Теоремы С4.1. Поскольку $\dim \operatorname{conv} S = d$, то и $\dim L_S = d$. Это означает, что система векторов линейно зависима, поскольку состоит из $d+1$ вектора. Тогда существуют $\beta_2, \ldots, \beta_n$, не равные нулю одновременно, такие, что

$$\sum_{i=2}^{n} \beta_i (x_i - x_1) = 0.$$

Обозначим

$$\gamma_1 = -\sum_{i=2}^{n} \beta_i,\ \gamma_i = \beta_i.$$

Это значит, что можно переписать исходное выражение как

$$\sum_{i=1}^{n} \gamma_i x_i = 0,\ \sum_{i=1}^{n} \gamma_i = 0,$$

причем не все γ_i равны нулю одновременно. Это означает, что среди них есть такие, которые принимают отрицательное значение. Обозначим

$$a = \min\left\{ -\frac{\alpha_i}{\gamma_i} : \gamma_i < 0 \right\} = -\frac{\alpha_k}{\gamma_k}.$$

Тогда $\forall i\ \alpha_i + a\gamma_i \geq 0$, причем

$$x = \sum_{i=1}^{n} (\alpha_i + a\gamma_i) x_i,\ \sum_{i=1}^{n} (\alpha_i + a\gamma_i) = 1,$$

то есть это выпуклая комбинация. Но поскольку $\alpha_k + a\gamma_k = 0$ по определению, то мы уменьшили число точек на одну. Этот процесс можно продолжать до тех пор, пока не окажется $n = d+1$. В таком случае продолжать итеративно будет невозможно, поскольку система d векторов в d-мерном пространстве не обязательно линейно зависима. ■

Теорема С4.4. (Теорема Радона) Пусть $\dim X = d$, тогда любое конечное множество $S \subseteq X$, содержащее не менее $d+2$ точек может быть разбито на два не пересекающихся множества, выпуклые оболочки которых пересекаются.

Доказательство. Рассмотрим набор точек $x_1, \ldots, x_n \in S$, $n \geq d+2$. Из доказательства Теоремы С4.3, утверждается линейная зависимость системы $x_2 - x_1, \ldots, x_n - x_1$. Тогда существует нетривиальная комбинация этих векторов с коэффициентами $\beta_2, \ldots, \beta_n$, равная нулю. Введем коэффициенты $\gamma_1, \ldots, \gamma_n$ аналогично предыдущему доказательству, получив

$$\sum_{i=1}^{n} \gamma_i x_i = 0, \ \sum_{i=1}^{n} \gamma_i = 0.$$

Выделим положительные и отрицательные коэффициенты в разные группы. Не ограничивая общность рассуждений, будем считать, что $\gamma_1, \ldots, \gamma_m \geq 0$, $\gamma_{m+1}, \ldots, \gamma_n < 0$. Тогда имеем

$$\sum_{i=1}^{m} \tilde{\gamma}_i x_i = \sum_{i=m+1}^{n} \tilde{\gamma}_i x_i,$$

где для $j = \overline{1, m}$ определили $\tilde{\gamma}_i = \frac{\gamma_i}{\sum_{j=1}^{m} \gamma_j}$, а для остальных — $\tilde{\gamma}_i = -\frac{\gamma_i}{\sum_{j=1}^{m} \gamma_j}$. Нетрудно видеть, что в обеих частях равенства стоят выпуклые комбинации. Таким образом, получили общую точку двух выпуклых оболочек. ■

Теорема С4.5. (Теорема Хелли для конечных семейств) Пусть $\dim X = d$ и есть семейство из $n \geq d+1$ выпуклых множеств $S_1, \ldots, S_n \subseteq X$, такое, что любые $d+1$ имеют общую точку. Тогда все эти множества имеют общую точку.

Доказательство.

- База при $n = d+1$ очевидна.
- Рассмотри $n \geq d+2$ выпуклых множеств $S_1, \ldots, S_n$. Предположим, что утверждение доказано для любых $n-1$ выпуклых множеств. Из каждого выберем некоторую точку:

  $$x_k = \bigcap_{i=1, i \neq k}^{n} S_i.$$

  Такие пересечения не пусты по предположению индукции. Но $n \geq d+2$, то есть можем воспользоваться теоремой Радона. Переставив индексы, получаем существование такого m, что $\operatorname{conv}\{x_1, \ldots, x_m\}$ и $\operatorname{conv}\{x_{m+1}, \ldots, x_n\}$ имеют общую точку x. Более того, каждая точка из множества $x_1, \ldots, x_m$ содержится в каждом из множеств $S_{m+1}, \ldots, S_n$. Это означает, что x содержится в каждом $S_{m+1}, \ldots, S_n$ в силу их выпуклости. Аналогично доказывается принадлежность x множествам $S_1, \ldots, S_m$. Таким образом, нашли общую точку для всех множеств.

■

С4.4 Операции, сохраняющие выпуклость

Определение С4.5. Функция $f : \mathbb{R}^d \to \mathbb{R}^n$ называется *аффинной*, если найдутся $b \in \mathbb{R}^m$ и $A \in \mathbb{R}^{m \times d}$, такие, что $f(x) = Ax + b$.

Ряд базовых операций имеет свойство сохранять выпуклость множеств.

Утверждение C4.4. Следующие операции оставляют множество выпуклым:

- **Пересечение:** Пусть $\{S_i\}_{i\in I} \subseteq X$ — семейство выпуклых множеств, тогда пересечение $\bigcap_{i\in I} S_i$ также является выпуклым.
- **Линейная комбинациия:** Пусть $S_1, S_2 \subseteq X$ — выпуклые множества и $c_1, c_2 \in \mathbb{R}$, тогда линейная комбинация $c_1S_1 + c_2S_2 = \{\, c_1x_1 + c_2x_2 \mid x_1 \in S_1,\ x_2 \in S_2 \,\}$ также является выпуклым множеством.
- **Взятие образа при аффинном преобразовании:** Пусть $S \subseteq X$ — выпуклое множество, f — аффинная функция, тогда $f\big(S\big)$ также является выпуклым множеством.
- **Взятие прообраза при аффинном преобразовании:** Пусть $S \subseteq X$ — выпуклое множество, f — аффинная функция, тогда $f^{-1}(S)$ также является выпуклым множеством.
- **Произведение:** Пусть $S_1, S_2, \ldots, S_n \subseteq X$ — выпуклые множества, тогда декартово произведение $S_1 \times S_2 \times \ldots \times S_n$ также выпукло.

Пример C4.11. Покажите, что многогранник:

$$S(A, b, C, d) = \left\{ x \in \mathbb{R}^d \;\middle|\; Ax \preceq b,\ Cx = d \right\},$$

где $A \in \mathbb{R}^{n\times d}$, $b \in \mathbb{R}^n$, $C \in \mathbb{R}^{k\times d}$ и $d \in \mathbb{R}^n$ является выпуклым.

Решение. Многогранник является выпуклым множеством как пересечение полуплоскостей и плоскостей. ■

Пример C4.12. Покажите, что множество

$$S(\alpha, \beta) = \left\{ a \in \mathbb{R}^d \;\middle|\; p(0) = 1,\ |p(t)| \le 1\ \forall t : \alpha \le t \le \beta \right\},$$

где

$$p(t) = a_0 + a_1 t + \ldots + a_{d-1}t^{d-1},$$

является выпуклым.

Решение. Переформулируем ограничения на коэффициенты a в виде линейных равенств и неравенств. Тогда

$$p(0) = 1 \iff a_0 = 1,$$

ограничение вида $|p(t)| \le 1$ переписывается как

$$-1 \le a_0 + a_1 t + \ldots + a_{d-1}t^{d-1} \le 1.$$

Обозначим это пересечение двух полуплоскостей за $\mathcal{H}_t$. Тогда множество коэффициентов переписывается в виде

$$\bigcap_{t\in[\alpha,\beta]} \mathcal{H}_t \cap \{\, a \mid a_0 = 1 \,\},$$

что является выпуклым как пересечение выпуклых множеств. ■

Пример С4.13. Покажите, что множество

$$S(A,b,c,d) = \left\{ x \;\middle|\; \|Ax+b\| \le c^\top x + d \right\},$$

где $A \in \mathbb{R}^{n\times d}$, $b \in \mathbb{R}^n$, $c \in \mathbb{R}^d$ и $d \in \mathbb{R}$ является выпуклым.

Решение. Докажем для начала, что множество

$$T = \{\, (x,t) \mid \|x\| \le t \,\}$$

является выпуклым. Это несложно делается по определению: возьмем $\alpha \in [0,1]$ и $(x_1,t_1),(x_2,t_2) \in T$. Тогда

$$\|\alpha x_1 + (1-\alpha)x_2\| \le \alpha t_1 + (1-\alpha)t_2,$$

что означает

$$(\alpha x_1 + (1-\alpha)x_2,\, t_1 + (1-\alpha)t_2) \in T.$$

Заметим, что $S = f^{-1}(T)$, где $f(x) = (Ax+b, c^\top x + d)$ — аффинная функция. Получим, что S выпукло как прообраз выпуклого множества при аффинном преобразовании. ■

Пример С4.14. Покажите, что множество

$$S(P,c) = \left\{ x \;\middle|\; x^\top P x \le (c^\top x)^2,\ c^\top x \ge 0 \right\},$$

где $P \in \mathbb{S}^d_+$ и $c \in \mathbb{R}^d$ является выпуклым.

Решение. Как уже известно, множество

$$T = \{\, (x,t) \mid \|x\| \le t \,\}$$

является выпуклым. Воспользуемся фактом, что

$$\forall P \in \mathbb{S}^d_+ \ \exists Q \in \mathbb{S}^d_+ : P = Q^2,$$

поэтому наше множество S переписывается как

$$S = \left\{ x \;\middle|\; \|Qx\| \le c^\top x \right\}.$$

Заметим, что $S = f^{-1}(T)$, где $f(x) = (Qx, c^\top x)$ — аффинная функция. Поэтому S выпукло как прообраз выпуклого множества при аффинном преобразовании. ■

С4.5 Конусы

Определение С4.6. Множество $K \subseteq X$ называется *конусом*, если для любых $x \in K$ и $\alpha \ge 0$ точка αx также принадлежит K.

Пример С4.15. В качестве простейшей иллюстрации определения, приведем ряд базовых примеров конусов:

- Множества $\emptyset$, $\{0\}$ и $\mathbb{R}^d$;

- Прямая, проходящая через начало координат;
- Луч, начинающийся из начала координат;
- Любое линейное подпространство.

Утверждение C4.5. $K \subseteq X$ является выпуклым конусом, тогда и только тогда, когда для любых $x_1, x_2 \in K$, $\alpha_1, \alpha_2 \geq 0$ выполнено $\alpha_1 x_1 + \alpha_2 x_2 \in K$.

Доказательство.
$\Rightarrow$ Пусть K — выпуклый конус. Тогда верно, что для $x_1, x_2 \in K$ выполняется включение $\alpha x_1 + (1-\alpha)x_2 \in K$, $0 \geq \alpha \geq 1$. Возьмем $\alpha_1, \alpha_2 \geq 0$ одновременно не равные нулю и представим через них α:

$$\alpha = \frac{\alpha_1}{\alpha_1 + \alpha_2}.$$

Тогда воспользуемся определением конуса и получим $\alpha_1 c_1 + \alpha_2 c_2 \in K$.
$\Leftarrow$ Пусть $\alpha_1 x_1 + \alpha_2 x_2 \in K$, тогда если взять $\alpha_1 = \alpha$ и $\alpha_2 = 1 - \alpha$ при $0 \leq \alpha \leq 1$, то получим, что K является выпуклым. А если взять $\alpha_1 = \alpha$ и $\alpha_2 = 0$, то получим, что K — конус. ■

Утверждение C4.6. Пересечение любого семейства множеств сохраняет свойство быть конусом, а также сохраняется свойство быть выпуклым конусом.

Доказательство.

- Пусть $x \in \bigcap_{i \in I} K_i$, тогда $\alpha x \in \bigcap_{i \in I} K_i$, так как $\alpha x \in K_i \ \forall i \in I$.
- Следует из того, что пересечение выпуклых множеств — выпукло и пересечение конусов — конус.

■

Пример C4.16. Покажите, что множество

$$M = \left\{ (x,t) \in R^{d+1} \;\middle|\; \|x\| \leq t \right\}$$

является выпуклым конусом.

Решение. Проверим по критерию. Пусть $\alpha_1, \alpha_2 \geq 0$, (x_1, t_1), $(x_2, t_2) \in M$, тогда

$$\|\alpha_1 x_1 + \alpha_2 x_2\| \leq \alpha_1 \|x_1\| + \alpha_2 \|x_2\| \leq \alpha_1 t_1 + \alpha_2 t_2.$$

Следовательно точка $\alpha_1(x_1, t_1) + \alpha_2(x_2, t_2)$ также из M, то есть это выпуклый конус. ■

Пример C4.17. Покажите, что $\mathbb{S}_+$ является выпуклым конусом.

Решение. Нетрудно заметить, что умножение матрицы на неотрицательный коэффициент не меняет определенность. Ровно как и суммирование положительно полуопределенных матриц дает положительно полуопределенную матрицу. ■

Пример C4.18. Матрица $D \in \mathbb{S}^n$, поэлементно определённая по правилу

$$D_{ij} = \|x_i - x_j\|_2^2$$

для каких-то $x_1, \ldots, x_n \in \mathbb{R}^d$, называется *евклидовой матрицей расстояний*. Покажите, что множество евклидовых матриц расстояний является выпуклым конусом.

Решение. Достаточно доказать, что если матрицы A и B евклидовы, то αA и $A + B$ также являются евклидовыми.

- Пусть векторы $x_1, \ldots, x_n \in \mathbb{R}^d$ определяют матрицу A, тогда векторы

  $$\sqrt{\alpha}x_1, \ldots, \sqrt{\alpha}x_n \in \mathbb{R}^d$$

  определяют матрицу αA, а значит она тоже евклидова.

- Пусть векторы $y_1, \ldots, y_n \in \mathbb{R}^m$ определяют матрицу B, тогда векторы

  $$(x_1, y_1), \ldots, (x_n, y_n) \in \mathbb{R}^{d+m}$$

  определяют матрицу $A + B$, поэтому она евклидова.

Таким образом, множество всех евклидовых матриц является выпуклым конусом. ■

Пример C4.19. Матрица $X \in \mathbb{S}^d$ называется *коположительной*, если $z^\top X z \geq 0$ для всех $z \succeq 0$. Покажите, что множество коположительных матриц является выпуклым конусом.

Решение. Обозначим множество коположительных матриц за C. Перепишем это множество как пересечение полуплоскостей:

$$C = \bigcap_{z \succeq 0} \left\{ X \in \mathbb{S}^d \;\middle|\; z^\top X z \geq 0 \right\}.$$

Полуплоскости являются выпуклыми конусами, а операция пересечения сохраняет это свойство, поэтому C также является выпуклым конусом. ■

C4.6 Теоремы об отделимости

Теорема C4.6 (**Теорема об отделимости**). Пусть S и T — не пересекающиеся непустые выпуклые множества в $\mathbb{R}^d$. Тогда найдется $\lambda \in \mathbb{R}^d \setminus \{0\}$, такое, что

$$\sup_{x \in S} \lambda^\top x \leq \inf_{y \in T} \lambda^\top y.$$

То есть любые два не пересекающиеся множества можно разделить гиперплоскостью.

Пример C4.20. Пусть S — выпуклое множество с непустой внутренностью. Точка x_0 такая, что $x_0 \notin \operatorname{int} S$. Покажите, что найдется $\lambda \in \mathbb{R}^d \setminus \{0\}$, такое, что

$$\sup_{x \in S} \lambda^\top x \leq \lambda^\top x_0.$$

Решение. Пусть $x \in \operatorname{int} S$, $B_{r_0}(x) \subseteq S$ и $y \in \partial S$, тогда весь интервал $[x, y)$ принадле-

жит внутренности, так как если $z \in [x, y)$, то $B_r(z) \subseteq \operatorname{int} S$, где

$$r = r_0 \frac{\|y - z\|_2}{\|y - x\|_2}.$$

Поэтому $\partial S \subseteq \operatorname{cl} \operatorname{int} S$. Получаем, что

$$\operatorname{cl} S = \operatorname{cl}(\operatorname{int} S \cup \partial S) \subseteq \operatorname{cl}(\operatorname{cl} \operatorname{int} S) = \operatorname{cl} \operatorname{int} S.$$

А так как $\operatorname{cl} S \supseteq \operatorname{cl} \operatorname{int} S$, то $\operatorname{cl} S = \operatorname{cl} \operatorname{int} S$. Заметим теперь, что $\operatorname{int} S$ и x_0 — непустые не пересекающиеся выпуклые множества. Применив для них теорему об отделимости, получим

$$\sup_{x \in \operatorname{int} S} \lambda^\top x \leq \lambda^\top x_0.$$

Требуемое утверждение получается взятием замыкания у $\operatorname{int} S$. ■

Пример С4.21. Рассмотрим систему строгих неравенств

$$Ax \prec b,\ A \in \mathbb{R}^{n \times d},\ b \in \mathbb{R}^n.$$

Она неразрешима тогда и только тогда, когда

$$\exists \lambda \in \mathbb{R}^n \setminus \{0\} :\ \lambda^\top A = 0,\ \lambda^\top b \leq 0,\ \lambda \succeq 0.$$

Доказательство. $\Rightarrow$ Пусть система $Ax \prec b$ неразрешима, тогда выпуклые множества $\left\{ b - Ax \;\middle|\; x \in \mathbb{R}^d \right\}$ и $\mathbb{R}^d_{++}$ не пересекаются. Воспользуемся теоремой об отделимости:

$$\exists \lambda \in \mathbb{R}^n \setminus \{0\} :\ \sup_{x \in \mathbb{R}^d} \lambda^\top (b - Ax) \leq \inf_{y \in \mathbb{R}^d_{++}} \lambda^\top y.$$

Так как правая часть неравенства не может быть равна $-\infty$, то $\lambda \succeq 0$. Помимо этого, устремляя $y \to 0$, получим

$$\sup_{x \in \mathbb{R}^d} \lambda^\top (b - Ax) \leq 0.$$

Так как супремум ограничен сверху, можем сделать вывод, что

$$\lambda^\top A = 0,\ \lambda^\top b \leq 0.$$

$\Leftarrow$ Предположим, что нашелся такой $x \in \mathbb{R}^d$, что $Ax \prec b$. Так как $\lambda \succeq 0$ и $\lambda \neq 0$, то

$$\lambda^\top Ax < \lambda^\top b \leq 0.$$

Отсюда мы получаем противоречие с тем, что $\lambda^\top A = 0$. ■

Сила данного утверждения в том, что неразрешимость какого-то линейного неравенства от x, который лежит в d-мерном пространстве, сводится к разрешимости системы равенств и неравенств от переменной из n-мерного пространства. Это может быть полезно в случае $n \ll d$.

Пример С4.22. Операция взятия сопряжения нормы является *инволюцией*, т.е. $\|\cdot\|_{**} = \|\cdot\|$.

Доказательство. Для начала заметим, что

$$\|y\|_* = \sup_{\|x\|\le 1} x^\top y \ge \left(\frac{x}{\|x\|}\right)^\top y.$$

Из этого следует

$$x^\top y \le \|x\|\|y\|_*.$$

Более того, так как $\{\, x \mid \|x\| \le 1 \,\}$ — компакт, то непрерывная функция $x^\top y$ по теореме Вейерштрасса достигает на нем максимума в какой-то точке, следовательно

$$\forall y\ \exists x \ne 0 :\ x^\top y = \|x\|\|y\|_*.$$

Отсюда следует, что

$$\forall x\ \exists y \ne 0 :\ \|x\|_{**}\|y\|_* = x^\top y \le \|x\|\|y\|_*.$$

Поэтому

$$\forall x \to \|x\|_{**} \le \|x\|.$$

Далее будем пользоваться теоремой об отделимости. Зафиксируем $x \ne 0$ (случай $x = 0$ очевиден) и рассмотрим два выпуклых не пересекающихся множества:

$$B(0,1) = \{\, x \mid \|x\| < 1 \,\},\ \left\{\frac{x}{\|x\|}\right\}.$$

По теореме об отделимости найдется такой $y \ne 0$, что

$$\sup_{\|x\|<1} x^\top y \le \left(\frac{x}{\|x\|}\right)^\top y.$$

Следовательно,

$$\|y\|_* = \sup_{\|x\|\le 1} x^\top y = \sup_{\|x\|<1} x^\top y \le \left(\frac{x}{\|x\|}\right)^\top y \le \frac{\|x\|_{**}\|y\|_*}{\|x\|}.$$

Получили неравенство в обратную сторону:

$$\forall x \to \|x\|_{**} \ge \|x\|.$$

Это рассуждение завершает доказательство. ■

Этот пример показывает, что иногда теорема об отделимости может возникать в совершенно неожиданных местах, и без ее применения некоторые утверждения доказать нелегко.

Теорема C4.7 (**Теорема о строгой отделимости**). Пусть S и T – не пересекающиеся непустые выпуклые множества в $\mathbb{R}^d$, причем S — компакт, а T — замкнуто. Тогда найдется $\lambda \in \mathbb{R}^d \setminus \{0\}$ такое, что

$$\sup_{x \in S} \lambda^\top x < \inf_{y \in T} \lambda^\top y.$$

Пример C4.23. Множества

$$A = \left\{ (x, y) \;\middle|\; x > 0,\ y \geq \frac{1}{x} \right\},$$
$$B = \left\{ (x, y) \;\middle|\; x > 0,\ y \leq -\frac{1}{x} \right\}.$$

не могут быть строго отделены гиперплоскостью.

Доказательство. Предположим нашелся такой $\lambda \neq 0$, который строго разделяет эти множества, то есть найдутся $c_1, c_2 \in \mathbb{R}$ такие, что

$$\sup_{a \in A} \lambda^\top a < c_1 < c_2 < \inf_{b \in B} \lambda^\top b.$$

Для любого $x > 0$ выполняются включения

$$\left(x, \frac{1}{x}\right) \in A,\ \left(x, -\frac{1}{x}\right) \in B.$$

Возьмем такой достаточно большой $x > 0$, что выполняется

$$2\left|\frac{\lambda_2}{x}\right| < c_2 - c_1.$$

Тогда

$$\lambda_1 x + \frac{\lambda_2}{x} < c_1 \implies \lambda_1 x - \frac{\lambda_2}{x} < c_1 - 2\frac{\lambda_2}{x} \leq c_1 + 2\left|\frac{\lambda_2}{x}\right| < c_2,$$

то есть

$$\lambda_1 x - \frac{\lambda_2}{x} < c_2.$$

Это приводит нас к противоречию. ■

C5 Выпуклые функции

Многие методы оптимизации опираются на свойства выпуклых функций. Для них локальный минимум всегда является глобальным, а если функция является сильно выпуклой, то минимум единственный.

C5.1 Основные понятия

Перед тем, как перейти к обсуждению основной темы главы, введём несколько вспомогательных определений.
Оказывается, что иногда довольно удобно разрешить функции принимать бесконечное значение и рассматривать её так, будто она задана на всем пространстве, а не только на области определения.

Определение C5.1. Множество $\overline{\mathbb{R}} = \mathbb{R} \cup \{+\infty\} \cup \{-\infty\}$ будем называть *расширенным множеством действительных чисел.*

Замечание C5.1. Операции, определенные в $\mathbb{R}$, естественным образом переносятся на $\overline{\mathbb{R}}$:

- Операции с числами из $\mathbb{R}$ понимаются в обычном смысле.
- Любое число из $\overline{\mathbb{R}} \setminus \{+\infty\}$ строго меньше $+\infty$. В частности, $-\infty < +\infty$.
- Сумма $+\infty$ и любого числа из $\overline{\mathbb{R}} \setminus \{-\infty\}$ равна $+\infty$.
- Сумма $-\infty$ и любого числа из $\overline{\mathbb{R}} \setminus \{+\infty\}$ равна $-\infty$.
- Сложение $+\infty$ и $-\infty$ не определено.
- Произведение $+\infty$ и положительного числа из $\overline{\mathbb{R}}$ (в том числе $+\infty$) равно $+\infty$.
- Произведение $+\infty$ и отрицательного числа из $\overline{\mathbb{R}}$ (в том числе $-\infty$) равно $-\infty$.
- Произведение $-\infty$ и положительного числа из $\overline{\mathbb{R}}$ (в том числе $+\infty$) равно $-\infty$.
- Произведение $-\infty$ и отрицательного числа из $\overline{\mathbb{R}}$ (в том числе $-\infty$) равно $+\infty$.
- Произведение нуля и $+\infty$, $-\infty$ не определено.

Определив $\overline{\mathbb{R}}$, мы готовы рассматривать функцию сразу в расширенном виде. Если функция определена на множестве, то вне этого множества доопределим её $+\infty$. Истинную область определения функции теперь будем называть эффективным множеством.

Определение C5.2. Рассмотрим функцию $f : \mathbb{R}^d \to \overline{\mathbb{R}}$. Её *эффективным множеством* будем называть множество всех точек, в которых она принимает конечные значения:

$$\operatorname{dom} f = \left\{ x \in \mathbb{R}^d \;\middle|\; |f(x)| < +\infty \right\}.$$

Введем еще одно определение, оно нам пригодится позже.

Определение C5.3. Рассмотрим функцию $f : \mathbb{R}^d \to \overline{\mathbb{R}}$. Будем называть её

собственной, если $\operatorname{dom} f \neq \varnothing$ и

$$\forall x \in \mathbb{R}^d : f(x) > -\infty.$$

C5.2 Выпуклые функции

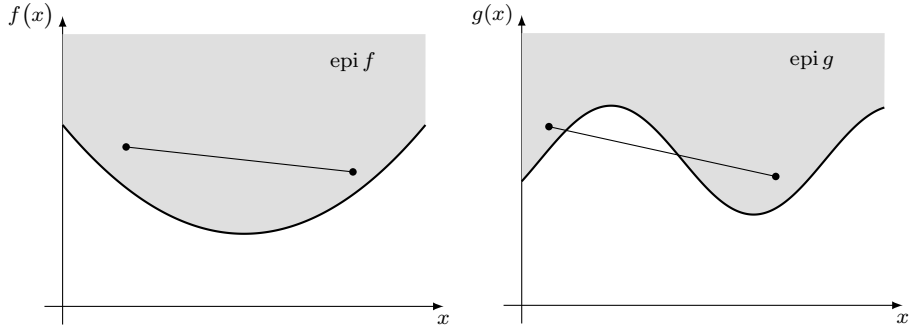
Предмет изучения этого главы — выпуклые функции. Поскольку ранее мы уже разобрали выпуклые множества, наиболее естественно ввести выпуклость функции, опираясь на понятие надграфика.

Определение C5.4. Пусть функция $f : \mathbb{R}^d \to \overline{\mathbb{R}}$. *Надграфиком (эпиграфом)* f называется множество

$$\operatorname{epi} f = \left\{ (x,t) \in \mathbb{R}^d \times \overline{\mathbb{R}} \;\middle|\; f(x) \leq t \right\}.$$

Нетрудно заметить, что график функции — есть граница эпиграфа. Довольно естественно называть функцию выпуклой, если её надграфик ограничивает выпуклое множество.

Определение C5.5. Рассмотрим функцию $f : \mathbb{R}^d \to \overline{\mathbb{R}}$. Будем называть её *выпуклой*, если надграфик $\operatorname{epi} f$ — есть выпуклое множество.



(a) Выпуклая парабола.

(b) Не выпуклый синус.

Рис. C5.1: Выпуклость через надграфик.

Пример C5.1. Покажите, что аффинная функция $f : \mathbb{R}^d \to \overline{\mathbb{R}}$:

$$f(x) = a^\top x + b,\ a \in \mathbb{R}^d,\ b \in \mathbb{R}$$

является выпуклой.

Решение. Посмотрим, как устроено множество $\operatorname{epi} f$. Это все точки, лежащие над гиперплоскостью, заданной аффинной функцией f. Покажем, что множество

$$S(a,b) = \left\{ (x,t) \in \mathbb{R}^d \times \overline{\mathbb{R}} \;\middle|\; a^\top x + b \leq t \right\}$$

является выпуклым по определению. Пусть $(x_1, t_1), (x_2, t_2) \in S$, тогда

$$a^\top(\alpha x_1 + (1-\alpha)x_2) + b = \alpha\big(a^\top x_1 + b\big) + (1-\alpha)\big(a^\top x_2 + b\big) \leq \alpha t_1 + (1-\alpha)t_2.$$

Это означает, что аффинная функция также выпукла по определению. ■

Пример C5.2. Покажите, что векторная норма $f : \mathbb{R}^d \to \overline{\mathbb{R}}$:

$$f(x) = \|x\|$$

является выпуклой.

Доказательство. Покажем, что множество

$$S = \left\{ (x,t) \in \mathbb{R}^d \times \overline{\mathbb{R}} \;\middle|\; \|x\| \leq t \right\}$$

является выпуклым по определению. Пусть $(x_1, t_1), (x_2, t_2) \in S$, тогда

$$\big\|\alpha x_1 + (1-\alpha)x_2\big\| \leq \alpha\big\|x_1\big\| + (1-\alpha)\big\|x_2\big\| \leq \alpha t_1 + (1-\alpha)t_2.$$

Это означает, что норма — выпуклая функция по определению. ■

Пример C5.3. Покажите, что функция $f : \mathbb{R}^d \to \overline{\mathbb{R}}$:

$$f(x) = \|x\|^4 - 4\|x\|^2$$

не является выпуклой.

Доказательство. Покажем, что множество

$$S = \left\{ (x,t) \in \mathbb{R}^d \times \overline{\mathbb{R}} \;\middle|\; \|x\|^4 - 4\|x\|^2 \leq t \right\}$$

не является выпуклым. Возьмем $\big((1,0,\ldots,0)^\top, -1\big), \big((-1,0,\ldots,0)^\top, -1\big) \in S$ и рассмотрим точку посередине и получим противоречие с неравенством во множестве: $-1 < 0$. ■

Работать непосредственно с эпиграфами далеко не всегда удобно и зачастую требует нетривиальных ходов. Чтобы подкрепить это утверждение, мы рассмотрим сложные примеры.

Пример C5.4. Покажите, что функция $f : \mathbb{R} \to \overline{\mathbb{R}}$:

$$f(x) = \begin{cases} x\ln x, & x > 0; \\ 0, & x = 0. \end{cases}, \quad \operatorname{dom} f = \mathbb{R}_+$$

является выпуклой.

Доказательство. Проанализировать надграфик предложенной функции непросто. Покажем, что он является пересечением выпуклых множеств.

- Рассмотрим вспомогательную функцию

$$g(x) = \sup_{y \in \mathbb{R}} \{\, xy - e^y \,\}.$$

Найдем явный вид этой функции. При $x < 0$ получаем, что $g\big(x\big) = +\infty$, при $x = 0$ получаем, что $g\big(x\big) = 0$, при $x > 0$ получаем, что $g\big(x\big) = x \ln x - x$. В итоге

$$g\big(x\big) = \begin{cases} x \ln x - x, & x > 0; \\ 0, & x = 0. \end{cases}$$

Получается, что $f\big(x\big) = g\big(x\big) + x$. В Примере C5.1 мы показали, что $\operatorname{epi} x$ — выпуклый, поэтому, если $\operatorname{epi} g$ — выпуклый, то и $\operatorname{epi} f$ — выпуклый.

- Надграфик функции g можно задать так:

$$\operatorname{epi} g = \bigcap_{y \in \mathbb{R}} \left\{ (x,t) \in \mathbb{R}^d \times \overline{\mathbb{R}} \;\middle|\; xy - e^y \leq t \right\}.$$

При фиксированном $y \in \mathbb{R}$, множество $\left\{ (x,t) \in \mathbb{R}^d \times \overline{\mathbb{R}} \;\middle|\; xy - e^y \leq t \right\}$ — выпуклое, как полуплоскость. Но тогда и $\operatorname{epi} g$ — выпуклое, как сумма выпуклых множеств.

■

Пример C5.5. Покажите, что функция $f : \mathbb{R}^d \to \overline{\mathbb{R}}$:

$$f\big(p\big) = \mathrm{KL}(p|q) = \sum_{i=1}^{d} p_i \ln \frac{p_i}{q_i}, \; q \in \mathbb{R}^d_{++}, \quad \operatorname{dom} f = \left\{ p \in \mathbb{R}^d_{+} \;\middle|\; \sum_{i=1}^{d} p_i = 1 \right\},$$

является выпуклой.

Доказательство. Запишем $\operatorname{epi} f$:

$$\operatorname{epi} f = \left\{ (p,t) \in \mathbb{R}^d \times \overline{\mathbb{R}} \;\middle|\; \sum_{i=1}^{d} p_i \ln \frac{p_i}{q_i} \leq t, \; \sum_{i=1}^{d} p_i = 1, \; p_i \geq 0 \right\}.$$

Введём переменную s, расширив размерность надграфика:

$$M = \left\{ (s,p,t) \in \mathbb{R}^d \times \mathbb{R}^d \times \overline{\mathbb{R}} \;\middle|\; \sum_{i=1}^{d} s_i \leq t, \; p_i \ln \frac{p_i}{q_i} \leq s_i, \; \sum_{i=1}^{d} p_i = 1, \; p_i \geq 0 \right\}.$$

Если показать, что множество M — выпуклое, то $\operatorname{epi} f$ будет выпуклым, так как это ортогональная проекция множества M. Представим множество M как пересечение выпуклых. Множество

$$R = \left\{ (s,p,t) \in \mathbb{R}^d \;\middle|\; \sum_{i=1}^{d} s_i \leq t \right\}$$

выпукло как полуплоскость. Проверим множество

$$S = \left\{ (s,p,t) \in \mathbb{R}^d \times \mathbb{R}^d \times \overline{\mathbb{R}} \;\middle|\; \sum_{i=1}^{d} p_i = 1, \; p_i \geq 0 \right\}$$

на выпуклость. Пусть $(s^1,p^1,t^1), (s^2,p^2,t^2) \in S$, проверим сначала равенство

$$\sum_{i=1}^{d} \big(\alpha p_i^1 + (1-\alpha) p_i^2\big) = \alpha \sum_{i=1}^{d} p_i^1 + (1-\alpha) \sum_{i=1}^{d} p_i^2 = 1,$$

а теперь проверим неравенство:

$$\alpha p_i^1 + (1-\alpha)p_i^2 \geq 0$$

Осталось убедиться в выпуклости множеств вида

$$M_i = \left\{ (p,s,t) \in \mathbb{R}^d \times \mathbb{R}^d \times \overline{R} \;\middle|\; s_i \geq p_i \ln \frac{p_i}{q_i} \right\}.$$

Можно заметить, что M_i — есть эпиграф функции $g_i\big(x\big) = x\ln x - x\ln q_i$. Выпуклость $x\ln x$ была показана в предыдущем примере. Кроме того, в предыдущем примере мы отметили, что добавление линейной функции не влияет на выпуклость. Таким образом, множество M выпуклое. ■

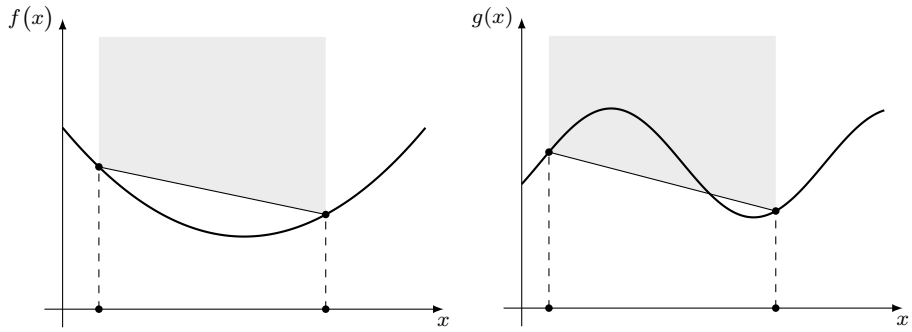
Из последних примеров можно видеть, что проверка выпуклости по определению зачастую довольно сильно затруднена. Известен ряд эквивалентных определений.

Определение C5.6. Рассмотрим функцию $f : \mathbb{R}^d \to \overline{\mathbb{R}}$. Функция f называется *выпуклой*, если для любых $x, y \in \mathbb{R}^d$ и любого $\alpha \in [0,1]$ выполняется

$$f\big(\alpha x + (1-\alpha)y\big) \leq \alpha f\big(x\big) + (1-\alpha) f\big(y\big). \tag{C5.1}$$

Замечание C5.2. Неравенство (C5.1) называется неравенством Йенсена для двух точек.

Неформально это означает, что хорда, соединяющая любые две точки графика функции, лежит над ним. Эта интуиция может быть знакома из курса математического анализа.



(a) Выпуклая парабола.

(b) Не выпуклый синус.

Рис. C5.2: Выпуклость через условие Йенсена для двух точек.

Замечание C5.3. Если в Определении C5.6 выполнено строгое неравенство, функцию f называют *строго выпуклой*.

Замечание C5.4. Если неравенство в Определении C5.6 (строго) выполнено в обратную сторону, f называется *(строго) вогнутой*.

Замечание C5.5. Из определений легко видеть, что функция f является (строго) выпуклой тогда и только тогда, когда функция $-f$ является (строго) вогнутой.

Утверждение C5.1. Определения C5.5 и C5.6 эквивалентны.

Доказательство. $\Rightarrow$ Покажем, что из Определения C5.5 следует Определение C5.6. Возьмем $\big(x_1, f(x_1)\big), \big(x_2, f(x_2)\big) \in \operatorname{epi} f$, тогда

$$f\big(\alpha x_1 + (1-\alpha)x_2\big) \le \alpha f\big(x_1\big) + (1-\alpha) f\big(x_2\big).$$

$\Leftarrow$ Покажем, что из Определения C5.6 следует Определение C5.5. Рассмотрим для точек $(x, t_x), (y, t_y) \in \operatorname{epi} f$ выпуклую комбинацию $\big(\alpha x + (1-\alpha)y, \alpha t_x + (1-\alpha)t_y\big)$. Тогда:

$$f\big(\alpha x + (1-\alpha)y\big) \le \alpha f\big(x\big) + (1-\alpha) f\big(y\big) \le \alpha t_x + (1-\alpha)t_y.$$

Получили, что $\operatorname{epi} f$ — выпуклый.

■

Пример C5.6. Покажите, что функция $f : \mathbb{R} \to \overline{\mathbb{R}}$:

$$f\big(x\big) = -\sqrt{x}, \quad \operatorname{dom} f = \mathbb{R}_+$$

является выпуклой.

Решение. Проверим по Определению C5.6, возьмем две точки x, y:

$$-\sqrt{\alpha x + (1-\alpha)y} \le -\alpha\sqrt{x} - (1-\alpha)\sqrt{y}.$$

Возведем в квадрат:

$$\alpha^2 x + 2\alpha(1-\alpha)\sqrt{xy} + (1-\alpha)^2 y \le \alpha x + (1-\alpha)y.$$

Упростим:

$$-\alpha(1-\alpha)x + 2\alpha(1-\alpha)\sqrt{xy} - \alpha(1-\alpha)y \le 0.$$

Нужно, показать, что:

$$2\sqrt{xy} \le x + y \implies 0 \le (x-y)^2.$$

■

Утверждение C5.2. Функция $f : \mathbb{R}^d \to \overline{\mathbb{R}}$ выпукла тогда и только тогда, когда функция одного аргумента $g\big(\alpha\big) = f\big(x + \alpha v\big)$ выпукла по α для любого ненулевого $v \in \mathbb{R}^d$.

Доказательство. $\Rightarrow$ Возьмём произвольные точку x, направление v и рассмотрим выпуклую комбинацию двух точек α_1, α_2. Тогда в в силу выпуклости функции $f\big(x\big)$ получаем:

$$\begin{aligned} g\big(\lambda\alpha_1 + (1-\lambda)\alpha_2\big) &= f\big(x + (\lambda\alpha_1 + (1-\lambda)\alpha_2)v\big) = f\big(\lambda(x+\alpha_1 v) + (1-\lambda)(x+\alpha_2 v)\big) \\ &\le \lambda f\big(x+\alpha_1 v\big) + (1-\lambda)(x+\alpha_2 v) = \lambda g\big(\alpha_1\big) + (1-\lambda) g\big(\alpha_2\big). \end{aligned}$$

Таким образом, функция $g\big(\alpha\big)$ выпукла.

$\Leftarrow$ Пусть теперь x_1 и x_2 — две отличные друг от друга точки, тогда направление $v = x_1 - x_2$ ненулевое. В силу выпуклости функции $g\big(\alpha\big)$ получаем:

$$\begin{aligned} f\big(x_2 + \alpha(x_1 - x_2)\big) &= f\big(\alpha x_1 + (1-\alpha)x_2\big) = f\big(x_2 + (\alpha \cdot 1 + (1-\alpha)\cdot 0)v\big) \\ &= g\big(\alpha \cdot 1 + (1-\alpha)\cdot 0\big) \leq \alpha g\big(1\big) + (1-\alpha) g\big(0\big) \\ &= \alpha f\big(x_1\big) + (1-\alpha) f\big(x_2\big). \end{aligned}$$

Таким образом, функция $f\big(x\big)$ выпукла.

■

Пример C5.7. Покажите, что функция $f : \mathbb{R}^d \to \overline{\mathbb{R}}$:

$$f\big(x\big) = -\sum_{i=1}^{n} \ln\Big(b_i - a_i^\top x\Big),\ A \in \mathbb{R}^{n\times d},\ b \in \mathbb{R}^n, \quad \operatorname{dom} f = \Big\{\, x \in \mathbb{R}^d \;\Big|\; a_i^\top x < b_i \,\Big\}$$

является выпуклой.

Доказательство. Покажем, что функция $g\big(\alpha\big)$:

$$g\big(\alpha\big) = -\sum_{i=1}^{n} \ln\Big(\big(b_i - a_i^\top x\big) - \alpha a_i^\top v\Big)$$

выпуклая. Покажем, что $-\ln(z)$ — выпуклая функция. Рассмотрим две точки x, y и проверим, что выполнено неравенство (C5.1):

$$-\ln(\alpha x + (1-\alpha)y) \leq -\alpha \ln x - (1-\alpha)\ln y.$$

Перепишем его:

$$x^\alpha y^{1-\alpha} \leq \alpha x + (1-\alpha)y.$$

и введём переменную $t = \frac{x}{y}$, тогда:

$$t^\alpha \leq \alpha t + 1 - \alpha.$$

Введем функцию $\phi(t)$:

$$\phi(t) = t^\alpha - \alpha t - 1 + \alpha.$$

Найдем её производную:

$$\phi'(t) = \alpha\big(t^{\alpha-1} - 1\big).$$

При $0 < t < 1$ производная положительная, при $t = 0$ производная равна нулю, при $t > 1$ производная отрицательная. Тогда, так как $\phi(1) = 0$, то неравенство выполняется. Покажем, что композиция выпуклой и аффинной функции — выпукла Пусть $h\big(x\big) = u(a(x))$, где u — выпуклая, a — аффинная. Применив Определение C5.6, получим

$$u(a(\lambda x + (1-\lambda)y)) = u(\lambda a(x) + (1-\lambda)a(y)) \leq \lambda u(a(x)) + (1-\lambda)u(a(y)).$$

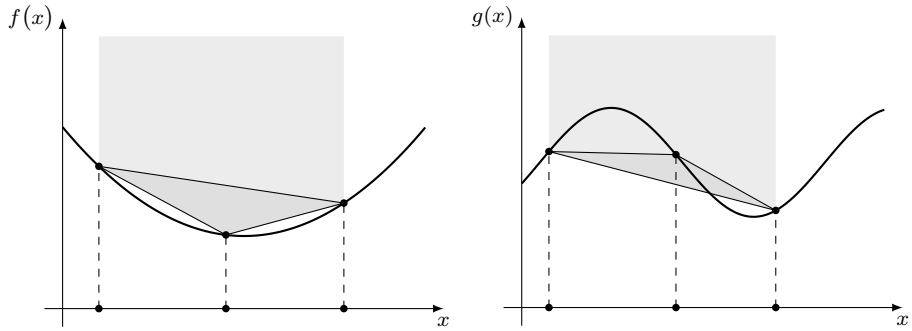
Сумма выпуклых функций выпукла по Определению C5.5, так как сумма выпуклых множеств — выпуклое. ■

Определение C5.6 можно обобщить на случай произвольного конечного числа точек.

Определение C5.7. Рассмотрим функцию $f : \mathbb{R}^d \to \overline{\mathbb{R}}$. Функция f называется *выпуклой*, если для любых $x_1, \ldots, x_k \in \mathbb{R}^d$ и любых $\alpha_1, \ldots, \alpha_k$, $\alpha_i \geq 0$, $\sum\limits_{i=1}^{k} \alpha_i = 1$ выполняется

$$f\left(\sum_{i=1}^{k} \alpha_i x_i\right) \leq \sum_{i=1}^{k} \alpha_i f\big(x_i\big). \tag{C5.2}$$

Замечание C5.6. Неравенство (C5.2) называется неравенством Йенсена.



(a) Выпуклая парабола.

(b) Не выпуклый синус.

Рис. C5.3: Выпуклость через условие Йенсена.

Замечание C5.7. С помощью данного определения не доказывают выпуклость функции, используют только в обратную сторону, так как есть определение для двух точек.

Замечание C5.8. Для любых фиксированных точек x_i неравенство (C5.2) переходит в равенство для любых α_i, удовлетворяющих условиям, тогда и только тогда, когда функция f является аффинной, или когда все точки x_i совпадают.

Утверждение C5.3. Определения C5.5, C5.6 и C5.7 эквивалентны.

Доказательство. Достаточно доказать эквивалентность нового определения какому-то из ранее введённых.

$\Leftarrow$ Покажем, что из Определения C5.6 следует Определение C5.7. Доказательство будем проводить по индукции. При $k = 1$ утверждение верно, а при $k = 2$ следует из Определения C5.6 выпуклой функции. Предположим, что неравенство (C5.7) верно для всех k и докажем его для $k + 1$. Будем считать, что $\alpha_{k+1} \in (0, 1)$, так как иначе все сводится к уже рассмотренным ранее случаям. Пусть

$$x = \sum_{i=1}^{k+1} \alpha_i x_i = \alpha_{k+1} x_{k+1} + \sum_{i=1}^{k} \alpha_i x_i = \alpha_{k+1} x_{k+1} + \big(1 - \alpha_{k+1}\big)\overline{x},$$

где

$$\overline{x} = \sum_{i=1}^{k} \frac{\alpha_i}{1 - \alpha_{k+1}} x_i,$$

В силу выпуклости функции $f(x)$ и предположения индукции

$$f\left(\sum_{i=1}^{k+1}\alpha_i x_i\right) = f\big(\alpha_{k+1}x_{k+1} + (1-\alpha_{k+1})\overline{x}\big) \le \alpha_{k+1}f(x_{k+1}) + (1-\alpha_{k+1})f(\overline{x})$$

$$\le \alpha_{k+1}f(x_{k+1}) + \sum_{i=1}^{k}\alpha_i f(x_i) = \sum_{i=1}^{k+1}\alpha_i f(x_i).$$

$\Leftarrow$ Покажем, что из Определения C5.7 следует Определение C5.6. Возьмём две произвольных точки и получим и неравенство (C5.1).

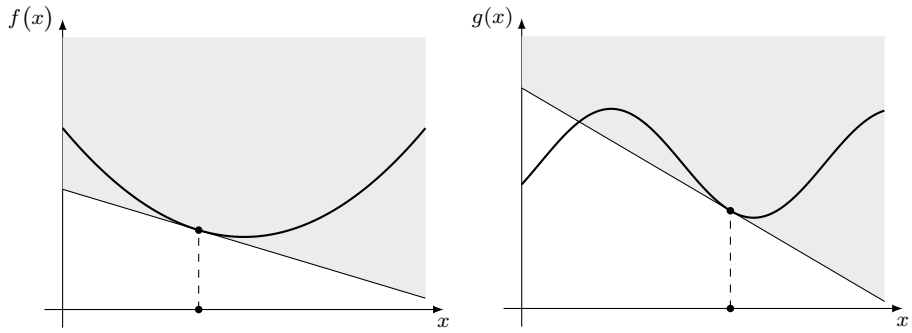
■

Помимо введённых ранее определений выпуклости, есть ещё одно, которое вводится для дифференцируемых функций.

Определение C5.8. Рассмотрим непрерывно дифференцируемую функцию $f : \mathbb{R}^d \to \mathbb{R}$, у которой $\operatorname{dom} f$ — выпуклое открытое множество. Функция f называется *выпуклой*, если выполняется

$$f(y) \ge f(x) + \langle \nabla f(x), y - x\rangle, \quad \forall x, y \in \operatorname{dom} f. \tag{C5.3}$$

Можно дать следующую геометрическую интерпретацию этому определению. Рассмотрим аффинную функцию $g(y) = f(x) + \langle \nabla f(x), y - x\rangle$ — касательную к графику функции $f(x)$ в точке x. Если функция выпуклая, то g является глобальной оценкой f снизу. Иными словами, график выпуклой функции лежит выше графика касательной, построенного в любой точке.



(a) Выпуклая парабола.

(b) Не выпуклый синус.

Рис. C5.4: Выпуклость через условие первого порядка.

Теорема C5.1. Определения C5.5, C5.6, C5.7 в случае дифференцируемой функции f и C5.8 эквивалентны.

Доказательство. Достаточно доказать эквивалентность нового определения какому-то из ранее введённых.

$\Rightarrow$ Покажем, что из Определения C5.6 следует Определение C5.8. Запишем неравенство (C5.1) из Определения C5.6

$$f\big(x+\alpha(y-x)\big) \le (1-\alpha)f\big(x\big) + \alpha f\big(y\big).$$

Поделив обе части неравенства на α, получим

$$f\big(y\big) \ge f\big(x\big) + \frac{f\big(x+\alpha(y-x)\big) - f\big(x\big)}{\alpha}.$$

Переходя к пределу при $\alpha \to 0$, получим требуемое утверждение.
$\Leftarrow$ Покажем, что из Определения C5.8 следует Определение C5.6. Выберем произвольные x, y и обозначим $z = \alpha x + (1-\alpha)y$. Воспользуемся неравенством (C5.3):

$$f\big(x\big) \ge f\big(z\big) + \big\langle \nabla f\big(z\big), x - z\big\rangle,$$
$$f\big(y\big) \ge f\big(z\big) + \big\langle \nabla f\big(z\big), y - z\big\rangle.$$

Сложив первое неравенство, умноженное на α, и второе, умноженное на $(1-\alpha)$, получим

$$\alpha f\big(x\big) + (1-\alpha) f\big(y\big) \ge f\big(z\big) + \big\langle \nabla f\big(z\big), \alpha x + (1-\alpha)y - z\big\rangle = f\big(z\big).$$

■

Пример C5.8. Покажите, что функция $f : \mathbb{R}^d \to \overline{\mathbb{R}}$:

$$f\big(x\big) = \sum_{i=1}^{d} \frac{1}{x_i}, \quad \operatorname{dom} f = \mathbb{R}^d_{++}.$$

является выпуклой.

Решение. Проверим неравенство (C5.3) для точек x, y:

$$\sum_{i=1}^{d} \frac{1}{y_i} \ge \sum_{i=1}^{d} \frac{1}{x_i} - \sum_{i=1}^{d} \frac{y_i - x_i}{x_i^2}$$

Перепишем:

$$\sum_{i=1}^{d} \frac{x_i^2 - x_i y_i + y_i^2 - x_i y_i}{x_i^2 y_i} = \sum_{i=1}^{d} \frac{(x_i - y_i)^2}{x_i^2 y_i} \ge 0.$$

■

Теорема C5.2 (Дифференциальное условие оптимальности для выпуклой функции). Пусть $f : \mathbb{R}^d \to \overline{\mathbb{R}}$ — собственная выпуклая функция, $\operatorname{dom} f$ является открытым множеством, и пусть $x^* \in \operatorname{dom} f$. Тогда x^* — глобальный минимум функции f, тогда и только тогда, когда $\nabla f\big(x^*\big) = 0$.

Доказательство. Следует из Определения C5.8. ■

Наиболее часто в упражнениях на проверку выпуклости мы будем вычислять гессиан функции и проверять, является ли он определённым. Это один из самых мощных подходов.

Теорема C5.3 (Критерий выпуклости второго порядка). Рассмотрим собственную дважды непрерывно дифференцируемую функцию $f : \mathbb{R}^d \to \mathbb{R}$, у которой $\operatorname{dom} f$ — выпуклое открытое множество. Функция f выпукла тогда и только тогда, когда

$$\nabla^2 f(x) \succeq 0, \quad \forall x \in \operatorname{dom} f.$$

Доказательство. $\Rightarrow$ По Определению C5.8 для любых $x, y \in \operatorname{dom} f$:

$$f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle,$$
$$f(x) \geq f(y) + \langle \nabla f(y), x - y \rangle.$$

Сложим и получим:

$$\langle \nabla f(x) - \nabla f(y), x - y \rangle \geq 0.$$

Возьмём x, направление v и $y = x + hv$:

$$\left\langle \frac{\nabla f(x + hv) - \nabla f(x)}{h}, v \right\rangle \geq 0.$$

Переходя к пределу по h, получим $v^\top \nabla^2 f(x) v \geq 0$ для любого v, то есть $\nabla^2 f(x) \succeq 0$.
$\Leftarrow$ Возьмём любые $x, y \in \operatorname{dom} f$ и рассмотрим $g(t) = f(x + t(y - x))$, тогда

$$g'(t) = \langle \nabla f(x + t(y - x)), y - x \rangle$$

По формуле Ньютона–Лейбница:

$$g(1) - g(0) = \int_0^1 g'(t)\,\mathrm{d}t = \langle \nabla f(x), y - x \rangle + \int_0^1 \int_0^t (y - x)^\top \nabla^2 f(x + s(y - x))(y - x)\,\mathrm{d}s\,\mathrm{d}t\,.$$

То есть получаем, что

$$f(y) - f(x) \geq \langle \nabla f(x), y - x \rangle.$$

■

Пример C5.9. Покажите, что функции $f : \mathbb{R} \to \overline{\mathbb{R}}$:

- $f(x) = \exp(ax)$ выпукла на $\mathbb{R}$ для любого $a \in \mathbb{R}$;
- $f(x) = -\ln x$ выпукла на $\mathbb{R}_{++}$;
- $f(x) = x \ln x$ выпукла на $\mathbb{R}_{++}$;
- $f(x) = x^p$ для $p \leq 0$ или $p \geq 1$ выпукла на $\mathbb{R}_{++}$

являются выпуклыми.

Решение.

- $f''(x) = a^2 \exp(ax) \geq 0$ на $\mathbb{R}$;
- $f''(x) = \frac{1}{x^2} \geq 0$ на $\mathbb{R}_{++}$;
- $f''(x) = \frac{1}{x} \geq 0$ на $\mathbb{R}_{++}$;
- $f''(x) = p(p-1)x^{p-2} \geq 0$ для $p \leq 0$ или $p \geq 1$ на $\mathbb{R}_{++}$;

■

Пример C5.10. Когда функция $f : \mathbb{R}^2 \to \overline{\mathbb{R}}$:

$$f\big(x\big) = ax_1^2 + bx_1x_2 + cx_2^2$$

является выпуклой?

Решение. Найдем гессиан:

$$\nabla^2 f\big(x\big) = \begin{pmatrix} 2a & b \\ b & 2c \end{pmatrix}.$$

Согласно критерию Сильвестра эта матрица неотрицательно определена, если $a \geq 0$, $c \geq 0$, $4ac \geq b^2$. ■

Пример C5.11. Покажите с помощью Определения C5.6 неравенство между средним арифметическим и средним геометрическим для двух переменных:

$$\sqrt{ab} \leq \frac{a+b}{2},$$

где $a, b \geq 0$.

Решение. Рассмотрим функцию выпуклую $f\big(x\big) = -\ln x$. Полагая в (C5.6) $\alpha = \frac{1}{2}$, получаем

$$-\ln\left(\frac{a+b}{2}\right) \leq \frac{-\ln a - \ln b}{2}.$$

Возьмём экспоненты от левой и правой части и получим ответ. ■

Пример C5.12. Когда функция $f : \mathbb{R}^d \to \overline{\mathbb{R}}$:

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

$A \in \mathbb{S}^d, b \in \mathbb{R}^d$ является выпуклой?

Решение. Заметим, что $\nabla^2 f\big(x\big) = A$. Следовательно, $f\big(x\big)$ является выпуклой тогда и только тогда, когда $A \succeq 0$. ■

Пример C5.13. Покажите, что функция $f : \mathbb{R}^d \to \overline{\mathbb{R}}$:

$$f\big(x\big) = \ln \sum_{i=1}^{d} e^{x_i}$$

является выпуклой.

Решение. Найдем, градиент:

$$\nabla f(x) = \left(\sum_{i=1}^{d} e^{x_i}\right)^{-1} (e^{x_1}, \ldots, e^{x_d})^\top = \frac{1}{\mathbf{1}^\top z} z,$$

где $z = (e^{x_1}, \ldots, e^{x_d})^\top$. Тогда гессиан:

$$\nabla^2 f(x) = \frac{1}{(\mathbf{1}^\top z)^2}\left((\mathbf{1}^\top z)\operatorname{diag}(z) - zz^\top\right).$$

Покажем, что для произвольного вектора v выполняется неравенство $v^\top \nabla^2 f(x) v \geq 0$:

$$v^\top \nabla^2 f(x) v = \frac{1}{(\mathbf{1}^\top z)^2}\left(\left(\sum_{i=1}^{d} z_i\right)\left(\sum_{i=1}^{d} v_i^2 z_i\right) - \left(\sum_{i=1}^{d} v_i z_i\right)^2\right) \geq 0.$$

Последнее неравенство следует из неравенства Коши-Буняковского-Шварца:

$$(a^\top a)(b^\top b) \geq (a^\top b)^2,$$

где $a_i = \sqrt{z_i}$, $b_i = v_i\sqrt{z_i}$. ■

C5.3 Операции, сохраняющие выпуклость

В этом параграфе рассмотрим преобразования, которые сохраняют выпуклость функций. Такие операции позволяют из простых выпуклых функций строить более сложные выпуклые функции.

C5.3.1 Неотрицательная взвешенная сумма

Сумма двух выпуклых функций выпукла, так как сумма выпуклых множеств — выпуклое. Также, если умножить выпуклую функцию на неотрицательное число c, то полученная функция $cf(x)$ также выпукла. Комбинируя эти свойства, получаем

Утверждение C5.4. Рассмотрим выпуклые функции $f_1, \ldots, f_n : \mathbb{R}^d \to \mathbb{R}$ и $c_1, \ldots, c_n \in \mathbb{R}_+$. Тогда функция $f : \mathbb{R}^d \to \mathbb{R}$:

$$f(x) = \sum_{i=1}^{n} c_i f_i(x)$$

является выпуклой.

Пример C5.14. Покажите, что функция $f : \mathbb{R}^d \to \overline{\mathbb{R}}$:

$$f(x) = \sum_{i=1}^{d} i^2 e^{x_i}$$

является выпуклой.

Решение. Функция выпуклая, как сумма выпуклых функций с положительными коэффициентами. ■

C5.3.2 Аффинная подстановка аргумента

Утверждение C5.5. Рассмотрим выпуклую функцию $f : \mathbb{R}^d \to \mathbb{R}$ и $A \in \mathbb{R}^{d\times n}$, $b \in \mathbb{R}^d$. Тогда функция $g : \mathbb{R}^d \to \mathbb{R}$:

$$g(x) = f(Ax + b)$$

является выпуклой.

Пример C5.15. Покажите, что функция $f : \mathbb{R}^d \to \overline{\mathbb{R}}$:

$$f(x) = \sum_{i=1}^{n} c_i \exp\left(a_i^\top x + b_i\right),\ a \in \mathbb{R}^{n\times d},\ b \in \mathbb{R}^n,\ c \in \mathbb{R}^n_+.$$

является выпуклой.

Решение. Функция является выпуклой, как взвешенная сумма экспонент с аффинной подстановкой аргумента. ■

C5.3.3 Поточечный максимум и супремум

Пересечение эпиграфов двух выпуклых функций $f_1(x)$ и $f_2(x)$, является выпуклым множеством. Тогда и функция $f : \mathbb{R}^d \to \overline{\mathbb{R}}$:

$$f(x) = \max\{f_1(x), f_2(x)\}$$

является выпуклой. Но пересекая не только два, а произвольное число выпуклых множеств, мы опять получаем выпуклое множество. Поэтому если функция двух аргументов $g(x, y)$ выпукла по x для любого y, то следующая функция

$$f(x) = \sup_{(\cdot,y)\in \operatorname{dom} g} g(x, y)$$

так же является выпуклой.

Утверждение C5.6. Рассмотрим выпуклые функции $f_1, \ldots, f_n : \mathbb{R}^d \to \mathbb{R}$. Тогда функция $f : \mathbb{R}^d \to \mathbb{R}$:

$$f(x) = \max\{f_1(x), \ldots, f_n(x)\}$$

является выпуклой.
Рассмотрим выпуклую по x функцию $g(x, y) : \mathbb{R}^{d+n} \to \mathbb{R}$. Тогда функция $f : \mathbb{R}^d \to \mathbb{R}$:

$$f(x) = \sup_{(\cdot,y)\in \operatorname{dom} g} g(x, y)$$

является выпуклой.

Пример C5.16. Покажите, что кусочно-линейная функция $f : \mathbb{R}^d \to \mathbb{R}$:

$$f(x) = \max\left\{ a_1^\top x + b_1, \ldots, a_n^\top x + b_n \right\},\ a \in \mathbb{R}^{n\times d},\ b \in \mathbb{R}^d$$

является выпуклой.

Решение. Функция является выпуклой, как поточеченый максимум линейных функций. ■

Пример C5.17. Обозначим $x_{[i]}$ i-ю максимальную координату вектора $x \in \mathbb{R}^d$. Покажите, что функция $f : \mathbb{R}^d \to \mathbb{R}$:

$$f\big(x\big) = \sum_{i=1}^{r} x_{[i]}$$

является выпуклой.

Решение. Запишем функцию $f\big(x\big)$ как

$$f\big(x\big) = \sum_{i=1}^{r} x_{[i]} = \max\left\{\, x_{i_1} + \ldots + x_{i_r} \mid 1 \leq i_1 < i_2 < \ldots < i_r \leq d \,\right\},$$

то есть максимум из всех возможных сумм r различных координат вектора x. Видно, что $f\big(x\big)$ — это поточечный максимум линейных функций, следовательно, $f\big(x\big)$ — выпуклая функция. ■

Пример C5.18. Покажите, что опорная функция $S_C : \mathbb{R}^d \to \mathbb{R}$:

$$S_C(y) = \sup_{x \in C} \big\langle y, x \big\rangle,\ C \subseteq \mathbb{R}^d$$

является выпуклой.

Решение. Функция является выпуклой, как поточечный супремум от аффинных функций. ■

Пример C5.19. Покажите, что функция $f : \mathbb{R}^d \to \mathbb{R}$ расстояния от точки x до самой удалённой точки множества C:

$$f\big(x\big) = \sup_{y \in C} \big\|x - y\big\|,\ C \subseteq \mathbb{R}^d$$

является выпуклой.

Решение. Функция является выпуклой, как поточечный супремум от выпуклых функций. ■

C5.3.4 Монотонная суперпозиция

Утверждение C5.7. Рассмотрим выпуклую функции $h : \mathbb{R}^d \to \mathbb{R}$ и выпуклую неубывающая функцию $g : \mathbb{R} \to \mathbb{R}$. Тогда композиция этих двух функций $f\big(x\big) = g\big(h\big(x\big)\big)$ является выпуклой функцией.

Доказательство. Возьмём произвольные точки x_1, x_2. В силу выпуклости функции $h\big(x\big)$ выполняется неравенство

$$h\big(\alpha x_1 + (1-\alpha)x_2\big) \leq \alpha h\big(x_1\big) + (1-\alpha)h\big(x_2\big).$$

Но внешняя функция $g\big(y\big)$ является неубывающей и выпуклой, поэтому

$$\begin{aligned} f\big(\alpha x_1 + (1-\alpha)x_2\big) &= g\big(h\big(\alpha x_1 + (1-\alpha)x_2\big)\big) \leq g\big(\alpha h\big(x_1\big) + (1-\alpha)h\big(x_2\big)\big) \\ &\leq \alpha g\big(h\big(x_1\big)\big) + (1-\alpha)g\big(h\big(x_2\big)\big) = \alpha f\big(x_1\big) + (1-\alpha)f\big(x_2\big). \end{aligned}$$

■

Пример C5.20. Рассмотрим выпуклую $g : \mathbb{R}^d \to \mathbb{R}$. Покажите, что функция $f : \mathbb{R}^d \to \mathbb{R}$:

$$f(x) = \exp g(x)$$

является выпуклой.

Решение. Функция является выпуклой, как композиция выпуклой и выпуклой неубывающей функции. ■

Пример C5.21. Покажите, что функция $f : \mathbb{R}^d \to \mathbb{R}$:

$$f(x) = \|x\|^p$$

является выпуклой при $p \geq 1$.

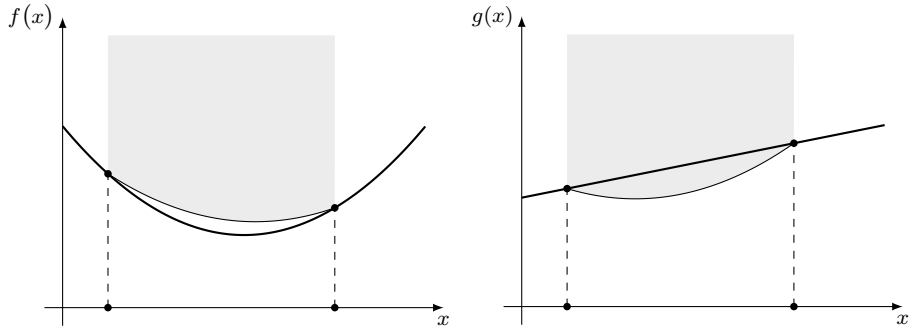
Решение. Функция является выпуклой, как композиция выпуклой и выпуклой неубывающей функции. ■

C5.4 Сильно выпуклые функции

Рассмотрим ещё один класс функций — μ-сильно выпуклые функции. Начнём с определения через неравенство Йенсена для двух точек.

Определение C5.9. Рассмотрим функцию $f : \mathbb{R}^d \to \overline{\mathbb{R}}$. Функция f называется *μ-сильно выпуклой*, если для любых $x, y \in \mathbb{R}^d$ и любого $\alpha \in [0, 1]$ выполняется

$$f\big(\alpha x + (1-\alpha)y\big) \leq \alpha f\big(x\big) + (1-\alpha) f\big(y\big) - \alpha(1-\alpha)\frac{\mu}{2}\|x-y\|_2^2. \qquad \text{(C5.4)}$$



(a) μ-сильно выпуклая парабола. (b) Не сильно выпуклая прямая.

Рис. C5.5: μ-сильная выпуклость через Йенсена для двух точек.

Пример C5.22. Покажите, что функция $f : \mathbb{R}^d \to \overline{\mathbb{R}}$:

$$f(x) = \ln \sum_{i=1}^{d} e^{x_i} + \|x\|_2^2$$

является 2-сильно выпуклой.

Решение. Выпуклость $\ln \sum_{i=1}^{d} e^{x_i}$ была показана в Примере C5.13. Покажем, что $\|x\|_2^2$ 2-сильно выпукла. Запишем неравенство (C5.4) для точек x, y:

$$\|\alpha x + (1-\alpha)y\|_2^2 \leq \alpha\|x\|_2^2 + (1-\alpha)\|y\|_2^2 - \alpha(1-\alpha)\|x-y\|_2^2.$$

Перепишем:

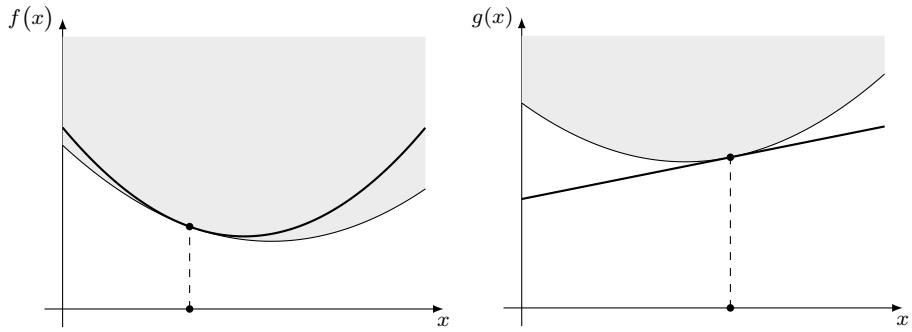
$$\alpha(1-\alpha)\|x-y\|_2^2 \leq \alpha(1-\alpha)\|x\|_2^2 - 2\alpha(1-\alpha)\langle x, y\rangle + \alpha(1-\alpha)\|y\|_2^2.$$

Получили, что неравенство верное. ■

Есть ещё одно определение, которое вводится для дифференцируемых функций.

Определение C5.10. Рассмотрим непрерывно дифференцируемую функцию $f : \mathbb{R}^d \to \mathbb{R}$, у которой $\operatorname{dom} f$ — выпуклое открытое множество. Функция f называется *μ-сильно выпуклой*, если выполняется

$$f(y) \geq f(x) + \langle \nabla f(x), y - x\rangle + \frac{\mu}{2}\|y-x\|_2^2, \quad \forall x, y \in \operatorname{dom} f. \tag{C5.5}$$



(a) μ-сильно выпуклая парабола.

(b) Не сильно выпуклая прямая.

Рис. C5.6: μ-сильная выпуклость через условие первого порядка.

Покажем эквивалентность определений.

Теорема C5.4. Определения C5.9 в случае дифференцируемой функции f и C5.10 эквивалентны.

Доказательство. $\Rightarrow$ Покажем, что из Определения C5.9 следует Определение C5.10.

Запишем неравенство (C5.4) из Определения C5.9

$$f\big(x+\alpha(y-x)\big) \le (1-\alpha)f\big(x\big) + \alpha f\big(y\big) - \alpha(1-\alpha)\frac{\mu}{2}\big\|x-y\big\|_2^2.$$

Поделив обе части неравенства на α, получим

$$f\big(y\big) \ge f\big(x\big) + \frac{f\big(x+\alpha(y-x)\big) - f\big(x\big)}{\alpha} + (1-\alpha)\frac{\mu}{2}\big\|x-y\big\|_2^2.$$

Переходя к пределу при $\alpha \to 0$, получим требуемое утверждение.

$\Leftarrow$ Покажем, что из Определения C5.10 следует Определение C5.9. Выберем произвольные x, y и обозначим $z = \alpha x + (1-\alpha)y$. Воспользуемся неравенством (C5.5):

$$f\big(x\big) \ge f\big(z\big) + \big\langle \nabla f\big(z\big), x - z\big\rangle + \frac{\mu}{2}\big\|x-z\big\|_2^2,$$
$$f\big(y\big) \ge f\big(z\big) + \big\langle \nabla f\big(z\big), y - z\big\rangle + \frac{\mu}{2}\big\|x-z\big\|_2^2.$$

Сложив первое неравенство, умноженное на α, и второе, умноженное на $(1-\alpha)$, получим

$$\begin{aligned}
\alpha f\big(x\big) + (1-\alpha)f\big(y\big) &\ge f\big(z\big) + \big\langle \nabla f\big(z\big), \alpha x + (1-\alpha)y - z\big\rangle \\
&\quad + \frac{\mu}{2}\Big(\alpha\big\|x-z\big\|_2^2 + (1-\alpha)\big\|y-z\big\|_2^2\Big) \\
&\ge f\big(z\big) + \alpha(1-\alpha)\frac{\mu}{2}\Big(\big\|x-z\big\|_2^2 + \big\|y-z\big\|_2^2\Big) \\
&\ge f\big(z\big) + \alpha(1-\alpha)\frac{\mu}{2}\big\|x-y\big\|_2^2.
\end{aligned}$$

■

Пример C5.23. Получите условие, при котором квадратичная функция $f : \mathbb{R}^d \to \overline{\mathbb{R}}$:

$$f\big(x\big) = \frac{1}{2}x^\top Ax + b^\top x$$

является сильно выпуклой.

Решение. Напомним, что $\nabla f\big(x\big) = Ax + b$. Тогда

$$\begin{aligned}
f\big(y\big) - f\big(x\big) - \big\langle \nabla f\big(x\big), y - x\big\rangle &= \frac{1}{2}y^\top Ay - \frac{1}{2}x^\top Ax - x^\top A(y-x) \\
&= \frac{1}{2}\,(y-x)^\top A(y-x).
\end{aligned}$$

Пользуясь Определением C5.10, получим условие, при котором $f\big(x\big)$ μ-сильно выпукла:

$$(y-x)^\top A(y-x) \ge \mu\big\|y-x\big\|_2^2 \implies (y-x)^\top (A-\mu I)(y-x) \ge 0.$$

Таким образом, должно выполняться $A \succeq \mu I_d$. ■

Теорема C5.5 (Критерий μ-сильно выпуклости второго порядка). Рассмотрим собственную дважды непрерывно дифференцируемую функцию $f : \mathbb{R}^d \to \mathbb{R}$, у которой $\operatorname{dom} f$ — выпуклое открытое множество. Функция μ-сильно

выпукла выпукла тогда и только тогда, когда

$$\nabla^2 f(x) \succeq \mu I_d, \ \forall x \in \operatorname{dom} f.$$

Доказательство. Если g — μ-сильно выпуклая функция, то $g(x) = f(x) - \frac{\mu}{2}\|x\|_2^2$ является выпуклой. Действительно, воспользуемся Определением С5.10:

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

Заметим, что $\|x\|_2^2 = \|y\|_2^2 + 2\langle y, x - y\rangle + \|x - y\|_2^2$ и получим

$$\begin{aligned} g(x) &\geq f(y) + \frac{\mu}{2}\|y\|_2^2 + \langle \nabla f(y) - \mu y, x - y\rangle + \Big(\frac{\mu}{2} - \frac{\mu}{2}\Big)\|x - y\|_2^2 \\ &= g(y) + \langle \nabla g(y), x - y\rangle. \end{aligned}$$

Подставим $g(x)$ в критерий выпуклости второго порядка и получим $\nabla^2 f(x) \succeq \mu I_d$. ■

Замечание С5.9. Константу сильной выпуклости можно искать как

$$\mu = \lambda_{\min}\big(\nabla^2 f(x)\big).$$

Замечание С5.10. Рассмотрим L-гладкую собственную дважды непрерывно дифференцируемую функцию $f : \mathbb{R}^d \to \mathbb{R}$. Запишем для неё L-гладкость (Определение Л2.5):

$$\big\|\nabla f(x) - \nabla f(y)\big\|_2 \leq L\|x - y\|_2 \implies \frac{\big\|\nabla f(x) - \nabla f(y)\big\|_2}{\|x - y\|_2} \leq L.$$

При $y \to x$ получаем: $\big\|\nabla^2 f(x)\big\|_2 \leq L$, то есть $\max_i |\lambda_i(\nabla^2 f(x))| \leq L$. Если функция $f(x)$ является ещё и μ-сильно выпуклой, то:

$$\mu I_d \preceq \nabla^2 f(x) \preceq L I_d.$$

C6 Субдифференциал

В приложениях теории оптимизации порой приходится иметь дело с негладкой целевой функцией. Например, когда она задана как поточечный максимум по конечному множеству индексов. В таком случае требуется обобщить определение производной для функций, не дифференцируемых в некоторых точках. Понятие субдифференциала позволяет формализовать условия оптимальности и строить методы, аналогичные градиентным, даже когда градиент в привычном смысле не существует.
В рамках этой главы мы вновь будем рассматривать функции, расширенные на всё пространство.

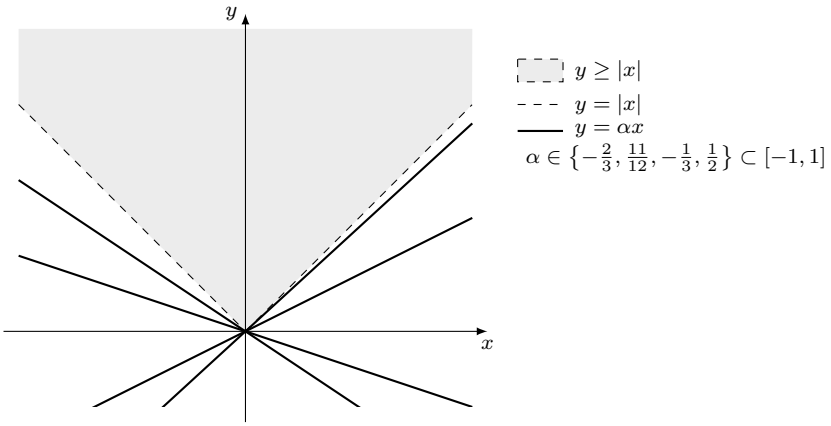
Определение C6.1. Пусть в евклидовом пространстве V определена функция $f : V \to \overline{\mathbb{R}}$. *Субградиентом функции f в точке $x_0 \in \operatorname{dom} f$* называется вектор $g \in V$ такой, что

$$f(x) \geq f(x_0) + \langle g, x - x_0 \rangle, \ \forall x \in V.$$

То есть субградиент в точке x_0 определяет гиперплоскость в пространстве V, которая подпирает снизу надграфик функции f в любой точке. Таким образом, для функций, определённых на множестве действительных чисел, субградиент достаточно легко находить графически. Рассмотрим соответствующий пример.

Пример C6.1. Найдите субдифференциал $f(x) = |x|$ в точке $x = 0$.

Решение. 1. Графический способ.
Рассмотрим график функции $f(x) = |x|$ и заметим, что любая прямая $y = ax$, где $a \in [-1, 1]$, проходит через точку $(0, 0)$ и в любой другой точке проходит не выше, чем $y = |x|$. Причём, если $|a| > 1$, то $y = ax$ уже лежит выше $y = |x|$. Таким образом $\partial f(0) = [-1, 1]$.



2. По определению.
$f(x) = |x|$ и $f(0) = 0$. По определению вектор $g \in \mathbb{R}$ является субградиентом для функции $f(x) = |x|$ в точке 0, если $\forall x \in \mathbb{R}\ |x| \geq gx$. Решая данное неравенство, мы получаем, что оно выполняется для всех $g \in [-1, 1]$ и только для таких g. ■

Определение C6.2. Пусть в евклидовом пространстве V определена функция $f : V \to \overline{\mathbb{R}}$. Множество всех субградиентов функции в точке $x_0 \in \operatorname{dom} f$ называется *субдифференциалом функции f в точке x_0* и обозначается $\partial f\big(x_0\big)$:

$$\partial f\big(x_0\big) = \left\{\, g \in V \;\middle|\; f\big(x\big) \geq f\big(x_0\big) + \big\langle g, x - x_0 \big\rangle,\ \forall x \in \operatorname{dom} f \,\right\}.$$

Для $x_0 \notin \operatorname{dom} f$ будем полагать $\partial f\big(x_0\big) = \varnothing$.

Замечание C6.1. Если субдифференциал в некоторой точке содержит хотя бы один элемент, то функция называется *субдифференцируемой* в этой точке.

Замечание C6.2. Из определений выше можно заметить, что субградиент зависит от введённого скалярного произведения. Поэтому при нахождении субдифференциала важно указывать, как задано скалярное произведение. Мы же далее будем рассматривать стандартное скалярное произведение, если не оговорено другое.

Замечание C6.3. Понятие субградиента можно обобщить для функций $f : V \to \overline{\mathbb{R}}$, где V не обязательно евклидово (т.е. скалярное произведение на этом пространстве может быть не определено). Тогда для функции f можно определить субградиент $g \in V^*$ как элемент сопряжённого пространства к V (т.е. пространства всевозможных линейных непрерывных функционалов на V), при этом операция $\big\langle g, x \big\rangle$ интерпретируется как действие линейного функционала g на элемент x.

Пример C6.2. Рассмотрите отображение $f : \mathbb{S}^d \to \mathbb{R}$ на пространстве симметричных матриц, определенное как

$$f\big(X\big) = \lambda_{\max}(X).$$

Вычислите аналитически субградиент $g \in \partial f\big(X\big)$.

Решение. Пусть v — нормированный собственный вектор, соответствующий максимальному собственному числу матрицы X. Рассмотрим еще одну матрицу $Y \in \mathbb{S}^d$. Заметим, что

$$\lambda_{\max}(Y) = \max_{\|u\|=1} \big\{u^\top Y u\big\} \geq v^\top Y v.$$

Добавим $v^\top X v$, в качестве умного нуля к правой части неравенства. Получим

$$\lambda_{\max}(Y) \geq v^\top X v + v^\top (Y - X) v = \lambda_{\max}(X) + \operatorname{Tr}\Big(v v^\top (Y - X)\Big).$$

Осталось заметить, что $\operatorname{Tr}\big(v v^\top (Y - X)\big) = \big\langle v v^\top, Y - X \big\rangle$. Таким образом, $v v^\top \in \partial f\big(X\big)$. ■

Следующий пример чуть более прикладной, его суть станет в полной мере понятна на одной из следующих глав.

Пример C6.3. Пусть заданы функции $f : \mathbb{R}^d \to \mathbb{R}$, $h : \mathbb{R}^d \to \mathbb{R}^n$. Рассмотрите функцию

$$d(\lambda) = -\min_{x \in \mathbb{R}^d} \big\{ f\big(x\big) + \lambda^\top h\big(x\big) \big\}.$$

Будем считать f такой, что d — выпуклая на

$$\operatorname{dom} d = \left\{\, \lambda \in \mathbb{R}^n \;\middle|\; \lambda > 0,\ d(\lambda) > -\infty \,\right\}.$$

Вычислите аналитически субградиент $g \in \partial d(\lambda)$.

Решение. Рассмотрим $\lambda_0 \in \operatorname{dom} d$ и пусть

$$\min_{x \in \mathbb{R}^d} \{f(x) + \lambda_0^\top h(x)\}$$

достигается в точке $x_0 \in \mathbb{R}^d$. Тогда распишем:

$$\begin{aligned} d(\lambda) &= -\min_{x \in \mathbb{R}^d} \{f(x) + \lambda^\top h(x)\} = -f(x_0) - \lambda^\top h(x_0) \\ &= -f(x_0) - \lambda_0^\top h(x_0) + (\lambda_0 - \lambda)^\top h(x_0) \\ &= d(\lambda_0) + \langle -h(x_0), \lambda - \lambda_0 \rangle. \end{aligned}$$

Таким образом, $-h(x_0) \in \partial f(x_0)$. ■

Важнейшим примером является субдифференциал произвольной нормы $\|\cdot\|$. Известно, что норма — выпуклая функция, дифференцируемая всюду кроме нуля. Таким образом, ноль — единственная точка, в которой субдифференциал не одноточечное множество.

Пример С6.4. Пусть $\|\cdot\|$ — произвольная норма в евклидовом пространстве V и пусть $\|\cdot\|_*$ — сопряжённая к ней норма. Рассмотрите $f(x) = \|x\|$ и покажите, что субградиент равен:

$$\partial f(x) = \begin{cases} \mathcal{B}_{\|\cdot\|_*}(0,1), & x = 0, \\ \frac{x}{\|x\|}, & x \neq 0. \end{cases}$$

где $\mathcal{B}_{\|\cdot\|_*}(0,1) = \{\, g \in V \mid \|g\|_* \leq 1 \,\}$ — единичный шар в сопряжённой норме с центром в точке 0.

Решение. Пусть $g \in V$ и g является субградиентом $f(x) = \|x\|$ в точке x_0. Тогда верно неравенство

$$\|x\| \geq \|x_0\| + \langle g, x - x_0 \rangle \ \forall x \in V.$$

Перепишем данное неравенство

$$\langle g, x \rangle - \|x\| \leq \langle g, x_0 \rangle - \|x_0\| \ \forall x \in V.$$

Так как данное неравенство верно $\forall x \in V$, то оно верно и для супремума левой части по множеству V:

$$\sup_{x \in V} \{\langle g, x \rangle - \|x\|\} \leq \langle g, x_0 \rangle - \|x_0\|.$$

Но так как в точке x_0 данное неравенство переходит в равенство, то супремум левой части достигается в точке x_0. Поэтому

$$\sup_{x \in V} \{\langle g, x \rangle - \|x\|\} = \langle g, x_0 \rangle - \|x_0\|.$$

Теперь попробуем вычислить супремум аналитически. Предположим сначала, что $\|g\|_* > 1$. Тогда по определению сопряжённой нормы существует такой $\hat{x}$, что $\|\hat{x}\| \leq 1$ и $\langle g, \hat{x} \rangle > 1$. Возьмем $x = t\hat{x}$. Тогда

$$\langle g, x \rangle - \|x\| = t(\langle g, \hat{x} \rangle - \|\hat{x}\|) \to \infty, \ t \to \infty,$$

поскольку выражение в скобках положительно, то есть супремум неограничен. Предположим теперь, что $\|g\|_* \le 1$. Тогда, пользуясь неравенством Гёльдера, запишем

$$\langle g, x\rangle - \|x\| \le \|x\|\big(\|g\|_* - 1\big) \le 0.$$

Таким образом, имеем два условия на субдифференицал в точке $x_0 \neq 0$:

$$\|g\|_* \le 1,\ \langle g, x_0\rangle = \|x_0\|.$$

Это соответствует единственному решению

$$g = \frac{x_0}{\|x_0\|}.$$

Если $x_0 = 0$, то остаёся условие:

$$\|g\|_* \le 1.$$

Итого, доказали требуемое. ■

C6.1 Свойства субдифференциала

С точки зрения численных методов, в первую очередь, субдифференциал должен состоять хотя бы из одного элемента, чтобы было возможно сделать эффективный шаг. Во-вторых, для устойчивой сходимости хотелось бы уметь гарантировать, что выбранный субградиент не может быть неограниченно большим. Таким образом, существуют два критических вопроса: непустота субдифференциала и его ограниченность. Ниже мы сформулируем несколько достаточных условий, начиная с более сильных и постепенно ослабляя их.

Теорема C6.1. Пусть в евклидовом пространстве V определена собственная выпуклая функция $f : V \to \overline{\mathbb{R}}$, $\operatorname{dom} f$ — выпуклое множество. Если $x_0 \in \operatorname{int}\operatorname{dom} f$, то $\partial f\big(x_0\big)$ — непустое ограниченное множество.

Доказательство. Рассмотрим точку $\big(x_0, f\big(x_0\big)\big)$. По определению надграфика $\operatorname{epi} f$, она находится на его границе. Тогда по теореме отделимости существует ненулевой вектор $(p, -\lambda)$, такой что

$$\langle p, x_0\rangle - \lambda f\big(x_0\big) \ge \langle p, x\rangle - \lambda t,\ \forall (x,t) \in \operatorname{epi} f. \tag{C6.1}$$

Здесь мы отделяем выпуклое множество от точки, не принадлежащей его внутренности. Заметим, что $\big(x_0, f\big(x_0\big) + 1\big) \in \operatorname{epi} f$. В совокупности с теоремой отделимости это означает

$$\langle p, x_0\rangle - \lambda f\big(x_0\big) \ge \langle p, x_0\rangle - \lambda(f\big(x_0\big) + 1),$$

откуда следует неотрицательность λ. Из курса математического анализа известно, что из выпуклости f и условия $x_0 \in \operatorname{int}\operatorname{dom} f$ следует существование окрестности $\mathcal{B}_\varepsilon(x_0) \subset \operatorname{int}\operatorname{dom} f$, такой что

$$\big|f\big(\hat{x}\big) - f\big(x_0\big)\big| \le M\big\|\hat{x} - x_0\big\|,\ \forall \hat{x} \in \mathcal{B}_\varepsilon(x_0).$$

Резюмируя, запишем

$$\langle p, \hat{x} - x_0\rangle \le \lambda\hat{t} - \lambda f\big(x_0\big) \le \lambda M\big\|\hat{x} - x_0\big\|.$$

По следствию из теоремы Рисса-Фреше выберем такое направление $\hat{p}$, что $\|\hat{p}\| = 1$, $\langle p, \hat{p}\rangle = \|p\|_*$ и $x_0 + \varepsilon\hat{p} \in \mathcal{B}_\varepsilon(x_0)$. Положим $\hat{x} = x_0 + \varepsilon\hat{p}$. Тогда

$$\varepsilon\|p\|_* \le \lambda M\varepsilon.$$

Ранее мы уже установили $\lambda \ge 0$. Положим теперь $\lambda = 0$. Тогда автоматически получим $p = 0$, что находится в противоречии с теоремой об отделимости, поскольку $(p, -\lambda)$ — ненулевой вектор. Таким образом, $\lambda > 0$. Выбирая $t = f(x)$ в (C6.1) и осуществляя деление на λ, получаем

$$f(x) \ge f(x_0) + \Big\langle \frac{p}{\lambda}, x - x_0 \Big\rangle.$$

Это доказывает не пустоту субдифференциала. Чтобы показать неограниченность, для произвольного субградиента $g \in \partial f(x_0)$ выберем направление $\hat{g}$, такое что $\|\hat{g}\| = 1$ и $\langle g, \hat{g}\rangle = \|g\|_*$. Выбирая $x = x_0 + \varepsilon\hat{g}$, получаем:

$$\varepsilon\|g\|_* \le f(x) - f(x_0) \le M\|x - x_0\| = M\varepsilon.$$

Таким образом, субдифференциал в точке ограничен локальной константой Липшица выпуклой функции. ■

Замечание C6.4. Отметим, что для следования не пустоты субдифференциала из выпуклости f существенно, чтобы x_0 принадлежало внутренности эффективного множества. Проиллюстрируем это примером.

Пример C6.5. Рассмотрим функцию $f : \mathbb{R} \to \mathbb{R}$:

$$f(x) = \begin{cases} -\sqrt{x}, & x \ge 0, \\ +\infty, & x < 0. \end{cases}$$

Это выпуклая функция, которая, тем не менее, не субдифференцируема в нуле. Предположим, что существует такой $g \in \mathbb{R}$, что:

$$-\sqrt{y} \ge gy, \ \forall y \ge 0.$$

Подставив $y = 1$, получим $g \le -1$. Подставив затем $y = \frac{1}{2g^2}$, получим новое условие на g (с учетом $g < -1$):

$$\frac{1}{2g^2} \le \frac{1}{4g^2}.$$

Это неравенство не может быть выполнено.

Доказанную теорему об ограниченности субдифференциала можно расширить, показав, что даже объединение субдифференциалов по всем точкам компактного подмножества внутренности $\operatorname{dom} f$ ограничено.

Теорема C6.2. Пусть в евклидовом пространстве V определена собственная выпуклая функция $f : V \to \overline{\mathbb{R}}$. Пусть $X \subseteq \operatorname{int}\operatorname{dom} f$ — непустой компакт. Тогда $Y = \cup_{x\in X}\partial f(x)$ — непустое ограниченное множество.

Доказательство. Не пустота Y очевидна, поскольку не пуст каждый субдифференциал, поэтому перейдём сразу к доказательству ограниченности. Предположим, что

существует последовательность $\{x_k\}_{k=1}^{\infty}$, такая, что $\|g_k\|_* \to \infty$, где $g_k \in \partial f(x_k)$. Для каждого g_k выберем вектор $\hat{g}_k$, такой что $\|\hat{g}_k\| = 1$ и $\langle g_k, \hat{g}_k \rangle = \|g_k\|_*$. Тогда выполнено

$$f(x_k + \varepsilon \hat{g}_k) - f(x_k) \geq \varepsilon \|g_k\|_*.$$

Тогда

$$f(x_k + \varepsilon \hat{g}_k) - f(x_k) \to \infty. \tag{C6.2}$$

Поскольку обе последовательности существуют на компакте в евклидовых пространствах, из них можно выделить сходящиеся подпоследовательности $\{x_m\}_{m=1}^{\infty}, \{g_m\}_{m=1}^{\infty}$:

$$f(x_m + \varepsilon \hat{g}_m) - f(x_m) \to f(x_{\lim} + \varepsilon \hat{g}_{\lim}) - f(x_{\lim}).$$

Здесь дополнительно воспользовались переходом к пределу под знаком непрерывной функции. Это находится в противоречии с (C6.2). Тогда произвольная $\{g_k\}_{k=1}^{\infty}$ ограничена. ■

Таким образом, во внутренности эффективного множества субдифференциал обладает рядом полезных свойств, позволяющих строить эффективные методы. Однако, есть и ряд патологических случаев. Например, если целевая функция задана на подмножестве, имеющем меньшую размерность, чем объемлющее пространство.

Теорема C6.3. Пусть в евклидовом пространстве V определена собственная функция $f : V \to \overline{\mathbb{R}}$ Пусть $\dim \operatorname{dom} f < \dim V$ и $\partial f(x_0) \neq \varnothing$ для некоторой $x_0 \in \operatorname{dom} f$. Тогда $\partial f(x_0)$ — неограниченное множество.

Доказательство. Зафиксируем произвольный $g \in \partial f(x_0)$. Напомним, что размерность множества мы определяли как размерность его аффинной оболочки. Тогда для $\widehat{V} = \operatorname{aff} \operatorname{dom} f - \{x_0\}$ имеем

$$\dim \widehat{V} < \dim V.$$

В частности, это значит, что существует ненулевой вектор $v \in V$, такой что

$$\langle v, w \rangle = 0, \ \forall w \in \widehat{V}.$$

Воспользуемся определением субградиента:

$$f(y) \geq f(x) + \langle g, y - x \rangle = f(x) + \langle g + \lambda v, y - x \rangle.$$

Это означает, что $g + \lambda v \in \partial f(x_0)$ для любых значений λ. Таким образом, субдифференциал — неограниченное множество. ■

Поговорим теперь про случай дифференцируемой функции. Интуитивно понятно, что субдифференциал либо пуст, либо является одноточечным множеством, состоящим из градиента. Формализуем это в виде утверждения.

Утверждение C6.1. Пусть в евклидовом пространстве V определена функция $f : V \to \overline{\mathbb{R}}$, $\operatorname{dom} f$ — выпуклое множество, и пусть задана точка $x_0 \in \operatorname{int} \operatorname{dom} f$, в которой функция f является дифференцируемой. Тогда $\partial f(x_0) = \varnothing$ или $\partial f(x_0) = \nabla f(x_0)$. Если функция f является выпуклой, тогда $\partial f(x_0) = \nabla f(x_0)$.

Доказательство. Если функция дифференцируема, то g определено единственным образом. В случае выпуклой функции, линейная аппроксимация с g подпирает график функции снизу — субдифференциал состоит из одного элемента. В противном случае, единственная касательная не подпирает график снизу, и субдифференциал оказывается пустым. ■

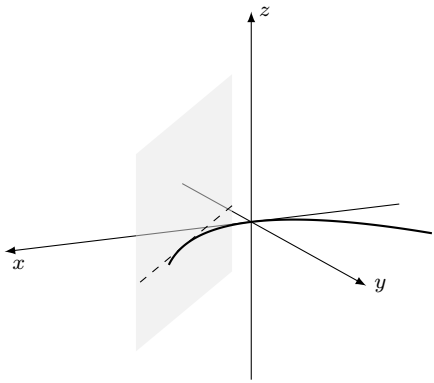


Рис. С6.1: Пример, когда субдифференциал — неограниченное множество.

Пример С6.6. Найдите субдифференциал функции $f : \mathbb{R} \to \mathbb{R}$: $f\big(x\big) = |x|$.

Решение. Заметим, что данная функция является дифференцируемой при $x \neq 0$ и выпуклой. Тогда применяя Теорему С6.1 мы получаем, что $\partial f\big(x\big) = f'(x) = \operatorname{sign}(x)$, при $x \neq 0$. Примере С6.1 мы получили, что $\partial f\big(0\big) = [-1, 1]$. Таким образом,

$$\partial |x| = \begin{cases} -1, & x < 0 \\ [-1, 1], & x = 0 \\ 1, & x > 0. \end{cases}$$

■

Пример С6.7. Найдите субдифференциал функции $f : \mathbb{R}^2 \to \mathbb{R}$:

$$f(x, y) = \cos(xy).$$

Решение. Заметим, что функция $f\big(x, y\big)$ дифференцируема на всём $\mathbb{R}^2$, а значит её субдифференциал есть пустое множество, либо состоит только из градиента, осталось понять, где градиент дает нижнюю оценку функции. Для этого найдем её минимум:

$$\min f(x, y) = -1, \text{ при } y = \frac{\pi(1 + 2k)}{x}, k \in \mathbb{Z}.$$

Так как субградиент — это касательная плоскость, подпирающая нашу функцию снизу, то он может существовать только в точках, где $\cos(xy) = -1$, остается только найти в них градиент:

$$\frac{\partial f}{\partial x} = -\sin(xy) \cdot y = -\sin(\pi(1 + 2k)) \cdot y = 0,$$
$$\frac{\partial f}{\partial y} = -\sin(xy) \cdot x = -\sin(\pi(1 + 2k)) \cdot x = 0.$$

В точках $\left(x, \frac{\pi(1+2k)}{x}\right), k \in \mathbb{Z}$ субградиент единственен и равен $(0, 0)$. В остальных точках субградиента не существует. ■

Заметим, что, хоть не пустота субдифференциала и не является необходимым условием выпуклости функции, её можно использовать как достаточное условие.

Утверждение С6.2. Пусть в евклидовом пространстве V определена собственная функция $f : V \to \overline{\mathbb{R}}$, $\operatorname{dom} f$ — выпуклое множество. Если для любого $x \in \operatorname{dom} f$ субдифференциал $\partial f(x)$ не пуст, то f — выпуклая функция.

Доказательство. Зафиксируем произвольные $x, y \in \operatorname{dom} f$ и определим $z = \lambda x + (1-\lambda)y$, где $\lambda \in [0,1]$. Поскольку $\operatorname{dom} f$ выпуклое, имеем $z \in \operatorname{dom} f$. Из не пустоты субдифференциала в любой точке эффективного множества имеем

$$f(y) \geq f(z) + \langle g, y - z\rangle = f(z) + \lambda\langle g, y - x\rangle,$$
$$f(x) \geq f(z) + \langle g, x - z\rangle = f(z) - (1-\lambda)\langle g, y - x\rangle.$$

Умножим первое неравенство на $1-\lambda$, второе — на λ. Сложив получившиеся неравенства, получим

$$f(z) \leq \lambda f(x) + (1-\lambda) f(y).$$

Тогда f выпукла по определению. ■

Утверждение С6.3. Пусть в евклидовом пространстве V определена собственная выпуклая функция $f : V \to \overline{\mathbb{R}}$. Точка $x_0 \in \operatorname{dom} f$ является минимумом функции $f(x)$ тогда и только тогда, когда

$$\exists g \in \partial f(x_0) : \ \langle g, x - x_0\rangle \geq 0 \ \forall x \in V.$$

В случае $\operatorname{dom} f = V$ имеем в качестве критерия $0 \in \partial f(x_0)$.

Пример С6.8. Пусть функция $f : \mathbb{R} \to \mathbb{R}$ задана следующим образом $f(x) = |x|$. Покажите, что у данной функции минимум достигается в точке 0.

Решение. Воспользуемся Теоремой С6.3 и получим, что минимум достигается в точке $x = 0$ и только в ней, так как $0 \in \partial f(0) = [-1, 1]$ и при этом $0 \notin \partial f(x) = \operatorname{sign} x$ при $x \neq 0$. ■

Теорема С6.4. Пусть в евклидовом пространстве V определена собственная выпуклая функция $f : V \to \overline{\mathbb{R}}$, $\operatorname{dom} f$ — выпуклое множество. Если $x_0 \in \operatorname{int} \operatorname{dom} f$, то $\partial f(x_0)$ — выпуклый компакт.

Доказательство. В евклидовом случае достаточно показать, что $\partial f(x_0)$ — ограниченно и замкнуто. Ограниченность мы уже показали ранее. Заметим, что субдифференциал можно представить как

$$\partial f(x_0) = \bigcap_{y \in V} \left\{ g \in V \mid f(y) \geq f(x) + \langle g, y - x\rangle \right\}.$$

Таким образом, субдифференциал — есть пересечение полуплоскостей. Как следствие, он выпуклый и замкнутый. В совокупности с ограниченностью получаем компактность. ■

С6.2 Субдифференциальное исчисление

Для начала, сформулируем определение относительной окрестности.

Определение С6.3. Рассмотрим подмножество S некоторого евклидова пространства V. Множество $\mathrm{ri}(x_0, \varepsilon) = \mathcal{B}_\varepsilon(x_0) \cap \operatorname{aff} S$ называется *относительной окрестностью точки* $x_0 \in S$.

Неформально, относительная окрестность точки — есть обыкновенная окрестность в объемлющем пространстве, суженная на аффинное продолжение интересующего нас множества. Этот объект позволяет ввести, например, внутренние точки для отрезка на плоскости.

Определение С6.4. Рассмотрим подмножество S некоторого евклидова пространства V. Точка x_0 называется *относительно внутренней точкой множества* S, если существует $\varepsilon > 0$, такое что $\mathrm{ri}(x_0, \varepsilon) \subset S$.

Определение С6.5. Рассмотрим подмножество S некоторого евклидова пространства V. *Относительной внутренностью* $\operatorname{relint}(S)$ *множества* S называется множество всех его относительно внутренних точек.

Отметим, что субдифференциал линеен относительно умножения на неотрицательную константу.

Утверждение С6.4. Пусть в евклидовом пространстве V определена функция $f : V \to \overline{\mathbb{R}}$, $\alpha > 0$ и пусть $x_0 \in \operatorname{dom} f$. Тогда

$$\partial(\alpha f)(x_0) = \alpha \partial f\big(x_0\big).$$

Очевидно, что для отрицательной константы это правило не работает, так как меняется знак неравенства и функция перестает быть субдифференцируемой.
В случае привычного нам дифференцирования производная была линейным оператором. Аналог этого свойства для субдифференциалов дает теорема Моро-Рокафеллара.

Теорема С6.5 (**Моро-Рокафеллара**)**.** Пусть в евклидовом пространстве V определены n функций $f_i : V \to \overline{\mathbb{R}}$. Определим их взвешенную сумму

$$f\big(x\big) = \sum_{i=1}^{n} \alpha_i f_i\big(x\big),\ \alpha_i > 0.$$

Тогда в точке $x_0 \in \cap_{i=1}^{n} \operatorname{dom} f_i$:

$$\partial f\big(x_0\big) \supseteq \sum_{i=1}^{n} \alpha_i \partial f_i\big(x_0\big),$$

где под суммой субдифференциалов подразумевается сумма Минковского. Если дополнительно известно, что $\cap_{i=1}^{n} \operatorname{relint}(\operatorname{dom} f_i) \neq \varnothing$ и все функции выпуклые, то

$$\partial f\big(x\big) = \sum_{i=1}^{n} \alpha_i \partial f_i\big(x\big).$$

Таким образом, линейность субдифференциалов относительно сложения требует наложения дополнительных условий на эффективные множества.

Пример С6.9. Рассмотрим функцию $f : [0, +\infty) \to \mathbb{R}$: $f(x) = -\sqrt{x}$ и функцию $h : (-\infty, 0] \to \infty$: $h(x) = -\sqrt{-x}$. Тогда $\partial(f + h)(0) = \mathbb{R}$, так как определена только в точке 0 и $(f + h)(0) = 0$. Но при этом $\partial f(0) = \partial h(0) = \varnothing$. Это было разобрано в Примере С6.5.

Часто встречаемый в оптимизации пример — композиция какой-то функции с линейным отображением.

Теорема С6.6. Пусть V и W — евклидовы пространства. Определим функцию $T : V \to W$: $T(x) = Ax + b$ и функцию $f : W \to \overline{\mathbb{R}}$. Тогда для $x_0 \in T^{-1}(\operatorname{dom} f)$ выполнено

$$\partial(fT)(x_0) \supseteq A^\top \partial f(z)\Big|_{z=T(x_0)}.$$

Если дополнительно известно, что функция f выпуклая и при этом $T(V) \cap \operatorname{int} \operatorname{dom} f \neq \varnothing$, то

$$\partial(fT)(x_0) = A^\top \partial f(z)\Big|_{z=T(x_0)}.$$

Пример С6.10. Найдите субдифференциал функции $f : \mathbb{R}^d \to \mathbb{R}$:

$$f(x) = \|Ax + b\|,\ A \in \mathbb{R}^{n\times d},\ b \in \mathbb{R}.$$

Решение. Норма — выпуклая функция, заданная на всем пространстве. Таким образом, можем воспользоваться вторым случаем Теоремы С6.6 и записать

$$\partial f(x) = A^\top \partial(\|z\|)\Big|_{z=Ax+b}.$$

Субдифференциал нормы был посчитан ранее:

$$\partial(\|z\|) = \begin{cases} \mathcal{B}_{\|\cdot\|_*}(0,1), & z = 0, \\ \frac{z}{\|z\|}, & z \neq 0. \end{cases}$$

Таким образом, имеем

$$\partial f(x) = \begin{cases} \{\, A^\top(Ax+b) \mid \|Ax+b\|_* \leq 1 \,\}, & Ax + b = 0 \\ \frac{A^\top(Ax+b)}{\|Ax+b\|}, & Ax + b \neq 0. \end{cases}$$

■

В матрично-векторном дифференцировании ключевую роль играло правило дифференцирования сложной функции. Его аналог для субдифференциалов даёт следующая теорема.

Теорема С6.7. Пусть в евклидовом пространстве V определены n выпуклых функций $g_i : V \to \overline{\mathbb{R}}$ и функция $\varphi : \mathbb{R}^n \to \mathbb{R}$ — монотонная неубывающая функция, $\operatorname{dom} \phi = \cap_{i=1}^n \operatorname{dom} f_i$. Рассмотрим $f(x) = \varphi(g(x))$, где $g(x) = (g_1(x), \ldots, g_m(x))$.

Тогда

$$\partial f(x) = \cup_{p \in \partial\varphi(u)} \left(\sum_{i=1}^{n} p_i \partial g_i(x) \right) \Bigg|_{u=g(x)}.$$

Пример С6.11. Найдите субдифференциал функции $f : \mathbb{R}^d \to \mathbb{R}$: $f(x) = \|x\|_2^2$.

Решение. Норма — выпуклая функция, $\varphi(u) = u^2$ — монотонная неубывающая на области значений нормы. Таким образом, имеем

$$\partial f(x) = 2\|x\|_2 \partial(\|x\|_2).$$

Выражение для субдифференциала нормы было посчитано ранее. ■

Пример С6.12. Найдите субдифференциал функции $f : \mathbb{R}^d \to \mathbb{R}$:

$$f(x) = \exp(\|Ax + b\|), \ A \in \mathbb{R}^{n \times d}, \ b \in \mathbb{R}.$$

Решение. $\|Ax + b\|$ — выпуклая функция как композиция выпуклой и аффинной. Экспонента — монотонная возрастающая функция. Запишем

$$\partial f(x) = \exp(\|Ax + b\|) \partial(\|Ax + b\|).$$

Воспользуемся правилом субдифференцирования композиции выпуклой и аффинной функции

$$\partial f(x) = \exp(\|Ax + b\|) A^\top \partial(\|z\|) \Big|_{z=Ax+b}.$$

Как результат, получим

$$\partial f(x) = \begin{cases} \left\{ \exp(\|Ax + b\|) A^\top (Ax + b) \mid \|Ax + b\|_* \leq 1 \right\}, & z = 0 \\ \frac{\exp(\|Ax+b\|) A^\top (Ax+b)}{\|Ax+b\|}, & z \neq 0. \end{cases}$$

■

Последний теоретический результат, который мы рассмотрим в рамках главы — теорема Дубовицкого-Милютина.

Теорема С6.8 (Дубовицкого-Милютина). Пусть в евклидовом пространстве V определены n выпуклых функций $f_i : V \to \overline{\mathbb{R}}$. Определим

$$f(x) = \max_i f_i(x).$$

Тогда для $x_0 \in \cap_{i=1}^n \operatorname{dom} f_i$ выполнено

$$\partial f(x_0) \supseteq \operatorname{cl}\left(\operatorname{conv}\left\{\cup_{i \in I(x_0)} \partial f_i(x_0)\right\}\right),$$

где $I(x_0) = \left\{ i \in \overline{1,n} \mid f(x_0) = f_i(x_0) \right\}$ — активное множество в точке x_0.
Если дополнительно известно, что $x_0 \in \cap_{i=1}^n \operatorname{int} \operatorname{dom} f_i$ и все функции под максимумом выпуклые, то

$$\partial f(x_0) = \operatorname{conv}\left\{\cup_{i \in I(x_0)} \partial f_i(x_0)\right\}.$$

Эта теорема оказывается крайне полезной для получения более сильных выражений. Для начала вернемся к простому примеру из начала главы.

Пример C6.13. Рассмотрим функцию $f : \mathbb{R}^d \to \mathbb{R}$: $f(x) = |x|$. Вычислим $\partial f(x)$. Заметим, что $f(x) = |x| = \max\{x, -x\}$. Для $x < 0$ и $x > 0$ активное множество состоит из одного индекса. Учитывая, что модуль — выпуклая функция, определённая на всем пространстве, получаем

$$\partial f(x) = \begin{cases} 1, & x > 0 \\ -1, & x < 0. \end{cases}$$

Если $x = 0$, то активное множество включает в себя оба индекса. Это означает

$$\partial f(0) = [-1, 1].$$

Перейдём к более сложному примеру.

Пример C6.14. Найдите субдифференциал функции $f : \mathbb{R}^d \to \mathbb{R}$: $f(x) = \|x\|_1$.

Решение. Заметим, что $f(x) = \sum\limits_{i=1}^{d} |x_i|$. Это выпуклая функция, определённая на всём пространстве. Воспользуемся теоремой Моро-Рокафеллара:

$$\partial f(x) = \sum_{i=1}^{d} \partial(|x_i|).$$

Далее покомпонентно воспользуемся утверждением прошлого примера и запишем ответ в форме суммы Минковского:

$$\partial f(x) = \sum_{i \in I_{\neq}(x)} e_i \operatorname{sign} x_i + \sum_{i \in I_{=}(x)} [-e_i, e_i],$$

где $I_{\neq}(x) = \{\, i \mid x_i \neq 0 \,\}$, $I_{=}(x) = \{\, i \mid x_i = 0 \,\}$. ■

Замечание C6.5. На практике, если необходим лишь какой-то элемент субдифференциала, можно брать $g = \operatorname{sign} x$ — покомпонентное взятие знака.

Пример C6.15. Найдите субдифференциал функции $f : \mathbb{R}^d \to \mathbb{R}$:

$$f(x) = \|Ax + b\|_1, \; A \in \mathbb{R}^{n \times d}, \; b \in \mathbb{R}.$$

Решение. По правилу субдифференцирования аффинной композиции имеем

$$\partial f(x) = A^\top \partial(\|z\|_1)\Big|_{z=Ax+b}.$$

Введём активные множества как

$$I_{\neq}(x) = \left\{ i \;\middle|\; a_i^\top x_i + b_i \neq 0 \right\},$$
$$I_{=}(x) = \left\{ i \;\middle|\; a_i^\top x_i + b_i = 0 \right\},$$

где a_i — соответствующий столбец матрицы A. Таким образом,

$$\partial\big(\|z\|_1\big)\Big|_{z=Ax+b} = \sum_{i\in I_{\neq}(x)} e_i \operatorname{sign}(a_i^\top x_i + b_i) + \sum_{i\in I_{=}(x)} [-e_i, e_i].$$

Объединяя полученные результаты, запишем ответ:

$$\begin{aligned}\partial f(x) &= \sum_{i\in I_{\neq}(x)} A^\top e_i \operatorname{sign}(a_i^\top x_i + b_i) + \sum_{i\in I_{=}(x)} [-A^\top e_i, A^\top e_i] \\ &= \sum_{i\in I_{\neq}(x)} a_i \operatorname{sign}(a_i^\top x_i + b_i) + \sum_{i\in I_{=}(x)} [-a_i, a_i].\end{aligned}$$

■

C7 Сопряжённые функции Фенхеля

Сопряжённые функции Фенхеля — важный инструмент, лежащий в основе теории двойственности. В нашем курсе мы познакомимся с определением сопряжения по Фенхелю, его интуитивной интерпретацией и основными свойствами. Выясним, как оно помогает имплементировать сопряжённые методы, получать нижнюю оценку на значение целевой функции в оптимуме или даже сводить оптимизационную задачу к более простой, имеющей очевидное аналитическое решение.

C7.1 Определение сопряжённой функции Фенхеля

Определение C7.1. Пусть $f : \mathbb{R}^d \to \overline{\mathbb{R}}$. Функция $f^* : \mathbb{R}^d \to \overline{\mathbb{R}}$, определённая как

$$f^*(y) = \sup_{x \in \mathbb{R}^d} \{\langle x, y\rangle - f(x)\}$$

называется *сопряжённой функцией Фенхеля* к f.

Далее в рамках этой главы сопряжённые функции Фенхеля будем называть просто сопряжёнными функциями. Сопряжённая функция в точке $y \in \mathbb{R}^d$ показывает, насколько нужно сдвинуть график функции f, чтобы линейная функция $g(x) = \langle x, y\rangle$ подпирала его снизу. Иными словами, Определение C7.1 показывает, насколько та или иная функция отличается от линейной.

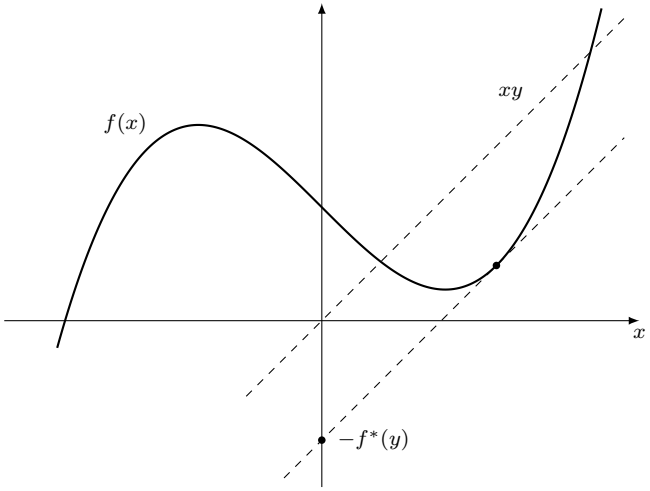


Рис. C7.1: Зазор показан пунктирными линиями. Если $f(x)$ дифференцируема, то пунктирная линия касается f в точке x, где $f'(x) = y$.

Из интуитивного объяснения определения должно быть ясно, что сопряжённая функция к линейной функции с константным сдвигом — есть функция, тождественно равная сдвигу с обратным знаком.

Пример C7.1. Найдите сопряжённую функцию к линейной функции $f : \mathbb{R}^d \to \mathbb{R}$:

$$f(x) = \langle a, x \rangle + b,\ a \in \mathbb{R}^d,\ b \in \mathbb{R}.$$

Решение. Воспользуемся определением сопряжённой функции:

$$f^*(y) = \sup_{x \in \mathbb{R}^d} \{\langle y, x \rangle - \langle a, x \rangle - b\} = \sup_{x \in \mathbb{R}^d} \{\langle y - a, x \rangle - b\}.$$

Понятно, что в случае $y \neq a$ имеем две гиперплоскости с разными нормалями. Как следствие, ни одна из них не может подпирать другую и $f^*(y) = +\infty$. Если же $y = a$, то $f^*(y) = -b$. Таким образом, имеем

$$f^*(y) = \begin{cases} -b, & y = a, \\ +\infty, & y \neq a. \end{cases}$$

■

Обратим внимание, что сопряжённая функция принимает значения в расширенной числовой прямой. Понятно, что с точки зрения практики интересен случай, когда она гарантированно принимает конечные значения на каком-то множестве переменных. Это можно гарантировать, если наложить дополнительные условия на функцию f.

Утверждение C7.1. Пусть $f : \mathbb{R}^d \to \overline{\mathbb{R}}$ — выпуклая собственная функция. Тогда сопряжённая функция $f^* : \mathbb{R}^d \to \overline{\mathbb{R}}$ — тоже собственная.

Доказательство. Поскольку f — собственная функция, то существует такой $\hat{x} \in \mathbb{R}^d$, что

$$f(\hat{x}) < +\infty.$$

Тогда по определению сопряжённой функции имеем

$$f^*(y) \geq \langle y, \hat{x} \rangle - f(\hat{x}) > -\infty.$$

Осталось показать, что сопряжённая функция не может быть тождественно равна $+\infty$. Отметим, что из выпуклости f следует существование $\tilde{x} \in \operatorname{dom} f$, такого что $\partial f(\tilde{x}) \neq \emptyset$. Выбрав произвольный $g \in \partial f(\tilde{x})$, запишем для произвольного $y \in \mathbb{R}^d$:

$$f(y) \geq f(\tilde{x}) + \langle g, y - \tilde{x} \rangle.$$

Следовательно, имеем

$$f^*(g) = \sup_{y \in \mathbb{R}^d} \{\langle g, y \rangle - f(y)\} \leq \sup_{y \in \mathbb{R}^d} \{\langle g, \tilde{x} \rangle - f(\tilde{x})\} = \langle g, \tilde{x} \rangle - f(\tilde{x}) < +\infty.$$

Таким образом, f^* — собственная функция. ■

Замечание C7.1. В продолжение разговора о свойствах сопряжённой функции, нетрудно убедиться, что она выпуклая и замкнутая независимо от функции f. Действительно, согласно Определению C7.1, f^* — есть поточечный максимум аффинных функций, то есть выпуклая и замкнутая функция.

C7.2 Вычисление сопряжённой функции в одномерном случае

В этой секции рассмотрим ряд одномерных примеров, чтобы уловить интуицию и отработать технику. Примеры даны по возрастанию сложности.

Пример C7.2. Найдите сопряжённую функцию к функции $f : \mathbb{R} \to \mathbb{R}$:

$$f(x) = e^x.$$

Решение. Воспользуемся определением сопряжённой функции:

$$f^*(y) = \sup_{x \in \mathbb{R}} \{xy - e^x\}.$$

Попытаемся вычислить максимум вогнутой функции $g(x) = xy - e^x$. Запишем производную:

$$g'(x) = y - e^x = 0.$$

Рассмотрим случаи:

- $y > 0$. Уравнение разрешимо и $x = \ln y$. Подставляя в определение, получаем:

$$f^*(y) = y \ln y - y.$$

- $y = 0$. Поскольку $-f(x) \to 0$ при $x \to -\infty$, то:

$$f^*(y) = 0.$$

- $y < 0$. Поскольку $g(x) \to +\infty$ при $x \to -\infty$, то:

$$f^*(y) = +\infty.$$

Таким образом, имеем

$$f^*(y) = \begin{cases} y \ln y - y, & y \geq 0, \\ +\infty, & y < 0. \end{cases}$$

Здесь мы доопределяем $0 \ln 0 = 0$ пределом по непрерывности. ■

Замечание C7.2. Отметим, что далее в курсе мы будем регулярно использовать факт $0 \ln 0 = 0$.

Пример C7.3. Найдите сопряжённую функцию к функции $f : \mathbb{R} \to \mathbb{R}$:

$$f(x) = \begin{cases} -\ln x, & x > 0, \\ +\infty, & x \leq 0. \end{cases}$$

Решение. Воспользуемся определением сопряжённой функции:

$$f^*(y) = \sup_{x > 0} \{xy + \ln x\}.$$

Попытаемся вычислить максимум вогнутой функции $g(x) = xy + \ln x$. Запишем производную:

$$g'(x) = y + \frac{1}{x} = 0.$$

Рассмотрим случаи:

- $y < 0$. Уравнение разрешимо и $x = -\frac{1}{y}$. Подставляя в определение, получаем:

$$f^*(y) = -1 - \ln(-y).$$

- $y \geq 0$. Поскольку $g(x) \to +\infty$ при $x \to -\infty$, то:

$$f^*(y) = +\infty.$$

Таким образом:

$$f^*(y) = \begin{cases} -1 - \ln(-y), & y < 0, \\ +\infty, & y \geq 0. \end{cases}$$

■

Пример С7.4. Найдите сопряжённую функцию к функции hinge loss $f : \mathbb{R} \to \mathbb{R}$:

$$f(x) = \max\{1 - x, 0\}.$$

Решение. Воспользуемся определением сопряжённой функции:

$$f^*(y) = \sup_{x \in \mathbb{R}} \{yx - \max\{1 - x, 0\}\}.$$

Вспомним полезное свойство

$$-\max\{a, b\} = \min\{-a, -b\}.$$

Занося yx под минимум, получим

$$f^*(y) = \sup_{x \in \mathbb{R}} \{\min\{(1 + y)x - 1, yx\}\}.$$

Рассмотрим $g(x) = \min\{(1 + y)x - 1, yx\}$ более детально. Заметим, что

$$g(x) = \begin{cases} (1 + y)x - 1, & x \leq 1, \\ yx, & x > 1. \end{cases}$$

Таким образом, $g(x)$ — кусочно линейная функция. Нетрудно видеть, что она имеет конечный максимум тогда и только тогда, когда наклон левой ветви неотрицательный, а наклон правой — не положительный. Таким образом, сопряжённая функция конечна только для y, удовлетворяющих системе неравенств

$$\begin{cases} 1 + y \geq 0 \\ y \leq 0. \end{cases}$$

При этом ясно, что максимум в таком случае достигается в точке перелома $x = 1$ и равен y. Таким образом:

$$f^*(y) = \begin{cases} y, & y \in [-1, 0], \\ +\infty, & y \notin [-1, 0]. \end{cases}$$

■

Пример C7.5. Найдите сопряжённую функцию к функции $f : \mathbb{R} \to \mathbb{R}$:

$$f(x) = \frac{1}{p}|x|^p,\ p > 1.$$

Решение. Воспользуемся определением сопряжённой функции:

$$f^*(y) = \sup_{x \in \mathbb{R}} \left\{ xy - \frac{1}{p}|x|^p \right\}.$$

Попытаемся вычислить максимум вогнутой функции $g(x) = xy - \frac{1}{p}|x|^p$. Запишем производную:

$$g'(x) = y - \operatorname{sign} x |x|^{p-1} = 0.$$

Попытаемся решить уравнение. Записав $y = \operatorname{sign} y |y|$, поймем, что $\operatorname{sign} x = \operatorname{sign} y$ и $|x|^{p-1} = |y|$. Таким образом:

$$|x| = \operatorname{sign}(y) |y|^{\frac{1}{p-1}}.$$

Тогда сопряжённая функция к f:

$$f^*(y) = \left(1 - \frac{1}{p}\right) \operatorname{sign}(y) y^{\frac{p}{p-1}} = \frac{1}{q} \operatorname{sign}(y) y^q,$$

где q удовлетворяет уравнению

$$\frac{1}{p} + \frac{1}{q} = 1.$$

■

Замечание C7.3. Пример выше намекает на связь между сопряжёнными нормами и сопряжением по Фенхелю.

Пример C7.6. Найдите сопряжённую функцию к функции $f : \mathbb{R} \to \mathbb{R}$:

$$f(x) = \ln\left(1 + e^x\right).$$

Доказательство. Воспользуемся определением сопряжённой функции:

$$f^*(y) = \sup_{x \in \mathbb{R}} \left\{ xy - \ln\left(1 + e^x\right) \right\}.$$

Попытаемся вычислить максимум вогнутой функции $g(x) = xy - \ln\left(1 + e^x\right)$. Запишем производную:

$$g'(x) = y - \frac{e^x}{1 + e^x} = 0.$$

Рассмотрим случаи:

- $y \in (0, 1)$. Уравнение разрешимо и $x = \ln y - \ln(1 - y)$. Подставляя в определение, получаем:

  $$f^*(y) = y \ln y + (1 - y) \ln(1 - y).$$

  Получили бинарную кросс-энтропию.

- $y = 0$. Поскольку $-f(x) \to 0$ при $x \to -\infty$, то:

$$f^*(y) = 0.$$

- $y < 0$. Из монотонности логарифма и того, что $e^x < 1$ при $x < 0$, следует

$$\ln\left(1 + e^x\right) < \ln 2,\ \forall x < 0.$$

Это означает, что

$$xy - \ln\left(1 + e^x\right) > xy - \ln 2.$$

Поскольку $yx \to +\infty$ при $x \to -\infty$, то $xy - \ln\left(1 + e^x\right) \to +\infty$ при $x \to -\infty$. Таким образом

$$f^*(y) = +\infty.$$

- $y = 1$. Поскольку

$$\ln(1 + e^x) \geq x,\ \forall x \in \mathbb{R},$$

супремум не может быть больше нуля, а он равен нулю, так как

$$\ln\left(1 + e^x\right) = x + \ln\left(1 + e^{-x}\right),\ \forall x \in \mathbb{R}$$

и $\ln\left(1 + e^{-x}\right) \to 0$ при $x \to +\infty$. Поэтому:

$$f^*(y) = 0.$$

- $y > 1$. Поскольку:

$$\ln\left(1 + e^x\right) < \ln\left(e^x + e^x\right) = \ln 2 + x,\ \forall x > 0.$$

Отсюда следует

$$xy - \ln\left(1 + e^x\right) > (y - 1)x - \ln 2,\ \forall x > 0.$$

Устремляя $x \to +\infty$, получаем, что

$$f^*(y) = +\infty.$$

Таким образом:

$$f^*(y) = \begin{cases} y \ln y + (1 - y)\ln(1 - y), & y \in [0, 1] \\ +\infty, & y \notin [0, 1]. \end{cases}$$

■

C7.3 Вычисление сопряжённой функции в многомерном случае

Теперь, когда мы рассмотрели достаточно много одномерных примеров, перейдем к многомерному случаю. Как и в предыдущем параграфе, начнем с более простых примеров и двинемся к более сложным.

Пример C7.7. Найдите сопряжённую функцию к функции $f : \mathbb{R}^d \to \mathbb{R}$:

$$f(x) = \frac{1}{2}x^\top Ax + b^\top x,\ A \in \mathbb{S}^d_{++},\ b \in \mathbb{R}^d.$$

Решение. Воспользуемся определением сопряжённой функции:

$$f^*(y) = \sup_{x \in \mathbb{R}^d} \left\{ y^\top x - \frac{1}{2}x^\top Ax - b^\top x \right\} = \sup_{x \in \mathbb{R}^d} \left\{ -\frac{1}{2}x^\top Ax - (b-y)^\top x \right\}.$$

Вычислим максимум сильно вогнутой функции $g(x) = -\frac{1}{2}x^\top Ax - (b-y)^\top x$. Запишем градиент:

$$\nabla g(x) = -Ax - (b-y) = 0 \implies x = A^{-1}(y-b).$$

Тогда сопряжённая функция к f:

$$f^*(y) = \frac{1}{2}(y-b)^\top A^{-1}(y-b).$$

■

Пример C7.8. Найдите сопряжённую функцию к функции $f : \mathbb{R}^d \to \mathbb{R}$:

$$f(x) = \ln\left(\sum_{i=1}^d e^{x_i}\right).$$

Решение. Воспользуемся определением сопряжённой функции:

$$f^*(y) = \sup_{x \in \mathbb{R}} \left\{ y^\top x - \ln\left(\sum_{i=1}^d e^{x_i}\right) \right\}.$$

Попытаемся вычислить максимум вогнутой функции $g(x) = y^\top x - \ln\left(\sum_{i=1}^d e^{x_i}\right)$. Запишем градиент:

$$\nabla g_i(x) = y_i - \frac{e^{x_i}}{\sum\limits_{j=1}^d e^{x_j}} = 0,\ i = \overline{1,d}.$$

Рассмотрим случаи:

- $y \succ 0$ и $\mathbf{1}^\top y = 1$. Уравнение разрешимо и

  $$\ln y_i = x_i - \ln\left(\sum_{j=1}^d e^{x_j}\right) = x_i - f(x).$$

  Подставляя в определение, получаем:

  $$f^*(y) = \sum_{i=1}^d y_i(\ln y_i + f(x)) - f(x) = \sum_{i=1}^d y_i \ln y_i,$$

  так как $\mathbf{1}^\top y = 1$.

- $y \succeq 0$ и $\mathbf{1}^\top y = 1$. Пусть существует $y_k = 0$, тогда $x_k \to -\infty$ и ответ не поменяется:

$$f^*\big(y\big) = \sum_{i=1}^{d} y_i\big(\ln y_i + f\big(x\big)\big) - f\big(x\big) = \sum_{i=1}^{d} y_i \ln y_i,$$

- $y \not\succeq 0$ и $\mathbf{1}^\top y = 1$. Пусть существует $y_k < 0$, тогда выберем $x_k = -t$ и $x_i = 0$ $(i \neq k)$, а затем устремим $t \to +\infty$. Таким образом

$$f^*\big(y\big) = +\infty.$$

- $y \succeq 0$ и $\mathbf{1}^\top y \neq 1$. Возьмём $x = t\mathbf{1}$:

$$y^\top x - f\big(x\big) = t y^\top \mathbf{1} - t - \ln(d).$$

  В случае, если $\mathbf{1}^\top y > 1$ выражение не ограничено при $t \to +\infty$. В случае, если $\mathbf{1}^\top y < 1$ выражение не ограничено при $t \to -\infty$. Таким образом

$$f^*\big(y\big) = +\infty.$$

Таким образом:

$$f^*\big(y\big) = \begin{cases} \sum\limits_{i=1}^{d} y_i \ln y_i, & y \in \triangle_{d-1} \\ +\infty, & y \notin \triangle_{d-1}. \end{cases}$$

■

Пример C7.9. Найдите сопряжённую функцию к функции $f : \mathbb{S}^d_{++} \to \mathbb{R}$:

$$f\big(X\big) = \ln\det X^{-1}.$$

Решение. Воспользуемся определением сопряжённой функции:

$$f^*\big(Y\big) = \sup_{X \succ 0} \big\{\operatorname{Tr}(XY) + \ln\det X\big\},$$

где мы воспользовались, что $\det X^{-1} = \big(\det X\big)^{-1}$.
Попытаемся вычислить максимум вогнутой функции $g(X) = \operatorname{Tr}(XY) + \ln\det X$. Запишем градиент:

$$\nabla g(X) = Y + X^{-1} = 0,$$

Рассмотрим случаи:

- $Y \prec 0$. Уравнение разрешимо и $X = -Y^{-1}$. Подставляя в определение, получаем:

$$f^*\big(Y\big) = \operatorname{Tr}(-I_d) + \ln\det(-Y)^{-1} = \ln\det(-Y)^{-1} - d.$$

- $Y \not\prec 0$. Y имеет нормированный собственный вектор v и собственное значение $\lambda \geq 0$. Возьмем $X = I + tvv^\top$, тогда

$$\operatorname{Tr}(XY) + \ln\det X = \operatorname{Tr}(Y) + t\lambda + \ln\det\Big(I + tvv^\top\Big) = \operatorname{Tr}(Y) + t\lambda + \ln(1+t).$$

Таким образом:

$$f^*(Y) = +\infty.$$

Таким образом:

$$f^*(Y) = \begin{cases} \ln\det(-Y)^{-1} - d, & Y \prec 0, \\ +\infty, & Y \not\prec 0. \end{cases}$$

■

Пример С7.10. Найдите сопряжённую функцию к функции $f : \mathbb{R}^d \to \mathbb{R}$:

$$f(x) = \begin{cases} -\sqrt{1 - \|x\|^2}, & \|x\| \leq 1, \\ +\infty, & \|x\| < 1. \end{cases}.$$

Решение. Поскольку мы мало что знаем о сопряжении в случае сложных функций, решить задачу «в лоб» будет крайне сложно. Вместо этого, мы сведем задачу к одномерному виду:

$$f^*(y) = \sup_{\|x\| \leq 1} \left\{ \langle y, x \rangle + \sqrt{1 - \|x\|^2} \right\} = \sup_{\alpha \in [0,1]} \sup_{\|x\| = \alpha} \left\{ \langle y, x \rangle + \sqrt{1 - \alpha^2} \right\}.$$

Пользуясь определением сопряжённой нормы, раскроем один из супремумов:

$$f^*(y) = \sup_{\alpha \in [0,1]} \left\{ \alpha \|y\|_* + \sqrt{1 - \alpha^2} \right\},$$

откуда получим

$$\alpha = \frac{\|y\|_*}{\sqrt{\|y\|_*^2 + 1}}.$$

Таким образом:

$$f^*(y) = \sqrt{\|y\|_*^2 + 1}.$$

■

Пример С7.11. Найдите сопряжённую функцию к функции $f : \mathbb{R}^d \to \mathbb{R}$:

$$f(x) = \frac{1}{2}\|x\|^2.$$

Решение. Поступим аналогично предыдущему примеру:

$$f^*(y) = \sup_{\alpha \geq 0} \sup_{\|x\| = \alpha} \left\{ \langle y, x \rangle - \frac{\alpha^2}{2} \right\} = \sup_{\alpha \geq 0} \left\{ \alpha \|y\|_* - \frac{\alpha^2}{2} \right\} = \frac{1}{2}\|y\|_*^2.$$

■

Пример C7.12. Найдите сопряжённую функцию к функции $f : \mathbb{R}^d \to \mathbb{R}$:

$$f(x) = \|x\|.$$

Доказательство. Рассмотрим случаи:

- $\|y\|_* > 1$. По определению двойственной нормы существует $z \in \mathbb{R}^d$ с $\|z\| \leq 1$ и $y^\top z > 1$. Беря $x = tz$ и устремляя $t \to +\infty$, получаем:

  $$y^\top x - \|x\| = t(y^\top z - \|z\|) \to +\infty.$$

  Таким образом:

  $$f^*(y) = +\infty.$$

- $\|y\|_* \leq 1$. Воспользуемся неравенством Гёльдера: $y^\top x \leq \|x\|\|y\|_*$. Тогда

  $$y^\top x - \|x\| \leq 0.$$

  При $x = 0$, выражение $y^\top x - \|x\| = 0$, то есть

  $$f^*(y) = 0.$$

Таким образом:

$$f^*(y) = \begin{cases} 0, & \|y\|_* \leq 1, \\ +\infty, & \|y\|_* > 1. \end{cases}$$

■

C7.4 Анализ на сопряжённых функциях

До этой секции мы строили сопряжённые функции, пользуясь только определением сопряжения по Фенхелю. Однако, встречая некоторые более сложные примеры, логично задаться вопросом: существуют ли правила, которые помогут облегчить работу с сопряжёнными функциями? К сожалению, взятие сопряжения устроено не так хорошо, как дифференцирование или даже субдифференцирование. Однако, мы все же можем получить некоторые утверждения, которые окажутся полезными.

Теорема C7.1. Пусть $f : \mathbb{R}^d \to \overline{\mathbb{R}}$ и $\alpha \in \mathbb{R}_{++}$. Тогда для $g(x) = \alpha f\big(x\big)$ и $h(x) = \alpha f\big(\frac{x}{\alpha}\big)$ выполнено:

$$g^*(y) = \alpha f^*\Big(\frac{y}{\alpha}\Big),\ h^*(y) = \alpha f^*\big(y\big).$$

Доказательство. Докажем по определению:

$$g^*\big(y\big) = \sup_{x \in \mathbb{R}^d} \big\{\langle y, x\rangle - \alpha f\big(x\big)\big\} = \alpha \sup_{x \in \mathbb{R}^d} \Big\{\Big\langle \frac{y}{\alpha}, x\Big\rangle - f\big(x\big)\Big\} = \alpha f^*\Big(\frac{y}{\alpha}\Big).$$

Аналогично поступим с $h(x)$:

$$h^*(x) = \sup_{x\in\mathbb{R}^d}\left\{\langle y, x\rangle - \alpha f\Big(\frac{x}{\alpha}\Big)\right\} = \alpha \sup_{x\in\mathbb{R}^d}\left\{\Big\langle y, \frac{x}{\alpha}\Big\rangle - f\Big(\frac{x}{\alpha}\Big)\right\}$$
$$= \alpha \sup_{z\in\mathbb{R}^d}\left\{\langle y, z\rangle - f(z)\right\} = \alpha f^*(y).$$

■

Таким образом, для операции взятия сопряжения по Фенхелю нет линейности относительно умножения на константу, но есть правила, которые облегчают жизнь при вычислении.

Пример С7.13. Найдите сопряжённую функцию к функции $f : \mathbb{R}^d \to \mathbb{R}$:

$$f(x) = \begin{cases} -\sqrt{\alpha^2 - \|x\|^2}, & \|x\| \le \alpha, \\ +\infty, & \|x\| > \alpha, \end{cases}$$

где $\alpha \in \mathbb{R}$.

Решение. Можно заметить, что мы уже решали очень похожую задачу в Примере С7.10. Воспользуемся Теоремой С7.1, тогда

$$f^*(y) = \alpha\sqrt{\|y\|_*^2 + 1}.$$

■

Теорема С7.2. Рассмотрим набор собственных функций $f_i : \mathbb{R}^d \to \overline{\mathbb{R}}$. Пусть функция $g : \mathbb{R}^d \times \ldots \times \mathbb{R}^d \to \mathbb{R}$ задана по правилу

$$g(x_1, \ldots, x_n) = \sum_{i=1}^n f_i(x_i).$$

Тогда сопряжённая функция к g имеет вид

$$g*(y_1, \ldots, y_n) = \sum_{i=1}^n f_i^*(y_i).$$

Доказательство. Распишем по определению:

$$g^*(y_1, \ldots, y_n) = \sup_{x_1,\ldots,x_n}\left\{\langle (y_1, \ldots, y_n)^\top, (x_1, \ldots, x_n)^\top\rangle - g(x_1, \ldots, x_n)\right\}$$
$$= \sup_{x_1,\ldots,x_n}\left\{\sum_{i=1}^n \left[\langle y_i, x_i\rangle - f_i(x_i)\right]\right\}.$$

Поскольку каждое слагаемое в сумме независимо, можем посчитать супремумы отдельно:

$$g^*(y_1, \ldots, y_n) = \sum_{i=1}^n \sup_{x_i\in\mathbb{R}^d}\left\{\langle y_i, x_i\rangle - f_i(x_i)\right\} = \sum_{i=1}^n f^*(y_i).$$

■

Таким образом, взятие сопряжения по Фенхелю линейно в случае сепарабельных функций. Разберем несколько примеров, в которых Теорема С7.2 оказывается полезной.

Пример С7.14. Найдите сопряжённую функцию к функции $f : \mathbb{R}^d \to \mathbb{R}$:

$$f(x) = \begin{cases} \sum\limits_{i=1}^{d} x_i \ln x_i, & x \succeq 0, \\ +\infty, & x \not\succeq 0. \end{cases}$$

Решение. Поскольку функция f сепарабельна, достаточно рассмотреть одномерный случай. Рассмотрим $g(t) = t \ln t,\ t \geq 0$.Воспользуемся определением сопряжённой функции:

$$g^*(s) = \sup_{t \in \mathbb{R}} \{ts - t \ln t\} \implies g^*(s) = e^{s-1}.$$

Пользуясь Теоремой С7.2, получим:

$$f^*(y) = \sum_{i=1}^{d} e^{y_i - 1}.$$

■

Пример С7.15. Найдите сопряжённую функцию к функции $f : \mathbb{R}^d \to \mathbb{R}$:

$$f(x) = \begin{cases} -\sum\limits_{i=1}^{d} \ln x_i, & x \succ 0, \\ +\infty, & x \not\succ 0. \end{cases}$$

Решение. Заметим, что опять имеем дело с сепарабельной функцией. Для одномерного случая мы уже вычислили сопряжение. Тогда по Теореме С7.2 получим

$$f^*(y) = \begin{cases} -\sum\limits_{i=1}^{d} \ln(-y_i) - d, & y \prec 0, \\ +\infty, & y \not\prec 0. \end{cases}$$

■

Замечание С7.4. К сожалению, в несепарабельном случае линейность не выполняется. Мы подробно обсудим это в следующей главе, когда разберем функцию Лагранжа и понятие свертки функций.

В случае дифференцирования и субдифференцирования мы умели удобно выражать результат применения операции к композиции функции с аффинным отображением. Аналогичная формула есть и для взятия сопряжения.

Теорема С7.3. Рассмотрим $f : \mathbb{R}^d \to \overline{\mathbb{R}}$. Пусть $A \in \mathbb{R}^{d \times d}$, $b, a \in \mathbb{R}^d$ и $\det A \neq 0$. Тогда для g, определенной по правилу

$$g(x) = f\big(A(x - a)\big) + \langle b, x \rangle$$

выполнено

$$g^*(y) = f^*\big(A^{-\top}(y - b)\big) + \langle a, y - b \rangle.$$

Доказательство. Запишем по определению:

$$g^*(y) = \sup_{x\in\mathbb{R}^d} \big\{\langle y, x\rangle - f\big(A(x-a)\big) - \langle b, x\rangle\big\}.$$

Сделаем замену переменных $z = A(x-a)$. Тогда

$$\begin{aligned} g^*(y) &= \sup_{z\in\mathbb{R}^d} \big\{\langle y, A^{-1}z + a\rangle - f\big(z\big) - \langle b, A^{-1}z + a\rangle\big\} \\ &= \sup_{z\in\mathbb{R}^d} \big\{\langle y - b, A^{-1}z\rangle - f\big(z\big)\big\} + \langle a, y - b\rangle \\ &= \sup_{z\in\mathbb{R}^d} \big\{\langle A^{-\top}(y-b), z\rangle - f\big(z\big)\big\} + \langle a, y - b\rangle \\ &= f^*\big(A^{-\top}(y-b)\big) + \langle a, y - b\rangle. \end{aligned}$$

■

C7.5 Связь с субдифференциалом

Теория, построенная в этом параграфе, может пригодиться для вычисления субдифференциалов. Перед тем, как выяснить связь между этими объектами, введем понятие второй сопряжённой функции Фенхеля.

Определение C7.2. Рассмотрим функцию $f : \mathbb{R}^d \to \overline{\mathbb{R}}$. Функция $f^{**} : \mathbb{R}^d \to \overline{\mathbb{R}}$, определённая как

$$f^{**}\big(x\big) = \sup_{y\in\mathbb{R}^d} \big\{\langle x, y\rangle - f^*\big(y\big)\big\}.$$

называется *второй сопряжённой функцией по Фенхелю* к f.

Замечание C7.5. Нетрудно заметить, что $f^{**}\big(x\big) \le f\big(x\big)$ для любого $x \in \mathbb{R}^d$. Действительно, по определению сопряжённой функции $f\big(x\big) \ge \langle x, y\rangle - f^*\big(y\big)$. Переходя к супремуму по y, получаем требуемое.

Чтобы выяснить связь сопряжения с субдифференциалом потребуется выделить класс функций, для которых $f^{**} = f$.

Теорема C7.4. Пусть $f : \mathbb{R}^d \to \overline{\mathbb{R}}$ — собственная замкнутая выпуклая функция. Тогда

$$f^{**}\big(x\big) = f\big(x\big),\ \forall x \in \mathbb{R}^d.$$

Доказательство. Учитывая Замечание C7.5, остается показать обратное неравенство. Пусть нашелся такой $x \in \mathbb{R}^d$, что $f^{**}\big(x\big) < f\big(x\big)$. Поскольку f собственная замкнутая выпуклая функция, её надграфик $\operatorname{epi} f$ — непустое замкнутое выпуклое множество, причем точка $\big(x, f^{**}\big(x\big)\big)$ ему не принадлежит. По второй теореме об отделимости существуют такие $a \in \mathbb{R}^d$, $b, c_1, c_2 \in \mathbb{R}$, что для любых $(z, s) \in \operatorname{epi} f$ выполнено

$$\langle a, z\rangle + bs \le c_1 < c_2 \le \langle a, x\rangle + bf^{**}\big(x\big).$$

Переставляя члены, сделаем вывод, что

$$\langle a, z - x\rangle + b(s - f^{**}\big(x\big)) < 0.$$

- $b > 0$. Неравенство нарушается, если взять достаточно большое s при фиксированном z.

- $b < 0$. Тогда разделим последнее неравенство на b и обозначим $y = -\frac{a}{b}$. Получим

  $$\langle y, z - x\rangle + f^{**}(x) - s < 0.$$

  Выберем $s = f(z)$, тогда

  $$\langle y, z\rangle - f(z) - \langle y, x\rangle + f^{**}(x) < 0.$$

  Взяв максимум по z, воспользуемся определением сопряжённой функции. Тогда

  $$f^*(y) + f^{**}(x) < \langle y, x\rangle.$$

  В то же время, из определения сопряжённой функции легко заметить

  $$f^*(y) + f^{**}(x) \geq \langle y, x\rangle.$$

  Получили противоречие.

- $b = 0$. Поскольку f — собственная выпуклая функция, существует $\hat{y} \in \operatorname{dom} f$. Определим $\hat{a} = a + \varepsilon\hat{y}$, $\hat{b} = -\varepsilon$. Тогда для $z \in \operatorname{dom} f$ выполнено

  $$\begin{aligned}\langle \hat{a}, z - x\rangle + \hat{b}(s - f^{**}(x)) &= \langle a, z - x\rangle + \varepsilon\big[\langle \hat{y}, z\rangle - f(z) + f^{**}(x) - \langle \hat{y}, x\rangle\big] \\ &\leq (c_1 - c_2) + \varepsilon\big[f^*(\hat{y}) - f^{**}(x) - \langle \hat{y}, x\rangle\big].\end{aligned}$$

  Здесь мы воспользовались второй теоремой об отделимости и взяли максимум по $\hat{y}$. Поскольку $c_1 - c_2 < 0$, а выбор ε произвольный, то его можно взять достаточно малым, чтобы сделать правую часть неравенства отрицательной. Таким образом,

  $$\langle \hat{a}, z - x\rangle + \hat{b}(s - f^{**}(x)) < 0,$$

  и мы свели случай $b = 0$ к первому пункту, который уже был рассмотрен. Получили противоречие.

■

Теорема C7.4 позволяет доказать желаемое утверждение о связи между сопряжением и субдифференцированием

Теорема C7.5. Пусть $f : \mathbb{R}^d \to \overline{\mathbb{R}}$ — собственная замкнутая выпуклая функция. Тогда следующие три условие эквивалентны:

1. $f(x) + f^*(y) = \langle x, y\rangle$;
2. $y \in \partial f(x)$;
3. $x \in \partial f^*(y)$.

Доказательство.

- Пусть $y \in \partial f(x)$. Тогда по определению выпуклости для любого $z \in \mathbb{R}^d$ выполнено

$$f(z) \geq f(x) + \langle y, z - x \rangle.$$

Переставляя части неравенства и максимизируя по z, получим

$$\langle y, x \rangle - f(x) \geq f^*(y).$$

Учитывая полученное в предыдущем доказательстве обратное неравенство, заключаем

$$f(x) + f^*(y) = \langle x, y \rangle.$$

Обратим внимание, что все переходы можно проделать и в обратном направлении, что доказывает эквивалентность.

- Поскольку функция f собственная выпуклая и замкнутая, по Теореме C7.4 имеем $f^{**} = f$. Это означает, что $f(x) + f^*(y) = \langle x, y \rangle$ эквивалентно равенству

$$f^{**}(x) + f^*(y) = \langle x, y \rangle.$$

По доказанной в первом пункте эквивалентности имеем $x \in \partial f^*(y)$.

■

Рассмотрим условие $f(x) + f^*(y) = \langle x, y \rangle$ более детально. По определению сопряжённой функции его можно эквивалентно переписать в виде

$$x \in \operatorname*{argmax}_{\hat{x} \in \mathbb{R}^d} \{\langle y, \hat{x} \rangle - f(\hat{x})\}.$$

Аналогично, учитывая равенство f со второй сопряжённой, запишем

$$y \in \operatorname*{argmax}_{\hat{y} \in \mathbb{R}^d} \{\langle x, \hat{y} \rangle - f^*(\hat{y})\}.$$

Но мы доказали эквивалентность указанного равенства и включения точек в субдифференциалы. Тогда

$$\partial f(x) = \operatorname*{argmax}_{\hat{y} \in \mathbb{R}^d} \{\langle x, \hat{y} \rangle - f^*(\hat{y})\},$$
$$\partial f^*(y) = \operatorname*{argmax}_{\hat{x} \in \mathbb{R}^d} \{\langle y, \hat{x} \rangle - f(\hat{x})\}.$$

Пример C7.16. Найдите субградиент функции $f : \mathbb{R}^d \to \mathbb{R}$:

$$f(x) = \|x\|.$$

Решение. Норма — собственная замкнутая выпуклая функция. Таким образом, имеем

$$\partial f(0) = \operatorname*{argmin}_{y \in \mathbb{R}^d} \|y\|^*.$$

Из Примера C7.12 известно, что минимум сопряжённой функции к норме равен нулю и достигается в шаре по сопряжённой норме. Таким образом

$$\partial f(0) = \mathcal{B}_{\|\cdot\|_*}(0, 1).$$

Аналогичный результат был получен в предыдущей главе. ■

C8 Двойственность по Лагранжу

Некоторые оптимизационные задачи требуют не столько отыскания подходящих параметров модели, сколько оценки на значение целевой функции в оптимуме. Например, в задаче логистики не обязательно требуется отыскать точный маршрут для каждой машины, но важно знать, сколько в принципе будет стоить оптимальная доставка. Поскольку решение задачи со сложными ограничениями может быть сопряжено с рядом трудностей, на помощь приходит аппарат двойственных функций Лагранжа.

C8.1 Лагранжиан

Начнем с рассмотрения постановки задачи оптимизации стандартной формы:

$$\begin{aligned} \min_x \quad & f_0(x) \\ \text{s.t.} \quad & f_i(x) \le 0, \quad i = \overline{1,n} \\ & h_j(x) = 0, \quad j = \overline{1,m}. \end{aligned} \tag{C8.1}$$

Чтобы (C8.1) имела смысл, потребуем не пустоту пересечения эффективных множеств функций-ограничений и целевой функции. Перепишем ее в эквивалентном виде:

$$\min_x \quad f_0(x) + \sum_{i=1}^{n} I_-(f_i(x)) + \sum_{j=1}^{m} I_0(h_j(x)), \tag{C8.2}$$

где I_- и I_0 — индикаторные функции не положительных и нулевых значений соответственно:

$$I_-(x) = \begin{cases} 0, & x \le 0, \\ +\infty, & x > 0, \end{cases} \qquad I_0(x) = \begin{cases} 0, & x = 0, \\ +\infty, & x \ne 0. \end{cases}$$

В задаче (C8.2) целевая функция регуляризуется индикаторами, которые служат жестким штрафом за нахождение вне требуемого множества. Поскольку отыскать минимум (C8.2) невозможно в силу не дифференцируемости функции, имеет смысл перейти к более мягким ограничениям на x. Заменим индикаторы на линейные функции $\lambda_i f_i$ и $\nu_j h_j$. Тогда функция, которую мы минимизируем, становится *лагранжианом*.

Определение C8.1. *Лагранжиан* $\mathcal{L}$ относительно задачи оптимизации (C8.1) задается следующим образом:

$$\mathcal{L}(x, \lambda, \nu) = f_0(x) + \sum_{i=1}^{n} \lambda_i f_i(x) + \sum_{j=1}^{m} \nu_j h_j(x).$$

Лагранжиан является линейной аппроксимацией ограничений исходной задачи. Переменные $\lambda \in \mathbb{R}^n$ и $\nu \in \mathbb{R}^m$ будем называть *двойственными переменными*, в то время как x — прямой переменной.

C8.2 Двойственная функция по Лагранжу

Основная идея подхода через функцию Лагранжа заключается в переходе от оптимизационной задачи по прямым переменным к задаче по двойственным. Как результат минимизации лагранжиана по x, определим *двойственную функцию Лагранжа*. Иногда ее также называют *сопряженной функцией по Лагранжу*.

Определение C8.2. *Двойственная функция по Лагранжу* к задаче (C8.1) $g : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}$ задаётся следующим образом:

$$g(\lambda, \nu) = \inf_{x \in \mathbb{R}^d} \left(f_0(x) + \sum_{i=1}^{n} \lambda_i f_i(x) + \sum_{i=1}^{m} \nu_j h_j(x) \right). \tag{C8.3}$$

Замечание C8.1. Возможна ситуация, что при некоторых наборах двойственных переменных (λ, ν) лагранжиан L является неограниченным снизу по переменной x. Тогда будем говорить, что двойственная функция принимает значение $-\infty$ на этих наборах.

Далее мы докажем одно из ключевых практических свойств сопряжения по Лагранжу.

Утверждение C8.1. Двойственная функция, построенная по определению (C8.3), всегда является вогнутой по переменным (λ, ν), вне зависимости от того, является ли $f_0(x)$ выпуклой или нет.

Доказательство. Рассмотрим $(\lambda^1, \nu^1), (\lambda^2, \nu 62)$ и $\alpha \in [0, 1]$, тогда:

$$\begin{aligned}
&g\big(\alpha\lambda^1 + (1-\alpha)\lambda^2, \alpha\nu^1 + (1-\alpha)\nu^2\big) \\
&= \inf_{x \in \mathbb{R}^d} \left(f_0(x) + \sum_{i=1}^{n} \big(\alpha\lambda_i^1 + (1-\alpha)\lambda_i^2\big) f_i(x) + \sum_{j=1}^{m} \big(\alpha\nu_j^1 + (1-\alpha)\nu_j^2\big) h_j(x) \right) \\
&= \inf_{x \in \mathbb{R}^d} \big(\alpha \mathcal{L}(x, \lambda^1, \nu^1) + (1-\alpha)\mathcal{L}(x, \lambda^2, \nu^2)\big) \\
&\geq \alpha \inf_{x \in \mathbb{R}^d} \mathcal{L}(x, \lambda^1, \nu^1) + (1-\alpha) \inf_{x \in \mathbb{R}^d} \mathcal{L}(x, \lambda^2, \nu^2) \\
&= \alpha g(\lambda^1, \nu^1) + (1-\alpha) g(\lambda^2, \nu^2).
\end{aligned}$$

Следовательно, $g(\lambda, \nu)$ является вогнутой по определению. ■

Замечание C8.2. Можно доказывать этот факт, ссылаясь на доказанное ранее свойство поточечного супремума аффинных функций.

В начале главы было упомянуто, что аппарат сопряженных функций Лагранжа полезен для получения оценки на значение функции в оптимуме (обозначим его как p^*). Ниже мы доказываем, что на любом наборе (λ, ν), сопряженная функция Лагранжа оценивает p^* снизу.

Утверждение C8.2. Рассмотрим задачу (C8.1) как p^*. Для любого $\lambda \succeq 0$ и любого ν выполняется:

$$g(\lambda, \nu) \leq p^*.$$

Доказательство. Пусть $\overline{x}$ — достижимая точка, то есть $f_i(\overline{x}) \leq 0$ и $h_j(\overline{x}) = 0$. Тогда, при условии $\lambda \succeq 0$, мы имеем

$$\sum_{i=1}^{n} \lambda_i f_i(\overline{x}) + \sum_{j=1}^{m} \nu_j h_j(\overline{x}) \leq 0.$$

Тогда, используя это неравенство в определении лагранжиана, мы получаем

$$\mathcal{L}(\overline{x}, \lambda, \nu) = f_0(\overline{x}) + \sum_{i=1}^{n} \lambda_i f_i(\overline{x}) + \sum_{j=1}^{m} \nu_j h_j(\overline{x}) \leq f_0(\overline{x}).$$

Как следствие,

$$g(\lambda, \nu) = \inf_{x \in \mathbb{R}^d} \mathcal{L}(x, \lambda, \nu) \leq \mathcal{L}(\overline{x}, \lambda, \nu) \leq f_0(\overline{x}).$$

Так как это неравенство выполняется для любой достижимой точки (т.е. выполняются ограничения типа неравенств и равенств), то мы получаем искомый результат. ■

C8.3 Двойственная задача

Утверждение C8.2 доказывает, что двойственная функция Лагранжа дает нижнюю оценку на оптимальное значение задачи (C8.1), напрямую зависящую от λ и ν. Интуитивно понятно, что лучшая нижняя оценка на p^* получается, если решить задачу максимизации по двойственным переменным:

$$\begin{aligned} \max_{\lambda, \nu} \quad & g(\lambda, \nu) \\ \text{s.t.} \quad & \lambda \succeq 0. \end{aligned}$$

Такая задача называется *двойственной задачей* к задаче (C8.1).
Обозначим оптимальное значение двойственной задачи относительно начальной как d^*. Утверждение C8.2 доказывает соотношение

$$d^* \leq p^*.$$

Это неравенство выполняется всегда. Соответствующее свойство называется *слабой двойственностью*. В частности, слабая двойственность выполняется и при бесконечных d^* и p^*. Если $p^* = -\infty$, то начальная задача не ограничена снизу и, как следствие, d^* обязана равняться $-\infty$. То же самое происходит и в случае $d^* = +\infty$.
Когда оптимальные значения двойственной и прямой задачи совпадают, говорят о *сильной двойственности*. Существует множество условий, позволяющих проверить выполнение сильной двойственности. Одним из них является *условие Слейтера*.

Утверждение C8.3. Рассмотрим задачу следующего вида:

$$\begin{aligned} \min_{x} \quad & f_0(x) \\ \text{s.t.} \quad & f_i(x) \leq 0, \quad i = \overline{1, n} \\ & Ax = b, \end{aligned} \tag{C8.4}$$

где $f_0, \ldots, f_n$ — выпуклые функции. Если существует такой $\overline{x} \in \operatorname{relint} \mathbb{R}^d$, что

$$f_i(\overline{x}) < 0, \ i = \overline{1, n}, \ A\overline{x} = b,$$

то для задачи (C8.4) выполняется сильная двойственность.

C8.4 Примеры

Пример C8.1. Рассмотрите задачу оптимизации:

$$\begin{aligned} \min_{x} \quad & x^\top x \\ \text{s.t.} \quad & Ax = b, \end{aligned}$$

где $A \in \mathbb{R}^{n \times d}$, $b \in \mathbb{R}^n$. Поставьте для нее двойственную задачу.

Решение. Выпишем лагранжиан:

$$\mathcal{L}(x, \nu) = x^\top x + \nu^\top \big(Ax - b\big).$$

Здесь отсутствуют ограничения типа неравенства, поэтому из двойственных переменных у нас только ν. Минимизируем лагранжиан по x. Так как это квадратичная форма, которая к тому же является выпуклой, достаточно градиент приравнять к 0.

$$\nabla_x \mathcal{L}(x, \nu) = 2x + A^\top \nu = 0.$$

Следовательно, $x = -\frac{1}{2} A^\top \nu$. Тогда подставим этот x в $\mathcal{L}$ и получим $g(\nu)$:

$$g(\nu) = \frac{1}{4} \nu^\top A A^\top \nu - \frac{1}{2} \nu^\top A A^\top \nu - \nu^\top b = -\frac{1}{4} \nu^\top A A^\top \nu - \nu^\top b.$$

Тогда двойственная задача к исходной задаче оптимизации будет иметь следующий вид:

$$\max_{\nu} \quad \left[-\frac{1}{4} \nu^\top A A^\top \nu - \nu^\top b \right].$$

■

Пример C8.2. Рассмотрите задачу оптимизации:

$$\begin{aligned} \min_{x} \quad & c^\top x \\ \text{s.t.} \quad & Ax = b \\ & x \succeq 0, \end{aligned}$$

где $A \in \mathbb{R}^{n \times d}$, $b \in \mathbb{R}^n$. Поставьте для нее двойственную задачу.

Решение. Выпишем лагранжиан:

$$\mathcal{L}(x, \lambda, \nu) = c^\top x - \sum_{i=1}^{d} \lambda_i x_i + \nu^\top (Ax - b) = -\nu^\top b + \Big(c + A^\top \nu - \lambda\Big)^\top x.$$

Двойственная функция равна

$$g\big(\lambda, \nu\big) = \inf_{x \in \mathbb{R}^d} \mathcal{L}(x, \lambda, \nu) = -\nu^\top b + \inf_{x} \Big(c + A^\top \nu - \lambda\Big)^\top x.$$

Заметим, что если $c + A^\top \nu - \lambda \neq 0$, то инфимум равен $-\infty$, так как это линейная функция. Единственный случай, когда инфимум не равен $-\infty$: $c + A^\top \nu - \lambda = 0$. Тогда

$$g(\lambda, \nu) = \begin{cases} -\nu^\top b, & c + A^\top \nu - \lambda = 0, \\ -\infty, & \text{иначе.} \end{cases}.$$

Тогда, поскольку интересующий нас случай — случай, когда $g > -\infty$, то для построения двойственной задачи в более удобном для нас виде мы можем ввести искусственное ограничение:

$$\begin{aligned} \max_{\nu,\lambda} \quad & -\nu^\top b \\ \text{s.t.} \quad & c + A^\top \nu - \lambda = 0 \\ & \lambda \succeq 0. \end{aligned}$$

■

Замечание C8.3. Добавление такого искусственного ограничения валидно и с точки зрения математики, и с точки зрения идеи. Поскольку мы максимизируем двойственную функцию, то очевидно, что лучше рассматривать случай, когда $g\big(\lambda, \mu\big) > -\infty$. Условие, при котором это выполняется, выносится в ограничение, сразу избавляя от необходимости рассматривать «плохие» случаи.

Пример C8.3. Рассмотрим задачу оптимизации:

$$\begin{aligned} \min_{x \in \mathbb{R}^d} \quad & c^\top x \\ \text{s.t.} \quad & x^\top A x \leq 1, \end{aligned}$$

где $A \in \mathbb{S}^d_{++}$. Поставьте для нее двойственную задачу и вычислите значение p^*.

Решение. Выпишем лагранжиан:

$$\mathcal{L}(x, \lambda) = c^\top x + \lambda(x^\top A x - 1).$$

Заметим, что это выпуклая функция по x (сумма квадратичной функции с положительно полуопределенной матрицей и линейной функции с неотрицательными коэффициентами). Будем минимизировать ее аналитически.

$$\nabla_x \mathcal{L}(x, \lambda) = c + 2\lambda A x.$$

Приравняв градиент к нулю, получим $x = -\frac{1}{2\lambda} A^{-1} c$. Тогда двойственная функция имеет вид

$$g\big(\lambda\big) = -\frac{1}{2\lambda} c^\top A^{-1} c + \frac{1}{4\lambda} c^\top A^{-1} A A^{-1} c - \lambda = -\frac{1}{4\lambda} c^\top A^{-1} c - \lambda.$$

Ей соответствует двойственная задача

$$\begin{aligned} \max_{\lambda} \quad & \left[-\frac{1}{4\lambda} c^\top A^{-1} c - \lambda \right] \\ \text{s.t.} \quad & \lambda \geq 0. \end{aligned}$$

Отметим, что условие Слейтера для рассматриваемой задачи выполнено, поскольку выпуклое ограничение строго выполнено, например, при $x = 0$. Производная g по λ имеет вид

$$g'(\lambda) = \frac{1}{4\lambda^2} c^\top A^{-1} c - 1.$$

Приравнивая её к нулю, получим точку минимума:

$$\lambda = \frac{1}{2} \sqrt{c^\top A^{-1} c}.$$

Подставив её в двойственную задачу, получим:

$$p^* = -\sqrt{c^\top A^{-1} c}.$$

■

Пример C8.4. Рассмотрите задачу оптимизации:

$$\begin{aligned} \min_{x \in \mathbb{R}^d} \quad & c^\top x \\ \text{s.t.} \quad & l \preceq x \preceq u, \end{aligned}$$

где $l, u \in \mathbb{R}^d$. Поставьте для неё двойственную задачу.

Решение. Выпишем лагранжиан:

$$\mathcal{L}(x, \lambda_1, \lambda_2) = c^\top x + \lambda_1^\top (l - x) + \lambda_2^\top (x - u) = (c - \lambda_1 - \lambda_2)^\top x + \lambda_1^\top l - \lambda_2^\top u.$$

Это линейный функционал, заданный на всем множестве. Понятно, что при $c - \lambda_1 - \lambda_2 \neq 0$ он может быть устремлен к $-\infty$. Таким образом, приходим к двойственной функции вида

$$g(\lambda_1, \lambda_2) = \begin{cases} \lambda_1^\top l - \lambda_2^\top u, & c - \lambda_1 + \lambda_2 = 0, \\ -\infty, & \text{иначе.} \end{cases}$$

Внося $c - \lambda_1 + \lambda_2 = 0$ в ограничения, получим двойственную задачу:

$$\begin{aligned} \max_{\lambda_1, \lambda_2} \quad & \left[\lambda_1^\top l - \lambda_2^\top u\right] \\ \text{s.t.} \quad & \lambda_1, \lambda_2 \succeq 0 \\ & c - \lambda_1 + \lambda_2 = 0. \end{aligned}$$

■

Пример C8.5. Рассмотрите задачу оптимизации линейного функционала на вероятностном симплексе:

$$\begin{aligned} \min_{x \in \mathbb{R}^d} \quad & c^\top x \\ \text{s.t.} \quad & \mathbf{1}^\top x = 1 \\ & x \geq 0. \end{aligned}$$

Поставьте для нее двойственную задачу и вычислите значение p^*.

Решение. Выпишем лагранжиан:

$$\mathcal{L}(x, \lambda, \nu) = c^\top x + \nu(\mathbf{1}^\top x - 1) - \lambda^\top x = (c + \boldsymbol{\nu} - \lambda)^\top x - \nu.$$

Аналогично предыдущему примеру, получим выражение для двойственной функции:

$$g(\lambda, \nu) = \begin{cases} -\nu, & c + \boldsymbol{\nu} - \lambda = 0, \\ -\infty, & \text{иначе.} \end{cases}$$

Тогда двойственная задача имеет вид:

$$\begin{aligned} \max_{\lambda,\nu} \quad & -\nu \\ \text{s.t.} \quad & \lambda \succeq 0 \\ & c + \boldsymbol{\nu} - \lambda = 0. \end{aligned}$$

Обсудим структуру полученного ограничения вида равенства. Если некоторая координата c_i неотрицательна, это не вызывает проблем, поскольку есть очевидное допустимое значение $\lambda_i = c_i$, $\nu = 0$. Однако, случай $c_i < 0$ требует более тонкого анализа, потому что соответствующая координата λ_i не может быть выбрана отрицательной. Это означает, что требуется выбрать $\nu = -\min_i c_i$. Это единственный выбор, который не вызывает проблем с удовлетворением ограничений. Таким образом, имеем $d^* = \min_i c_i$. Условие Слейтера очевидно выполнено, поскольку ограничения имеют вид аффинного равенства, а также неравенства, заданного выпуклой функцией. Это означает $p^* = \min_i c_i$. В следующем параграфе, посвящённом теореме Каруша-Куна-Таккера мы получим решение прямой задачи и продемонстрируем, что двойственная задача действительно дает правильную оценку значения функционала в оптимуме. ■

Пример C8.6. Рассмотрите задачу оптимизации:

$$\begin{aligned} \min_{x \in \mathbb{R}^d} \quad & x^\top W x \\ \text{s.t.} \quad & x_i^2 = 1, \quad i = \overline{1,d}, \end{aligned}$$

где $W \in \mathbb{S}^d$. Постройте для нее двойственную задачу и получите нижнюю оценку на p^*.

Решение. Выпишем лагранжиан:

$$\mathcal{L}(x,\nu) = x^\top W x + \sum_{i=1}^{d} \nu_i \big(x_i^2 - 1\big) = x^\top \big(W + \operatorname{diag}(\nu)\big) x - \mathbf{1}^\top \nu.$$

Двойственная функция равна:

$$g(\nu) = \begin{cases} -\mathbf{1}^\top \nu, & W + \operatorname{diag}(\nu) \succeq 0, \\ -\infty, & \text{иначе}, \end{cases}$$

где мы воспользовались фактом, что минимум квадратичной формы равен нулю, если форма положительно полуопределена, или же, в противном случае, равен $-\infty$. Тогда двойственная задача имеет вид:

$$\begin{aligned} \max_{\nu} \quad & -\mathbf{1}^\top \nu \\ \text{s.t.} \quad & W + \operatorname{diag}(\nu) \succeq 0. \end{aligned}$$

В отличие от предыдущих примеров, в этом не так легко сразу вывести хорошую нижнюю оценку на p^*. Если взять $\nu = -\lambda_{min}(W)\mathbf{1}$, то можно заметить, что

$$W + \operatorname{diag}(\nu) = W - \lambda_{min}(W) I_d \succeq 0.$$

Посчитав значение двойственной функции при таком значении переменной, получим

$$g\big(\nu\big) = -\mathbf{1}^\top \nu = d\lambda_{min}(W) \leq p^*.$$

■

Замечание C8.4. В прикладных сценариях, когда невозможно аналитически минимизировать лагранжиан по прямым переменным, могут быть использованы методы поиска седловой точки. Инициализируя x^0, λ^0, ν^0, итеративный алгоритм обновляет прямые и двойственные переменные одновременно. В качестве критерия остановки обычно используют $f\big(x^k\big) - g\big(\lambda^k, \nu^k\big) \leq \varepsilon$. Поэтому, когда будет исследоваться график невязки между прямой и двойственной функцией, станет понятно, выполняется сильная двойственность, или же нет: критерий должен стремиться к $p^* - d^*$ — так называемому *оптимальному двойственному зазору.*

C8.5 Связь сопряжения Фенхеля и Лагранжа

В качестве мотивации, рассмотрим пример, который не имеет практического смысла, но поможет уловить связь между сопряжением по Фенхелю и Лагранжу. Рассмотрим задачу:

$$\begin{aligned} \min_{x \in \mathbb{R}^d} \quad & f\big(x\big) \\ \text{s.t.} \quad & x = 0. \end{aligned}$$

Лагранжиан имеет вид

$$\mathcal{L}(x, \nu) = f\big(x\big) + \nu^\top x.$$

Тогда двойственная функция Лагранжа есть

$$g\big(\nu\big) = \inf_{x \in \mathbb{R}^d} \big(f\big(x\big) + \nu^\top x\big) = -\sup_x \big(-\nu^\top x - f\big(x\big)\big) = -f^*\big(-\nu\big).$$

Проделаем то же самое для более общей задачи

$$\begin{aligned} \min_{x \in \mathbb{R}^d} \quad & f\big(x\big) \\ \text{s.t.} \quad & Ax \preceq b \\ & Cx = d. \end{aligned}$$

Ее лагранжиан имеет вид

$$\mathcal{L}(x, \lambda, \nu) = f\big(x\big) + \lambda^\top (Ax - b) + \nu^\top (Cx - d),$$

а двойственная функция равна

$$\begin{aligned} g\big(\lambda, \nu\big) &= \inf_{x \in \mathbb{R}^d} \big(f\big(x\big) + \lambda^\top (Ax - b) + \nu^\top (Cx - d)\big) \\ &= -\lambda^\top b - \nu^\top d + \inf_{x \in \mathbb{R}^d} \left(f\big(x\big) + \left(A^\top \lambda + c^\top \nu\right)^\top x\right) \\ &= -\lambda^\top b - \nu^\top d - f^*\big(-A^\top \lambda - c^\top \nu\big). \end{aligned}$$

Как можно заметить, в ряде случаев возможно «убрать» ограничения на прямую переменную в аргумент двойственной функции при построении двойственной задачи. Более того, для задач с линейными ограничениями удается выписать сопряжение по Лагранжу, зная лишь сопряженную функцию Фенхеля для целевого функционала.

Пример C8.7. Рассмотрите задачу оптимизации:

$$\begin{aligned} \min_{x \in \mathbb{R}^d} \quad & \|x\| \\ \text{s.t.} \quad & Ax = b, \end{aligned}$$

Напомним, что если $f(x) = \|x\|$, то

$$f^*(y) = \begin{cases} 0, & \|y\|_* \leq 1, \\ +\infty, & \|y\|_* > 1. \end{cases}$$

Постройте для нее двойственную задачу.

Решение. Используя результат из введения к этой секции, имеем

$$g(\nu) = -\nu^\top b - f^*(-A^\top \nu) = \begin{cases} -\nu^\top b, & \|A^\top \nu\|_* \leq 1, \\ -\infty, & \text{иначе.} \end{cases}$$

Соответственно, двойственная задача имеет следующий вид:

$$\begin{aligned} \max_{\nu} \quad & -\nu^\top b \\ \text{s.t.} \quad & \|A^\top \nu\|_* \leq 1. \end{aligned}$$

■

Пример C8.8. Рассмотрите задачу оптимизации:

$$\begin{aligned} \min_{x \in \mathbb{R}^d_{++}} \quad & \sum_{i=1}^d x_i \log x_i \\ \text{s.t.} \quad & Ax \preceq b \\ & \mathbf{1}^\top x = 1. \end{aligned}$$

Постройте для нее двойственную задачу.

Решение. Чтобы найти двойственную функцию Лагранжа, необходимо найти сопряженную функцию Фенхеля для отрицательной энтропии. Она была найдена в прошлой главе:

$$f^*(y) = \sum_{i=1}^d e^{y_i - 1}.$$

Пользуясь результатом о связи сопряженной и двойственной функции, имеем

$$g(\lambda, \nu) = -\lambda^\top b - \nu - \sum_{i=1}^d e^{-a_i^\top \lambda - \nu - 1},$$

где a_i — i-ый столбец матрицы A. Значит, двойственная задача имеет вид

$$\begin{aligned} \max_{\lambda, \nu} \quad & \left[-\lambda^\top b - \nu - \sum_{i=1}^d e^{-a_i^\top \lambda - \nu - 1} \right] \\ \text{s.t.} \quad & \lambda \succeq 0. \end{aligned}$$

■

Замечание C8.5. Стоит отметить, что формально нужно было занести ограничение $x \in \mathbb{R}^d_{++}$ в лагранжиан. Однако, это ограничение было автоматически учтено при поиске сопряженной функции.

C8.6 Ввод искусственных ограничений

Некоторые задачи могут оказаться более простыми для решения, если искусственным образом перенести часть целевой функции в ограничения.

Пример C8.9. Рассмотрите задачу оптимизации:

$$\min_{x \in \mathbb{R}^d} \quad \sum_{i=1}^{N} \left\|A_i x + b_i\right\|_2 + \frac{1}{2}\left\|x - x_0\right\|_2^2,$$

где $A_i \in \mathbb{R}^{n \times d}$, $b_i \in \mathbb{R}^n$, $x_0 \in \mathbb{R}^d$. Постройте для нее двойственную задачу.

Решение. Переформулируем задачу, введя искусственные ограничения:

$$\begin{aligned} \min_{x \in \mathbb{R}^d} \quad & \sum_{i=1}^{N} \left\|y_i\right\|_2 + \frac{1}{2}\left\|x - x_0\right\|_2^2 \\ \text{s.t.} \quad & A_i x + b_i = y_i, \quad i = \overline{1, N}. \end{aligned}$$

Соответственно, лагранжиан равен

$$\mathcal{L}(x, y, \nu_1, \ldots, \nu_N) = \sum_{i=1}^{N} \left\|y_i\right\|_2 + \frac{1}{2}\left\|x - x_0\right\|_2^2 + \sum_{i=1}^{N} \nu_i^\top \left(y_i - A_i x - b_i\right).$$

Минимизируем лагранжиан относительно y_i. Заметим, что:

$$\inf_{y_i} \left(\left\|y_i\right\|_2 + \nu_i^\top y_i\right) = \begin{cases} 0, & \left\|\nu_i\right\|_2 \leq 1, \\ -\infty, & \left\|\nu_i\right\|_2 > 1. \end{cases}.$$

Если $\left\|\nu_i\right\|_2 > 1$, то можно выбрать $y_i = -t\nu_i$ и устремить $t \to \infty$. Если же $\left\|\nu_i\right\| \leq 1$, то по неравенству Коши-Буняковского-Шварца:

$$\left\|y_i\right\|_2 + \nu_i^\top y_i \geq 0.$$

Тогда достаточно взять $y_i = 0$.
Для минимизации по x достаточно взять градиент по x и приравнять его к нулю. Получим:

$$x = x_0 + \sum_{i=1}^{N} A_i^\top \nu_i.$$

Подставляя найденные переменные в лагранжиан, получаем:

$$g\big(\nu_1, \ldots, \nu_N\big) = \begin{cases} -\sum\limits_{i=1}^{N} \nu_i^\top (A_i x_0 + b_i) - \frac{1}{2}\left\|\sum\limits_{i=1}^{N} A_i^\top \nu_i\right\|_2^2, & \left\|\nu_i\right\|_2 \leq 1, \quad i = \overline{1, N} \\ -\infty, & \text{иначе.} \end{cases}$$

Тогда двойственная задача имеет вид:

$$\begin{aligned} \max_{\nu} \quad & \left[-\sum_{i=1}^{N} \nu_i^\top (A_i x_0 + b_i) - \frac{1}{2} \left\| \sum_{i=1}^{N} A_i^\top \nu_i \right\|_2^2 \right] \\ \text{s.t.} \quad & \|\nu_i\|_2 \leq 1, \quad i = \overline{1,N}. \end{aligned}$$

■

Пример С8.10. Рассмотрите задачу оптимизации:

$$\min_{x \in \mathbb{R}^d} \max_{i=\overline{1,n}} \left(a_i^\top x + b_i\right),$$

где $a_i \in \mathbb{R}^d$ и $b_i \in \mathbb{R}$. Постройте для неё двойственную задачу.

Решение. Переформулируем задачу, введя искусственное ограничение:

$$\begin{aligned} \min_{x \in \mathbb{R}^d} \max_{i=\overline{1,n}} \quad & y_i \\ \text{s.t.} \quad & a_i^\top x + b_i = y_i, \quad i = \overline{1,n}. \end{aligned}$$

Двойственная функция равна:

$$g(\nu) = \inf_{x,y} \left(\max_{i=\overline{1,n}} y_i + \sum_{i=1}^{n} \nu_i^\top \left(a_i^\top x + b_i - y_i\right) \right).$$

Инфимум по x конечен только тогда, когда $\sum\limits_{i=1}^{n} \nu_i^\top a_i = 0$. Минимизируя по y, получим

$$\inf_{y} \left[\max_i y_i - \nu^\top y \right] = \begin{cases} 0, & \nu \succeq 0, \ \mathbf{1}^\top \nu = 1, \\ -\infty, & \text{иначе.} \end{cases}$$

Чтобы это доказать, покажем, что если $\nu \succeq 0$, $\mathbf{1}^\top \nu = 1$, то

$$\nu^\top y = \sum_{j=1}^{n} \nu_j y_j \leq \sum_{j=1}^{n} \nu_j \max_i y_i = \max_i y_i.$$

Тогда, как следствие, взяв $y = 0$, получаем инфимум. Если же $\nu \not\succeq 0$, то тогда без ограничения общности рассмотрим $\nu_j < 0$. Возьмем $y_i = 0$, $i \neq j$ и $y_j = -t$. Тогда при стремлении $t \to +\infty$ имеем

$$\max_i y_i - \nu^\top y = 0 + t\nu_j \to -\infty.$$

И наконец, если $\mathbf{1}^\top \nu \neq 1$, тогда выберем $y = \mathbf{1}t$ и получим

$$\max_i y_i - \nu^\top y = t\left(1 - \mathbf{1}^\top \nu\right) \to -\infty,$$

где в зависимости от знака $1 - \mathbf{1}^\top \nu$ устремляем t к $+\infty$ или $-\infty$. В итоге двойственная функция записывается следующим образом:

$$g(\nu) = \begin{cases} b^\top \nu, & \sum\limits_{i=1}^{n} \nu_i a_i = 0, \ \nu \succeq 0, \ \mathbf{1}^\top \nu = 1, \\ -\infty, & \text{иначе.} \end{cases}$$

Значит, двойственная задача имеет вид:

$$\begin{aligned} \max_{\nu} \quad & b^\top \nu \\ \text{s.t.} \quad & A^\top \nu = 0 \\ & \nu \succeq 0 \\ & \mathbf{1}^\top \nu = 1. \end{aligned}$$

■

C9 Условия оптимальности Каруша-Куна-Такера

Ранее мы ввели понятие лагранжиана задачи и построили аппарат двойственных функций Лагранжа. В рамках этого семинара мы, предполагая нулевой двойственный зазор, продолжим анализ свойств лагранжиана и выведем условия оптимальности оптимизационной задачи. Будем рассматривать задачу минимизации функции f_0 с прямой переменной $x \in \mathbb{R}^d$, ограничениями равенства h_j и неравенства f_i:

$$\begin{aligned} \min_{x \in \mathbb{R}^d} \quad & f_0(x) \\ \text{s.t.} \quad & f_i(x) \leq 0, \quad i = \overline{1,n} \\ & h_j(x) = 0, \quad j = \overline{1,m}. \end{aligned} \tag{C9.1}$$

А также её двойственную задачу с максимизируемой функцией

$$g(\lambda, \nu) = \inf_{x \in \mathbb{R}^d} \left(f_0(x) + \sum_{i=1}^{n} \lambda_i f_i(x) + \sum_{j=1}^{m} \nu_j h_j(x) \right)$$

и двойственными переменными $\lambda \in \mathbb{R}^n$, $\nu \in \mathbb{R}^m$:

$$\begin{aligned} \max_{\lambda, \nu} \quad & g(\lambda, \nu) \\ \text{s.t.} \quad & \lambda \succeq 0. \end{aligned} \tag{C9.2}$$

Далее везде предполагаем дифференцируемость f_0, f_i, h_j и существование x^* — переменную, доставляющих оптимум в прямой задаче и (λ^*, ν^*) — переменные, доставляющие оптимум в двойственной задаче.

C9.1 Условия Каруша-Куна-Такера

Определение C9.1. Пусть x^* является оптимумом прямой задачи (C9.1). Говорят, что выполнены *условия Каруша-Куна-Такера*, если выполнены:

- Ограничения:

$$\begin{aligned} f_i(x^*) \leq 0, \quad & i = \overline{1,n}, \\ h_j(x^*) = 0, \quad & j = \overline{1,m}; \end{aligned}$$

- Неотрицательность:

$$\lambda_i^* \geq 0, \quad i = \overline{1,n};$$

- Дополняющая нежёсткость:

$$\lambda_i^* f_i(x^*) = 0, \quad i = \overline{1,n};$$

- Стационарность:

$$\nabla f_0(x^*) + \sum_{i=1}^{n} \lambda_i^* \nabla f_i(x^*) + \sum_{j=1}^{m} \nu_j^* \nabla h_j(x^*) = 0.$$

Если оптимизационная задача имеет нулевой двойственный зазор, то условия из Определения C9.1 оказываются необходимыми условиями оптимальности.

Теорема C9.1. Пусть x^* является оптимумом прямой задачи (C9.1), а (λ^*, ν^*) являются оптимумом двойственной задачи (C9.2). При этом пусть выполняется сильная двойственность. Тогда, если функции f_0, f_i, h_j дифференцируемы в точке x^*, то выполняются условия ККТ.

Доказательство. По определению допустимой точки, в x^* выполнено

$$h_j(x^*) = 0, \ \forall j = \overline{1,m};$$
$$f_i(x^*) \le 0, \ \forall i = \overline{1,n}.$$

Поскольку задача имеет нулевой двойственный зазор, то

$$f_0(x^*) = g(\lambda^*, \nu^*).$$

Воспользовавшись определением двойственной функции Лагранжа, запишем

$$\begin{aligned} f_0(x^*) &= \inf_{x \in \mathbb{R}^d} \left(f_0(x) + \sum_{i=1}^{n} \lambda_i^* f_i(x) + \sum_{j=1}^{m} \nu_j^* h_j(x) \right) \\ &= f_0(x^*) + \sum_{i=1}^{n} \lambda_i^* f_i(x^*) + \sum_{j=1}^{m} \nu_j^* h_j(x^*) \\ &= f_0(x^*) + \sum_{i=1}^{n} \lambda_i^* f_i(x^*) \\ &\le f_0(x^*), \end{aligned}$$

где $\lambda_i^* \ge 0$. Получаем, что:

$$\sum_{i=1}^{n} \lambda_i^* f_i(x^*) = 0.$$

Все слагаемые в этой сумме являются неположительными, из чего следует условие:

$$\lambda_i^* f_i(x^*) = 0, \quad i = \overline{1,n};$$

Физически это означает, что хотя бы одна из двух переменных (прямая либо двойственная) лежит на границе множества ограничений. Далее, так как x^* минимизирует $\mathcal{L}(x, \lambda^*, \nu^*)$, дифференциал $\mathcal{L}$ по x в точке x^* должен быть равен 0:

$$\nabla f_0(x^*) + \sum_{i=1}^{n} \lambda_i^* \nabla f_i(x^*) + \sum_{j=1}^{m} \nu_j^* \nabla h_j(x^*) = 0.$$

■

Замечание C9.1. Если функции не дифференцируемы в x^*, то можно использовать субградиент:

$$\partial f_0(x^*) + \sum_{i=1}^{n} \lambda_i^* \partial f_i(x^*) + \sum_{j=1}^{m} \nu_j^* \partial h_j(x^*) \ni 0.$$

Замечание C9.2. Отметим, что выполнение сильной двойственности существенно, чтобы условия ККТ были необходимыми. Рассмотрим задачу минимизации

$$\begin{aligned} \min_{x\in\mathbb{R}} \quad & x \\ \text{s.t.} \quad & x^2 \le 0. \end{aligned}$$

В этом примере не выполнено ни одной из условий сильной двойственности (в том числе условие Слейтера). Лагранжиан задачи имеет вид

$$\mathcal{L}(x,\lambda) = x + \lambda x^2.$$

Единственная допустимая (и при этом оптимальная) точка — $x^* = 0$. Однако, она не обращает в ноль первую производную лагранжиана по x. Таким образом, имеем точку минимума, в которой система условий ККТ несовместна.

Накладывая дополнительные условия на рассматриваемые функции, получим достаточные условия ККТ. Мы будем пользоваться следующей теоремой.

Теорема C9.2. Пусть функции f_0, f_i являются выпуклыми, а h_j — аффинными. Тогда если для набора переменных $\left(x^*, \lambda^*, \nu^*\right)$ выполнены все условия ККТ, то x^* является точкой глобального минимума прямой задачи (C9.1), а (λ^*, ν^*) — точкой глобального максимума двойственной (C9.2). При этом сильная двойственность достигается автоматически.

Доказательство. Рассмотрим некоторый набор точек $\left(\tilde{x}, \tilde{\lambda}, \tilde{\nu}\right)$, удовлетворяющий условиям ККТ. Поскольку $\tilde{\lambda} \succeq 0$, лагранжиан — выпуклая функция относительно x. Тогда из последнего условия следует, что $\tilde{x}$ — точка глобального минимума функции $\mathcal{L}(x, \tilde{\lambda}, \tilde{\nu})$. Таким образом, имеем

$$g\left(\tilde{\lambda}, \tilde{\nu}\right) = \mathcal{L}(\tilde{x}, \tilde{\lambda}, \tilde{\nu}) = f_0\left(\tilde{x}\right).$$

Таким образом, $\tilde{x}$ и $\left(\tilde{\lambda}, \tilde{\nu}\right)$ обнуляют двойственный зазор, а потому являются оптимальными прямыми и двойственными переменными. ■

Условия Каруша-Куна-Такера играют важную роль в оптимизации. В некоторых специальных случаях возможно решить задачу оптимизации аналитически. Более того, некоторые алгоритмы оптимизации могут быть интерпретированы как решение системы условий ККТ.

C9.2 Примеры

В этом параграфе рассмотрим классические задачи на применение условий ККТ.

Пример C9.1. Решите задачу минимизации

$$\begin{aligned} \min_{x,y\in\mathbb{R}} \quad & (x+1)^2 + (y+1)^2 \\ \text{s.t.} \quad & 2x + 3y \ge 5. \end{aligned}$$

Решение. Это выпуклая задача, то есть можно пользоваться теоремой ККТ для достаточных условий. Выпишем Лагранжиан задачи:

$$\mathcal{L}(x,y,\lambda) = (x+1)^2 + (y+1)^2 + \lambda(5 - 2x - 3y).$$

Запишем ограничения:

$$2x + 3y \geq 5.$$

Неотрицательность:

$$\lambda \geq 0.$$

Дополняющая нежёсткость:

$$\lambda(2x + 3y - 5) = 0.$$

Стационарность:

$$\frac{\partial \mathcal{L}}{\partial x}\,(x, y, \lambda) = 2(x + 1) - 2\lambda = 0,$$
$$\frac{\partial \mathcal{L}}{\partial y}\,(x, y, \lambda) = 2(y + 1) - 3\lambda = 0.$$

Пользуясь условиями на частные производные, получим

$$x = \lambda - 1,\ y = \frac{3}{2}\lambda - 1.$$

Далее переберем все варианты выполнения условия дополняющей нежёсткости:

1. $\lambda = 0$. Получаем, что $x = -1$, $y = -1$, но ограничение: $2x + 3y = -5 < 5$ не выполнено.

2. $\lambda > 0$. Получаем, что $(2x+3y-5) = 0$, тогда $2(\lambda-1)+3\big(\frac{3}{2}\lambda - 1\big)-5 = \frac{13}{2}\lambda-10 = 0$, поэтому

$$\lambda = \frac{20}{13},\ x^* = \frac{7}{13},\ y^* = \frac{17}{13}.$$

■

Пример С9.2. Решите задачу минимизации

$$\begin{aligned} \min_{x,y\in\mathbb{R}} \quad & x + 3y \\ \text{s.t.} \quad & x - y \geq 0, \\ & (x-1)^2 + (y-1)^2 \leq 9. \end{aligned}$$

Решение. Рассматриваемая задача выпуклая. Её лагранжиан имеет вид:

$$\mathcal{L}(x, y, \lambda_1, \lambda_2) = x + 3y + \lambda_1(y - x) + \lambda_2\big[(x-1)^2 + (y-1)^2 - 9\big].$$

Запишем ограничения:

$$x - y \geq 0,$$
$$(x-1)^2 + (y-1)^2 \leq 9.$$

Неотрицательность:

$$\lambda_1, \lambda_1 \geq 0.$$

Дополняющая нежёсткость:

$$\lambda_1(x-y)=0,$$
$$\lambda_2\big[(x-1)^2+(y-1)^2-9\big]=0.$$

Стационарность:

$$\frac{\partial\mathcal{L}}{\partial x}\,(x,y,\lambda_1,\lambda_2)=1-\lambda_1+2\lambda_2(x-1)=0,$$
$$\frac{\partial\mathcal{L}}{\partial y}\,(x,y,\lambda_1,\lambda_2)=3+\lambda_1+2\lambda_2(y-1)=0.$$

Рассмотрим все варианты условия дополняющей нежёсткости:

1. $\lambda_2=0$. Из условия на градиент $\lambda_1=1$, $\lambda_1=3$ — противоречие.
2. $\lambda_1=0$, $\lambda_2>0$. Получаем, что $x-1=-\frac{1}{2\lambda_2}$, $y-1=-\frac{3}{2\lambda_2}$ и $(x-1)^2+(y-1)^2=9$. Тогда $\frac{9}{4\lambda_2^2}+\frac{1}{4\lambda_2^2}=9$. Получаем
$$\lambda_1^*=0,\ \lambda_2^*=\frac{\sqrt{10}}{6},\ x^*=1-\frac{3}{\sqrt{10}},\ y^*=1-\frac{9}{\sqrt{10}}.$$
3. $\lambda_1>0$, $\lambda_2>0$. Получаем, что $(x-1)^2+(y-1)^2=9$, $x=y$, но тогда $x-1=y-1$ и по условию на градиенты $1-\lambda_1=3+\lambda_1\implies\lambda=-1<0$ — противоречие.

■

Пример C9.3. Решите задачу минимизации

$$\begin{aligned}\min_{x\in\mathbb{R}^d}\quad & \frac{1}{2}x^\top Px+q^\top x+r\\ \text{s.t.}\quad & Ax=b,\end{aligned}$$

где $P\in\mathbb{S}^d_+$, $A\in\mathbb{R}^{n\times d}$.

Решение. Заметим, что задача выпуклая в силу положительной полуопределенности матрицы. Лагранжиан задачи имеет вид

$$\mathcal{L}(x,\nu)=\frac{1}{2}x^\top Px+q^\top x+r+\nu^\top(Ax-b).$$

Запишем ограничения:

$$Ax=b$$

Стационарность:

$$\nabla_x\mathcal{L}(x,\nu)=Px+q+A^\top\nu=0.$$

Их можно переписать в виде

$$\begin{pmatrix}P & A^\top\\ A & 0\end{pmatrix}\cdot\begin{pmatrix}x\\ \nu\end{pmatrix}=\begin{pmatrix}-q\\ b\end{pmatrix}$$

По сути, мы имеем систему из $n+d$ уравнений на $n+d$ переменных x, ν. Решая эту систему, мы получаем решение прямой и двойственной задач. ■

ККТ позволяет вычислять аналитические выражения для проекций на относительно простые множества. Следующий пример рассматривает l_2 шар.

Пример С9.4. Решите задачу минимизации

$$\begin{aligned} \min_{x \in \mathbb{R}^d} \quad & \|x - s\|_2^2 \\ \text{s.t.} \quad & \|x\|_2^2 \leq 1, \end{aligned}$$

где $s \in \mathbb{R}^d$ — фиксированный вектор.

Решение. Заметим, что рассматриваемая задача выпуклая, то есть условия ККТ являются достаточными. Запишем лагранжиан задачи:

$$\mathcal{L}(x, \lambda) = \|x - s\|_2^2 + \lambda\Big(\|x\|_2^2 - 1\Big).$$

Запишем ограничения:

$$\|x\|_2^2 \leq 1.$$

Неотрицательность:

$$\lambda \geq 0.$$

Дополняющая нежёсткость:

$$\lambda\Big(\|x\|_2^2 - 1\Big) = 0.$$

Стационарность:

$$\nabla_x \mathcal{L}(x, \lambda) = 2(x - s) + 2\lambda x = 0 \implies x = \frac{s}{1 + \lambda}.$$

Рассмотрим два варианта условия дополняющей нежесткости:

1. $\lambda = 0$. Получаем, что $x = s$, но такой вариант подходит только когда $\|s\|_2 \leq 1$.
2. $\lambda > 0$. Получаем, что $\|x\|_2 = 1$, тогда $\|x\|_2 = \frac{1}{(1+\lambda)}\|s\|_2 = 1$. Следовательно $(1 + \lambda) = \|s\|_2$, это условие подходит, когда $\|s\|_2 > 1$. В этом случае, $x = \frac{s}{\|s\|_2}$.

Таким образом, оператор проекции на l_2-шар имеет вид

$$x^* = \min\left\{\|s\|_2, 1\right\} \cdot \frac{s}{\|s\|_2}.$$

■

ККТ также позволяет решать некоторые оптимизационные задачи в явном виде. Важным примером являются линейные задачи, поскольку теоретический анализ многих алгоритмов предполагает доступ к оракулу точного минимума линейной задачи. Следующий пример рассматривает l_2-шар.

Пример C9.5. Решите задачу минимизации

$$\begin{aligned} \min_{x\in\mathbb{R}^d} \quad & c^\top x \\ \text{s.t.} \quad & \|x\|_2^2 \le 1, \end{aligned}$$

где $c \in \mathbb{R}^d$ — фиксированный вектор.

Решение. Это выпуклая задача, то есть можно пользоваться ККТ как достаточными условиями. Лагранжиан задачи имеет вид

$$\mathcal{L}(x,\lambda) = c^\top x + \lambda\Big(\|x\|_2^2 - 1\Big).$$

Запишем ограничения:

$$\|x\|_2^2 \le 1.$$

Неотрицательность:

$$\lambda \ge 0.$$

Дополняющая нежёсткость:

$$\lambda\Big(\|x\|_2^2 - 1\Big) = 0.$$

Стационарность:

$$\nabla_x \mathcal{L}(x,\lambda) = c + 2\lambda x = 0 \implies 2\lambda x = -c.$$

Рассмотрим два варианта условия дополняющей нежесткости

1. $\lambda = 0$. Такой вариант подходит только когда $c = 0$ с любым $\|x\|_2 \le 1$.
2. $\lambda > 0$. Тогда $\|x\|_2 = 1$ и

$$\|x\|_2 = \frac{\|c\|_2}{2\lambda} = 1 \implies 2\lambda = \|c\|_2.$$

Отсюда следует

$$x = -\frac{c}{\|c\|_2}.$$

■

Рассмотрим также линейную задачу на симплексе. Это важный пример, поскольку с поиском оптимального вероятностного распределения приходится сталкиваться регулярно. В прошлой главе мы уже оценили оптимальное значение через минимизацию сопряженной функции Лагранжа. Теперь найдем оптимальные параметры модели.

Пример С9.6. Решите задачу минимизации

$$\begin{aligned} \min_{x\in\mathbb{R}^d} \quad & c^\top x \\ \text{s.t.} \quad & \mathbf{1}^\top x = 1 \\ & x \succeq 0, \end{aligned}$$

где $c \in \mathbb{R}^d$ — фиксированный вектор.

Решение. Запишем лагранжиан задачи:

$$\mathcal{L}(x, \lambda, \nu) = c^\top x + \nu\big(\mathbf{1}^\top x - 1\big) - \lambda^\top x = 0.$$

Запишем ограничения:

$$\begin{aligned} &\mathbf{1}^\top x = 1, \\ &x \succeq 0. \end{aligned}$$

Неотрицательность:

$$\lambda \succeq 0.$$

Дополняющая нежёсткость:

$$\lambda_i x_i = 0.$$

Стационарность:

$$\nabla_x \mathcal{L}(x, \lambda) = c + \boldsymbol{\nu} - \lambda = 0.$$

Рассмотрим последнее условие. В прошлой главе мы подробно обсуждали, что единственный подходящий выбор ν это

$$\nu = -\min_j c_j.$$

Таким образом, имеем

$$\lambda_i = c_i - \min_j c_j \geq 0.$$

Понятно, что i-я компонента вектора λ принимает нулевое значение только при условии

$$i = \operatorname*{argmin}_j \{c_j\}.$$

Обозначим множество подходящих индексов за I. Для всех остальных индексов $\lambda_i > 0$ и $x_i = 0$. Чтобы вектор x удовлетворял всем условиям ККТ, осталось подобрать его так, чтобы его компоненты суммировались в единицу. Подойдет любой x вида

$$x = \operatorname{conv}\{\, e_i \mid i \in I \,\}.$$

■

Выше мы упоминали, что условия ККТ можно переформулировать и для негладких задач, записав его через субдифференциалы. Рассмотрим минимзацию линейного функционала на l_1 шаре.

Пример C9.7. Решите задачу минимизации

$$\begin{aligned} \min_{x\in\mathbb{R}^d} \quad & c^\top x \\ \text{s.t.} \quad & \|x\|_1 \le 1, \end{aligned}$$

где $c \in \mathbb{R}^d$ — фиксированный вектор.

Решение. Лагранжиан задачи имеет вид:

$$\mathcal{L}(x,\lambda) = c^\top x + \lambda\big(\|x\|_1 - 1\big).$$

Запишем ограничения:

$$\|x\|_1 \le 1.$$

Неотрицательность:

$$\lambda \ge 0.$$

Дополняющая нежёсткость:

$$\lambda\big(\|x\|_1 - 1\big) = 0.$$

Стационарность:

$$\partial_x \mathcal{L}(x,\lambda) = c + \lambda\partial\big(\|x\|_1\big) = \big\{\, c + \lambda v \mid \|v\|_\infty \le 1,\ \langle x, v\rangle = \|x\|_1 \,\big\} \ni 0.$$

Из условия на субградиент следует равенство $c = -\lambda v$. Рассмотрим два варианта условия дополняющей нежёсткости.

1. $\lambda = 0$. Получаем, что $c = 0$, и в этом случае подходит любой x, такой что $\|x\|_1 \le 1$.

2. $\lambda > 0$. Тогда $\|x\|_1 = 1$. Чтобы достигалось равенство $\langle x, v\rangle = 1$ при $\|v\|_\infty \le 1$, по неравенству Гёльдера требуется:

$$\|v\|_\infty = 1,\ \lambda = \|c\|_\infty,\ v = -\frac{c}{\|c\|_\infty}.$$

   В качестве решения, на котором достигается равенство, берём вектор

$$x = -\operatorname{sign}(c_i)\cdot e_i,\ i = \operatorname*{argmax}_i |c_i|.$$

■

Перейдем к более сложным задачам, выходящим за рамки вычисления операторов проекции и минимизации линейных функционалов.

Пример С9.8. Решите задачу минимизации

$$\begin{aligned} \min_{x\in\mathbb{R}^d} \quad & -\sum_{i=1}^{d} \ln(\alpha_i + x_i) \\ \text{s.t.} \quad & x \succeq 0 \\ & \mathbf{1}^\top x = 1, \end{aligned}$$

где $\alpha_i > 0$.

Решение. Имеем дело с задачей выпуклой минимизации, поэтому будем пользоваться условиями ККТ как достаточными условиями минимума. Выпишем лагранжиан:

$$\mathcal{L}(x, \lambda, \nu) = -\sum_{i=1}^{d} \log(\alpha_i + x_i) - \lambda^\top x + \nu\big(\mathbf{1}^\top x - 1\big).$$

Запишем ограничения:

$$\begin{aligned} & x \succeq 0, \\ & \mathbf{1}^\top x = 1. \end{aligned}$$

Неотрицательность:

$$\lambda \succeq 0.$$

Дополняющая нежёсткость:

$$\lambda_i x_i = 0.$$

Стационарность:

$$\frac{\partial \mathcal{L}}{\partial x_i}\,(x, \lambda, \nu) = -\frac{1}{\alpha_i + x_i} - \lambda_i + \nu = 0.$$

Заметим, что от λ можно избавиться. Тогда получим:

$$\left(\nu - \frac{1}{\alpha_i + x_i}\right) x_i = 0, \quad \nu - \frac{1}{\alpha_i + x_i} \geq 0.$$

Если $\nu \geq \frac{1}{\alpha_i}$, то ситуация $x_i > 0$ невозможна в силу условия дополняющей нежесткости, тогда $x_i = 0$. Если $\nu < \frac{1}{\alpha_i}$, тогда $x_i > 0$. Получим:

$$\nu = \frac{1}{\alpha_i + x_i} \implies x_i = \frac{1}{\nu} - \alpha_i.$$

Решение можно объединить следующим образом:

$$x_i = \begin{cases} \frac{1}{\nu} - \alpha_i, & \nu < \frac{1}{\alpha_i}, \\ 0, & \nu \geq \frac{1}{\alpha_i}. \end{cases} \iff x_i = \max\left(\frac{1}{\nu} - \alpha_i, 0\right).$$

Подставляя это выражение во второе ограничение на x, имеем

$$\sum_{i=1}^{d} \max\left(\frac{1}{\nu} - \alpha_i, 0\right) = 1.$$

Если взять за переменную $\frac{1}{\nu}$, то это кусочно-линейная возрастающая функция, имеющая уникальное решение, которое можно точно определить. ■

Замечание C9.3. Пример C9.8 соответствует решению классической задаче про заполнение полости водой. На картинке ниже предоставлена её интуиция. Уровень заполнения воды это $\frac{1}{\nu}$. Для каждого уровня местности α_i подбирается x_i.

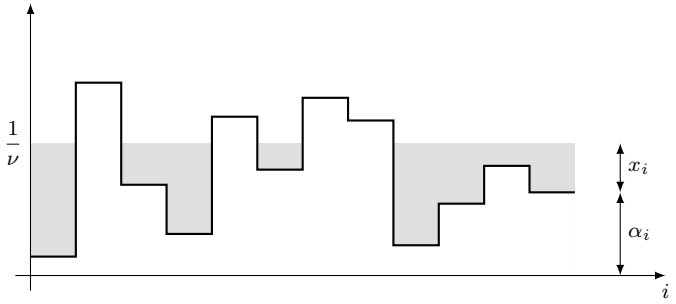


Рис. C9.1: Интуиция: уровень воды $1/\nu$, рельеф α_i, добавка x_i.

Пример C9.9. Решите задачу минимизации

$$\begin{aligned} \min_{x\in\mathbb{R}^d} \quad & \sum_{i=1}^d \frac{c_i}{x_i} \\ \text{s.t.} \quad & x \succeq \varepsilon \\ & \mathbf{1}^\top x = 1, \end{aligned}$$

где $c_i > 0$ и $\varepsilon > 0$.

Решение. Эта задача является задачей выпуклой оптимизации. Выпишем ее лагранжиан:

$$\mathcal{L}(x,\lambda,\nu) = \sum_{i=1}^d \frac{c_i}{x_i} + \sum_{i=1}^d \lambda_i(\varepsilon - x_i) + \nu\big(\mathbf{1}^\top x - 1\big).$$

Запишем ограничения:

$$\begin{aligned} & x \succeq \varepsilon, \\ & \mathbf{1}^\top x = 1. \end{aligned}$$

Неотрицательность:

$$\lambda \succeq 0.$$

Дополняющая нежёсткость:

$$\lambda_i(\varepsilon - x_i) = 0.$$

Стационарность:

$$\frac{\partial \mathcal{L}}{\partial x_i}\,(x,\lambda,\nu) = -\frac{c_i}{x_i^2} - \lambda_i + \nu = 0.$$

Пусть часть $x_i = \varepsilon$ и $\lambda_i \geq 0$ для $i \in I_1$, а другая $x_i > \varepsilon$ и $\lambda_i = 0$ для $i \in I_2$, так что $I_1 \cup I_2 = \{1, \ldots, d\}$ и $I_1 \cap I_2 = \varnothing$. Для $x_i > \varepsilon$, $i \in I_2$, получаем формулу:

$$x_i = \sqrt{\frac{c_i}{\nu}}.$$

Рассмотрим теперь остальные $x_i = \varepsilon$, $i \in I_1$. По условию нормировки имеем:

$$\sum_{i=1}^{d} x_i = \sum_{i \in I_1} \varepsilon + \sum_{i \in I_2} \sqrt{\frac{c_i}{\nu}} = 1 \implies \nu = \left(\frac{\sum_{i \in I_2} \sqrt{c_i}}{1 - |I_1|\varepsilon} \right)^2, \quad |I_1|\varepsilon < 1.$$

Далее посчитаем λ_i, $i \in I_1$ по условию на градиент:

$$\lambda_i = \nu - \frac{c_i}{\varepsilon^2}.$$

Итого, необходимо подобрать разбиение I_1, I_2, чтобы были выполнены следующие неравенства:

$$x_i = (1 - |I_1|\varepsilon) \frac{\sqrt{c_i}}{\sum_{i \in I_2} \sqrt{c_i}} > \varepsilon, \ \forall i \in I_2,$$
$$\lambda_i = \left(\frac{\sum_{i \in I_2} \sqrt{c_i}}{1 - |I_1|\varepsilon} \right)^2 - \frac{c_i}{\varepsilon^2} \geq 0, \ \forall i \in I_1.$$

Если все неравенства выполнены, получаем:

$$\begin{cases} x_i = (1 - |I_1|\varepsilon) \frac{\sqrt{c_i}}{\sum_{i \in I_2} \sqrt{c_i}}, & \forall i \in I_2, \\ x_i = \varepsilon, & \forall i \in I_1. \end{cases}$$

■

C9.3 Решение прямой задачи через двойственную

Предположим, что выполнена сильная двойственность. Выпишем задачу минимизации лагранжиана при оптимальных двойственных переменных (λ^*, ν^*):

$$\inf_x \mathcal{L}(x, \lambda^*, \nu^*).$$

Если у этой задачи единственный минимум, то он обязательно достигается в точке x^* глобального минимума прямой задачи (C9.1) ($f_0(x^*) = g(\lambda^*, \nu^*)$). Если минимум задачи минимизации лагранжиана не достигается, то и в прямой задаче он не достигается. Таким образом, зная оптимальные двойственные переменные (λ^*, ν^*), можно попробовать найти оптимальную прямую переменную x^* с помощью безусловной минимизации. Разберем несколько примеров.

Пример C9.10. Рассмотрите задачу минимизации отрицательной энтропии

$$\begin{aligned} \min_{x \in \mathbb{R}^d_{++}} \quad & \sum_{i=1}^{d} x_i \log x_i \\ \text{s.t.} \quad & Ax \preceq b \\ & \mathbf{1}^\top x = 1. \end{aligned}$$

Решение. Двойственная задача имеет вид:

$$\begin{aligned} \max_{\lambda,\nu} \quad & \left[-b^\top \lambda - \nu - e^{-\nu-1} \sum_{i=1}^{d} e^{-a_i^\top \lambda}\right] \\ \text{s.t.} \quad & \lambda \succeq 0, \end{aligned}$$

где a_i — столбцы матрицы A. Предположим, что выполняется условие Слейтера с линейными ограничениями неравенства, чтобы говорить о сильной двойственности. Запишем лагранжиан в оптимальных двойственных переменных:

$$\mathcal{L}(x, \lambda^*, \nu^*) = \sum_{i=1}^{d} x_i \log x_i + (\lambda^*)^\top (Ax - b) + \nu\big(\mathbf{1}^\top x - 1\big).$$

Эта функция является 1-сильно выпуклой, а значит имеет единственный минимум. Взяв градиент по x и приравняв его к нулю, получим:

$$x_i^* = \frac{1}{\exp\big\{\big(a_i^\top \lambda^* + \nu^* + 1\big)\big\}}, \; i = \overline{1, d}.$$

То есть, если x^* достижима, то она обязана доставлять оптимум. ■

Пример С9.11. Рассмотрим задачу минимизации сепарабельной функции

$$\begin{aligned} \min_{x \in \mathbb{R}^d} \quad & \sum_{i=1}^{d} f_i\big(x_i\big) \\ \text{s.t.} \quad & a^\top x = b, \end{aligned}$$

где f_i — выпуклые дифференцируемые функции. Будем предполагать, что пересечение области определения целевой функции и ограничения не пусто и минимум достигается в конечной точке. Это говорит о том, что решение существует и единственно.

Решение. Выпишем лагранжиан:

$$\mathcal{L}(x, \nu) = \sum_{i=1}^{d} f_i\big(x_i\big) + \nu(a^\top x - b) = -b\nu + \sum_{i=1}^{d} \big(f_i\big(x_i\big) + \nu a_i x_i\big).$$

Он также является сепарабельным по компонентам вектора x. Тогда двойственная функция имеет вид

$$\begin{aligned} g(\nu) &= -b\nu + \inf_x \left(\sum_{i=1}^{d} \big(f_i\big(x_i\big) + \nu a_i x_i\big)\right) \\ &= -b\nu + \sum_{i=1}^{d} \inf_{x_i} \big(f_i\big(x_i\big) + \nu a_i x_i\big) \\ &= -b\nu - \sum_{i=1}^{d} f_i^*\big(-\nu a_i\big), \end{aligned}$$

где f_i^* — сопряженные по Фенхелю функции. Имеем двойственную задачу:

$$\max_{\nu} \quad \left[-b\nu - \sum_{i=1}^{d} f_i^*\big(-\nu a_i\big)\right].$$

Предположим, что ν^* — решение двойственной задачи. Для его поиска можно пользоваться уже известными методами, например, методом дихотомии или золотого сечения. В силу показанного в начале параграфа, минимум прямой задачи совпадает с минимумом $\mathcal{L}(x, \nu^*)$. Тогда, для поиска x^* можно взять градиент лагранжиана в ν^* по x и приравнять его к нулю, то есть решать уравнения:

$$\frac{\partial \mathcal{L}}{\partial x_i}\ (x^*, \nu^*) = f_i'(x_i^*) + \nu^* a_i = 0.$$

■

C9.4 Условия KKT для локальных экстремумов невыпуклых задач

Ранее мы требовали выпуклость целевой функции и ограничений, чтобы дать достаточные условия минимума оптимизационной задачи. Однако, на практике часто приходится иметь дело с не выпуклыми постановками. Как и в случае минимизации без ограничений, условия, затрагивающие только первые производные, не могут быть достаточными для локальных минимумов в общем случае. Нужно вводить дополнительные условия на вторые производные, чтобы сделать условия KKT достаточными для локального минимума. Для набора переменных $\big(x, \lambda, \nu\big)$ определим следующие множества из активных неравенств:

$$I^0(x) = \big\{\ i \ \big|\ f_i\big(x\big) = 0,\ \lambda_i = 0\ \big\},$$
$$I^+(x) = \big\{\ i \ \big|\ f_i\big(x\big) = 0,\ \lambda_i > 0\ \big\}.$$

Определение C9.2. Если для любого вектора $z \neq 0$, такого что:

$$z^\top \nabla_x f_i\big(x\big) = 0,\ i \in I^+(x),$$
$$z^\top \nabla_x f_i\big(x\big) \leq 0,\ i \in I^0(x),$$
$$z^\top \nabla_x h_j\big(x\big) = 0,\ j = \overline{1, m}$$

верно, что

$$z^\top \nabla^2_{xx} \mathcal{L}(x, \lambda, \nu) z > 0,$$

то говорят, что для набора переменных (x, λ, ν) выполнено *достаточное условие второго порядка (second-order sufficient condition или SOSC).*

Используя условие второго порядка, сформулируем теорему.

Теорема C9.3. Пусть функции f_0, f_i, h_j являются дважды непрерывно дифференцируемыми. Тогда, если для набора переменных $\big(x^*, \lambda^*, \nu^*\big)$ выполнены все условия KKT и SOSC, то x^* является точкой локального минимума прямой задачи (C9.1).

Применим теорему на конкретном примере.

Пример C9.12. Решите задачу оптимизации

$$\begin{aligned} \min_{x \in \mathbb{R}^d} \quad & -x \\ \text{s.t.} \quad & x^2 + y^2 \leq 1 \\ & (x-1)^3 - y \leq 0. \end{aligned}$$

Решение. Запишем лагранжиан задачи:

$$\mathcal{L}(x, y, \lambda_1, \lambda_2) = -x + \lambda_1\big(x^2 + y^2 - 1\big) + \lambda_2\big((x-1)^3 - y\big).$$

Запишем ограничения:

$$\begin{aligned} x^2 + y^2 &\le 1, \\ (x-1)^3 - y &\le 0. \end{aligned}$$

Неотрицательность:

$$\lambda_1, \lambda_2 \ge 0.$$

Дополняющая нежёсткость:

$$\begin{aligned} \lambda_1\big(x^2 + y^2 - 1\big) &= 0, \\ \lambda_2\big((x-1)^3 - y\big) &= 0. \end{aligned}$$

Стационарность:

$$\begin{aligned} \frac{\partial \mathcal{L}}{\partial x}\,(x, y, \lambda_1, \lambda_2) &= -1 + 2\lambda_1 x + 3\lambda_2 (x-1)^2 = 0, \\ \frac{\partial \mathcal{L}}{\partial y}\,(x, y, \lambda_1, \lambda_2) &= 2\lambda_1 y - \lambda_2 = 0. \end{aligned}$$

Переберём четыре условия дополняющей нежёсткости:

1. $\lambda_1 = 0$, $\lambda_2 = 0$. Из условий на градиент по x получаем противоречие $0 = -1$.

2. $\lambda_1 = 0$, $\lambda_2 > 0$. Из условий на градиент по y получаем $\lambda_2 = 0$ — противоречие.

3. $\lambda_1 > 0$, $\lambda_2 = 0$. По градиенту по y получаем $y = 0$ и $x = \pm 1$ из дополняющей нежесткости. Точка $x = -1$, $y = 0$ в условии на градиент по x даёт оценку $\lambda_1 = \frac{1}{2x} = -\frac{1}{2} < 0$ — противоречие. Точка $x = 1$, $y = 0$ в условии на градиент по x даёт оценку $\lambda_1 = \frac{1}{2x} = \frac{1}{2} > 0$. Все ограничения выполняются. Тогда подходит набор переменных $\big(x, y, \lambda_1, \lambda_2\big) = \big(1, 0, \frac{1}{2}, 0\big)$.

4. $\lambda_1 > 0$, $\lambda_2 > 0$. Получим, что $x^2 + y^2 = 1$ и $(x-1)^3 - y = 0$. Только точки $(1, 0)$ и $(0, -1)$ подходят под эти условия. Точка $(1, 0)$ и условие градиента по y требуют $\lambda_2 = 0$ — противоречие. Для точки $(0, -1)$ решаем систему из условий на градиент и получаем $\lambda_2 = \frac{1}{3}$ и $\lambda_1 = -\frac{1}{6} < 0$ — противоречие.

Получили единственный набор переменных, удовлетворяющий ККТ. Запишем для него множества активных неравенств

$$I^+(1, 0) = \{1\},\ I^0(1, 0) = \{2\},$$

градиенты ограничений

$$\nabla f_1\big(1, 0\big) = \begin{pmatrix} 2x \\ 2y \end{pmatrix}\Bigg|_{(1,0)} = \begin{pmatrix} 2 \\ 0 \end{pmatrix},\ \nabla f_2\big(1, 0\big) = \begin{pmatrix} 3(x-1)^2 \\ -1 \end{pmatrix}\Bigg|_{(1,0)} = \begin{pmatrix} 0 \\ -1 \end{pmatrix}$$

и гессиан функции Лагранжа:

$$\nabla^2 \mathcal{L}\Big(1, 0, \frac{1}{2}, 0\Big) = \begin{pmatrix} 2\lambda_1 + 6\lambda_2(x-1) & 0 \\ 0 & 2\lambda_1 \end{pmatrix}\Bigg|_{\left(1,0,\frac{1}{2},0\right)} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}.$$

Для множества векторов $z \neq 0 \in \mathbb{R}^2$, таких что

$$z^\top \nabla f_1\big(1,0\big) = 2z_1 = 0,\ z^\top \nabla f_2\big(1,0\big) = -z_2 \leq 0,$$

выполнено

$$z^\top \nabla^2 \mathcal{L}\Big(1, 0, \frac{1}{2}, 0\Big) z = z_1^2 + z_2^2 > 0.$$

Таким образом, для набора $\big(1, 0, \frac{1}{2}, 0\big)$ KKT и SOSC выполнены, и точка $(1, 0)$ является локальным минимумом. ■

C10 Классические виды оптимизационных задач

До этого мы изучали оптимизационные задачи в общем виде. Однако, среди них можно выделить классы, для которых существуют специализированные методы решения. В этой главе мы рассмотрим основные классы оптимизационных задач и покажем, как некоторые прикладные проблемы могут быть к ним сведены.

C10.1 Линейное программирование

Первый и наиболее простой класс задач — *линейное программирование* (LP). Оно заключается в минимизации линейной функции при линейных ограничениях и широко применяется в различных областях, таких как планирование производства, транспорт, финансы, логистика. Существует несколько основных форм записи задачи LP.

Определение C10.1. Пусть $A \in \mathbb{R}^{n\times d}$, $b \in \mathbb{R}^n$, $G \in \mathbb{R}^{m\times d}$, $h \in \mathbb{R}^m$. Говорят, что задача *линейного программирования* записана

1. в *общей форме*, если она имеет вид
$$\begin{aligned} \min_{x\in\mathbb{R}^d} \quad & c^\top x \\ \text{s.t.} \quad & Ax = b \\ & Gx \preceq h; \end{aligned}$$
2. в *стандартной форме*, если она имеет вид
$$\begin{aligned} \min_{x\in\mathbb{R}^d} \quad & c^\top x \\ \text{s.t.} \quad & Ax = b \\ & x \succeq 0; \end{aligned}$$
3. в *канонической форме*, если она имеет вид
$$\begin{aligned} \min_{x\in\mathbb{R}^d} \quad & c^\top x \\ \text{s.t.} \quad & Gx \preceq h \\ & x \succeq 0. \end{aligned}$$

Та или иная форма записи выбирается в зависимости от ограничений решаемой задачи. Тем не менее, оказывается, что все три вида задачи LP эквивалентны.

Утверждение C10.1. Оптимизационная задача линейного программирования в любом виде может быть эквивалентно переписана в любой другой форме.

Доказательство.

1. Покажем, что любая задача линейного программирования, записанная в общей форме, представима в стандартной форме. Заметим, что неравенство $Gx \preceq h$ эквивалентно системе неравенств
$$Gx + v = h,\ v \succeq 0$$

с новой переменной $v \in \mathbb{R}^m$. Более того, любая переменная $x \in \mathbb{R}^d$ может представлена в виде

$$x = x^+ - x^-,$$

где

$$x_i^+ = \begin{cases} x_i, & x_i \geq 0, \\ 0, & x_i < 0, \end{cases} \quad x_i^- = \begin{cases} 0, & x_i \geq 0, \\ -x_i, & x_i < 0. \end{cases}$$

Тогда задача при помощи введения новых переменных может быть переписана в виде

$$\begin{aligned} \min_{\substack{x^+ \in \mathbb{R}^d \\ x^- \in \mathbb{R}^d \\ v \in \mathbb{R}^m}} \quad & \begin{pmatrix} c^\top & -c^\top & \mathbf{0}^\top \end{pmatrix} \begin{pmatrix} x^+ \\ x^- \\ v \end{pmatrix} \\ \text{s.t.} \quad & \begin{pmatrix} A & -A & \mathbf{0}_{n \times m} \\ G & -G & I_m \end{pmatrix} \begin{pmatrix} x^+ \\ x^- \\ v \end{pmatrix} = \begin{pmatrix} b \\ h \end{pmatrix} \\ & \begin{pmatrix} x^+ \\ x^- \\ v \end{pmatrix} \succeq 0. \end{aligned}$$

2. Имея задачу в стандартной форме, несложно свести ее к задаче в общей форме. Действительно, ограничение $x \succeq 0$ может быть представлено в виде $-Ix \preceq 0$. Таким образом, получили задачу в общем виде.

3. Покажем, что задача в стандартной форме может быть переписана в канонической форме. Заметим, что ограничение $Ax = b$ эквивалентно паре ограничений

   $$Ax \preceq b, \quad -Ax \preceq -b.$$

   Тогда сформируем матрицу G и вектор h как

   $$G = \begin{pmatrix} A \\ -A \end{pmatrix}, \quad h = \begin{pmatrix} b \\ -b \end{pmatrix}.$$

   Таким образом, ограничение $Ax = b$ принимает вид

   $$Gx \preceq h.$$

4. Покажем теперь, что каноническая постановка задачи может быть сведена к стандартной форме. Снова введем вектор $v \succeq 0$ и перепишем $Gx \preceq h$ в виде

   $$Gx + v = h.$$

   Получим задачу в стандартной форме:

   $$\begin{aligned} \min_{\substack{x \in \mathbb{R}^d \\ v \in \mathbb{R}^m}} \quad & \begin{pmatrix} c^\top & \mathbf{0}^\top \end{pmatrix} \begin{pmatrix} x \\ v \end{pmatrix}, \\ \text{s.t.} \quad & \begin{pmatrix} G & I_m \end{pmatrix} \begin{pmatrix} x \\ v \end{pmatrix} = h \\ & \begin{pmatrix} x \\ v \end{pmatrix} \succeq 0. \end{aligned}$$

■

Рассмотрим в качестве примеров несколько задач, которые сводятся к LP.

Пример C10.1. Пусть поставлена задача дробно-линейного программирования:

$$\begin{aligned} \min_{x\in\mathbb{R}^d} \quad & \left[f(x) := \frac{c^\top x + d}{e^\top x + f}\right] \\ \text{s.t.} \quad & Ax = b, \\ & Gx \preceq h, \end{aligned}$$

где $\operatorname{dom} f = \left\{\, x \mid e^\top x + f > 0 \,\right\}$. В общем случае данная задача не является выпуклой. Например, если $c = \mathbf{0}$, $d = -1$, $e = \mathbf{1}$, $f = 0$. Покажем, что она сводится к задаче из класса LP. Перепишем целевую функцию в виде

$$f(x) = c^\top \left(\frac{x}{e^\top x + f}\right) + d\left(\frac{1}{e^\top x + f}\right).$$

Поскольку на эффективном множестве целевого функционала выполнено $e^\top x + f > 0$, можем разделить ограничения на этот фактор. Получим

$$A\left(\frac{x}{e^\top x + f}\right) = b\left(\frac{1}{e^\top x + f}\right), \quad G\left(\frac{x}{e^\top x + f}\right) \leq h\left(\frac{1}{e^\top x + f}\right).$$

Введем

$$y = \frac{x}{e^\top x + f} \in \mathbb{R}^d, \quad z = \frac{1}{e^\top x + f} \in \mathbb{R}_+$$

и заметим, что целевая функция и ограничения линейны по этому набору переменных. Для разрешимости этой системы по x при данных y, z необходимо, чтобы выполнялись условия

$$e^\top y + fz = 1, \; z \geq 0.$$

Покажем, что задача дробно-линейного программирования эквивалента задаче вида:

$$\begin{aligned} \min_{\substack{y\in\mathbb{R}^d \\ z\in\mathbb{R}}} \quad & c^\top y + dz \\ \text{s.t.} \quad & Ay - bz = \mathbf{0} \\ & Gy - hz \preceq 0 \\ & e^\top y + fz = 1 \\ & z \geq 0. \end{aligned}$$

Выше показано, что для любого допустимого x в исходной задаче можно построить y, z во второй задаче. Соответственно, решение новой задачи не больше чем решение исходной. С другой стороны, если $z > 0$, то x восстанавливается по формуле

$$x = \frac{y}{z}.$$

Отдельного рассмотрения требует случай $z = 0$. Пусть x_0 допустим в исходной задаче. Тогда $x = x_0 + ty$ тоже допустим и имеем

$$f\big(x_0 + ty\big) = \frac{c^\top (x_0 + ty) + d}{e^\top (x_0 + ty) + f} = \frac{tc^\top y + c^\top x_0 + d}{t + e^\top x_0 + f} \to c^\top y, \; t \to +\infty.$$

Таким образом, показали, что для любой допустимой пары y, z можно построить либо x из исходной, либо последовательность допустимых x_t, таких что $f\big(x_t\big) \to c^\top y + dz$. Соответственно, задачи равносильны.

Пример C10.2. Рассмотрим задачу динамического планирования активности. Есть d экономических секторов, T временных периода и n различных товаров. Каждый экономический сектор j в момент времени $t \in \overline{0,T}$ характеризуется своей «активностью» $x_j(t) \geq 0$. Каждый экономический сектор производит и потребляет товары пропорционально соответствующей активности. Известно, что j-й сектор производит количество $a_{ij}x_j$ товара i и потребляет $b_{ij}x_j$. Задача состоит в оптимальном планировании активности экономических секторов $x(t)$, $t = \overline{1,T}$ при заданном $x(0)$. Покажем, что она лежит в классе LP. Введём $T+1$ новую переменную, соответствующую избыточным товарам:

$$\begin{aligned} &s(t) = Ax(t) - Bx(t+1), \quad t = \overline{0,T-1}, \\ &s(T) = Ax(T). \end{aligned}$$

Тогда целевую функцию, которую мы хотим максимизировать, можно записать в виде

$$f\big(x\big) = \sum_{t=0}^{T} \gamma^t c^\top s(t),$$

где $c \in \mathbb{R}^n$ — вектор стоимости товаров и $0 < \gamma \leq 1$ — дисконтирующий фактор, уменьшающий значимость будущего. Итого, задача динамического планирования активности в виде задачи линейного программирования записывается как

$$\begin{aligned} \min_{\{x(t)\}_{t=1}^{T} \subseteq \mathbb{R}^d} \quad & \sum_{t=0}^{T} \gamma^t c^\top s(t) \\ \text{s.t.} \quad & s(t) = Ax(t) - Bx(t+1), \quad t = \overline{0,T-1} \\ & s(T) = Ax(T) \\ & s(t) \geq 0, \quad t = \overline{0,T} \\ & x(t) \geq 0, \quad t = \overline{1,T}. \end{aligned}$$

C10.2 Квадратичное программирование с квадратичными ограничениями

Класс LP включает в себя задачи с линейным целевым функционалом и линейными ограничениями. Естественный способ его расширить — ввести в рассмотрение выпуклые квадратичные функции.

Определение C10.2. Задача *квадратичного программирования с квадратичными ограничениями* (QCQP) записывается в виде:

$$\begin{aligned} \min_{x \in \mathbb{R}^d} \quad & \frac{1}{2} x^\top P_0 x + q_0^\top x + r_0 \\ \text{s.t.} \quad & \frac{1}{2} x^\top P_i x + q_i^\top x + r_i \leq 0, \quad i = \overline{1,m} \\ & Gx = h, \end{aligned}$$

где $P_i \in \mathbb{S}_+^d$, $q_i \in \mathbb{R}^d$, $P_i \in \mathbb{R}$, $G \in \mathbb{R}^{n \times d}$.

Задачи такого вида часто встречаются в оптимальном управлении.

Замечание C10.1. Понятно, что в случае $A = P_i = \mathbf{0}_{d \times d}$ квадратичная задача

вырождается в линейную. Таким образом, имеем строгое вложение $\mathrm{LP} \subset \mathrm{QCQP}$.

Пример C10.3. Минимизация квадратичной функции на единичном евклидовом шаре лежит в классе QCQP:

$$\begin{aligned} \min_{x \in \mathbb{R}^d} \quad & \frac{1}{2} x^\top A x + c^\top x \\ \text{s.t.} \quad & x^\top x \leq 1. \end{aligned}$$

Замечание C10.2. В случае, когда ограничения вида неравенства линейны, говорят о задаче *квадратичного программирования* (QP).

Пример C10.4. Задача наименьших квадратов представима в виде задачи квадратичного программирования:

$$\min_{x \in \mathbb{R}^d} f(x) = \frac{1}{2} \|Ax - b\|_2^2 = \frac{1}{2} x^\top (A^\top A) x + \left(-A^\top b\right)^\top x + \left(\frac{1}{2} b^\top b\right).$$

Пример C10.5. Рассмотрим задачу составления оптимального портфеля. Есть d активов, в которые можно инвестировать. Классической модели предполагают, что цена i-го актива — есть случайная величина с известным математическим ожиданием p_i и известной дисперсией σ_i. Показатель ковариации между активами σ_{ij} также считается известным. Задача состоит в подборе такого распределения бюджета между активами, которое минимизирует риски при условии, что ожидаемая прибыль превышает некоторое значение. Покажем, что эта задача лежит в классе QP. Роль переменных в рассматриваемой задаче играют доли актива i в портфеле. Обозначим их как x_i. Понятно, что x удовлетворяют ограничениям

$$\sum_{i=1}^{d} x_i = 1, \; x \succeq 0.$$

Условие на то, что ожидаемая прибыль не меньше некоторого числа $\alpha \in \mathbb{R}$, есть условие вида

$$\sum_{i=1}^{d} p_i \geq \alpha.$$

Наконец, задача минимизации рисков есть задача минимизации дисперсии. Тогда задача оптимизации портфеля с ковариационной матрицей S приобретает вид

$$\begin{aligned} \min_{x \in \mathbb{R}^d} \quad & x^\top S x \\ \text{s.t.} \quad & \sum_{i=1}^{d} x_i = 1 \\ & x \succeq 0 \\ & \sum_{i=1}^{d} p_i x_i \geq \alpha. \end{aligned}$$

Пример C10.6. Рассмотрим задачу восстановления сферы по зашумленным данным. Пусть дан набор n точек $x_i \in \mathbb{R}^d$ и известно, что это зашумлённые точки сферы радиуса r с центром в x_c. Формально, это означает

$$x_i = \overline{x}_i + v,$$

где $\overline{x}_i \in S_d^2(x_c, r)$, $v \sim \mathcal{N}(0, \varepsilon^2)$. Покажем, что эта задача лежит в классе QP. Задачу восстановления x_c и r по зашумлённым данным можно записать как задачу оптимизации вида

$$\min_{\substack{x_c \in \mathbb{R}^d \\ r \in \mathbb{R}_+}} \sum_{i=1}^{n} \left(\|x_i - x_c\|_2^2 - r^2 \right)^2.$$

Целевой функционал в таком виде не является выпуклым, однако задача может быть сведена к задаче выпуклой оптимизации. Заметим, что целевая функция представима в виде:

$$f(x_c, t) = \sum_{i=1}^{n} \left(\|x_i - x_c\|_2^2 - r^2 \right)^2 = \sum_{i=1}^{n} \left(\|x_i\|_2^2 - 2x_i^\top x_c + \|x_c\|_2^2 - r^2 \right)^2.$$

Введём новую переменную $t = \|x_c\|^2 - r^2$. Тогда исходная задача минимизации может быть переписана в виде задачи наименьших квадратов:

$$\min_{\substack{x_c \in \mathbb{R}^d \\ t \in \mathbb{R}}} \sum_{i=1}^{n} \left(\|x_i\|_2^2 - 2x_i^\top x_c + t \right)^2.$$

Множество решений данной задачи шире, однако, если (x_c, t) — решение новой задачи, то выполнено

$$\nabla_t f(x_c, t) = 0 \implies \sum_{i=1}^{n} \left(\|x_i\|_2^2 - 2x_i^\top x_c + t \right) = 0,$$

откуда следует, что

$$\|x_c\|_2^2 - t = \frac{1}{n} \sum_{i=1}^{n} \|x_i - x_c\|_2^2 \geq 0.$$

Таким образом, изначальную невыпуклую задачу аппроксимации данных сферой удалось свести к задаче из класса QP.

C10.3 Коническое программирование второго порядка

Следующим шагом к расширению класса квадратичных выпуклых задач является обобщение на выпуклые ограничения более общего вида, в частности, на выпуклые конусы.

Определение C10.3. Задача *конического программирования второго порядка*

(SOCP) записывается в виде:

$$\begin{aligned} \min_{x\in\mathbb{R}^d} \quad & c^\top x \\ \text{s.t.} \quad & \left\|P_i x - q_i\right\|_2 \le e_i^\top x + f_i, \quad i = \overline{1,m} \\ & Gx = h, \end{aligned}$$

где $P_i \in \mathbb{R}^{k_i\times d}$, $q_i \in \mathbb{R}^{k_i}$, $e_i \in \mathbb{R}^d$, $f_i \in \mathbb{R}$, $G \in \mathbb{R}^{n\times d}$.

SOCP находит применение в проектировании фильтров, расчете веса антенных решеток и оптимизации силы захвата в робототехнике.

Замечание C10.3. Ограничение вида неравенства означает, что пары

$$\big(P_i x - q_i, e_i^\top x + f_i\big)$$

лежат в конусах второго порядка

$$K_i = \left\{ (y,t) \in \mathbb{R}^{k_i} \times \mathbb{R}_+ \;\middle|\; \|y\| \le t \right\}$$

соответствующей размерности.

Вложение QCQP $\subset$ SOCP не так очевидно, как LP $\subset$ QCQP, и требует формального доказательства.

Утверждение C10.2. Выполнено строгое вложение QCQP $\subset$ SOCP.

Доказательство. Заметим, что целевая функция может быть опущена в ограничение с помощью переформулировки оптимизационной задачи через надграфик. Действительно, рассмотрим минимизацию выпуклой функции

$$\min_{x\in\mathbb{R}^d} f\big(x\big).$$

Эта задача имеет такое же решение, как и

$$\begin{aligned} \min_{\substack{x\in\mathbb{R}^d \\ t\in\mathbb{R}}} \quad & t \\ \text{s.t.} \quad & f\big(x\big) \le t. \end{aligned}$$

Таким образом, достаточно рассмотреть квадратичное ограничение вида

$$\frac{1}{2} x^\top P x + q^\top x + r \le 0$$

и показать, что оно эквивалентно коническому. Так как $P \in \mathbb{S}^d_+$, то из неё можно извлечь корень, то есть существует $L \in \mathbb{R}^{d\times d}$ такой, что $P = 2L^\top L$. Тогда имеем

$$\frac{1}{2} x^\top P x + q^\top x + r = \left\|Lx\right\|_2^2 + q^\top x + r.$$

Рассмотрим линейную часть этого ограничения. Обозначая $t = q^\top x + r$ и используя тождество

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

получим

$$\left\|Lx\right\|_2^2 + \left(\frac{q^\top x + r + 1}{2}\right)^2 \leq \left(\frac{q^\top x + r - 1}{2}\right)^2.$$

Извлекая квадратный корень, получим:

$$\sqrt{\left\|Lx\right\|_2^2 + \left(\frac{q^\top x + r + 1}{2}\right)^2} \leq \frac{1 - q^\top x - r}{2}.$$

Рассмотрим преобразования

$$g(x) = \left(Lx, \frac{q^\top x + r + 1}{2}\right), \quad h(x) = \frac{1 - q^\top x - r}{2}.$$

Это аффинные преобразования. Таким образом, каждое выпуклое квадратичное ограничение сводится к коническому. Следовательно, QCQP $\subset$ SOCP. ■

Доказательство вложения QCQP $\subset$ SOCP использует переформулировку через надграфик. Этот же подход может быть использован и для приведения задач к виду SOCP. Рассмотрим простейшие примеры.

Пример C10.7. Рассмотрим задачу минимизации суммы норм аффинных преобразований:

$$\min_{x \in \mathbb{R}^d} \sum_{i=1}^m \left\|A_i x - b_i\right\|_2,$$

где $A_i \in \mathbb{R}^{n \times d}$, $b_i \in \mathbb{R}^n$. Покажем, что она лежит в классе SOCP. Точки (x, t_i) надграфика i-го аффинного преобразования задаются неравенством

$$\left\|A_i x - b_i\right\|_2 \leq t_i.$$

Тогда исходная задача может быть переписана в виде

$$\begin{aligned} \min_{\substack{x \in \mathbb{R}^d, \\ t \in \mathbb{R}^m}} \quad & \sum_{i=1}^m t_i \\ \text{s.t.} \quad & \left\|A_i x - b_i\right\|_2 \leq t_i, \quad i = \overline{1, m}. \end{aligned}$$

Пример C10.8. Рассмотрим задачу оптимизации

$$\min_{x \in \mathbb{R}^d} \max_{i = \overline{1,m}} \left\|A_i x - b_i\right\|_2,$$

где $A_i \in \mathbb{R}^{n \times d}$, $b_i \in \mathbb{R}^n$. Аналогично предыдущему примеру она может быть переписана в виде

$$\begin{aligned} \min_{\substack{x \in \mathbb{R}^d, \\ t \in \mathbb{R}^m}} \quad & t \\ \text{s.t.} \quad & \left\|A_i x - b_i\right\|_2 \leq t, \quad i = \overline{1, m}. \end{aligned}$$

Замечание C10.4. Примеры C10.7, C10.8 демонстрируют, что негладкие компоненты целевой функции могут быть вынесены в ограничения, что делает задачу более удобной для решения, например, барьерными методами.

В завершение обсуждения класса SOCP, перейдем к обсуждению более сложных практических примеров.

Пример С10.9. Пусть выполнено ограничение вида

$$c^\top x \le \alpha,\ \forall c \in \left\{ c \in \mathbb{R}^d \;\middle|\; \|c - c_0\|_2 \le \varepsilon,\ Ac \le b \right\},$$

где $c_0 \in \mathbb{R}^d$, $A \in \mathbb{R}^{n\times d}$, $b \in \mathbb{R}^n$. Ограничения такого рода крайне тяжело проверять, поэтому их нужно каким-либо образом упростить. Покажем, что эти ограничения сводятся к ограничениям из класса SOCP. Найдем наибольшее значение $c^\top x$:

$$\begin{aligned} \max_{\|c-c_0\|_2\le\varepsilon} \quad & c^\top x \\ \text{s.t.} \quad & Ac \le b. \end{aligned}$$

В силу того, что $\operatorname{int}\mathcal{C} \neq \varnothing$, выполняется сильная двойственность. То есть оптимальное значение прямой задачи совпадает с оптимальным значением двойственной. Выведем ее, пользуясь аппаратом двойственных функций Лагранжа. Запишем лагранжиан задачи:

$$\mathcal{L}(c, \lambda) = -c^\top x + \lambda^\top (Ac - b).$$

Далее, вычислим двойственную функцию:

$$\begin{aligned} g(\lambda) &= \inf_{\|c-c_0\|_2\le\varepsilon} \mathcal{L}(c,\lambda) \\ &= -c_0^\top x + \lambda^\top (Ac_0 - b) + \inf_{\|\Delta\|_2\le\varepsilon} \left(\lambda^\top A - x^\top\right)\Delta \\ &= -c_0^\top x + \lambda^\top (Ac_0 - b) - \varepsilon \left\|A^\top\lambda - x\right\|_2. \end{aligned}$$

Тогда двойственная задача имеет вид

$$\begin{aligned} \min_{\lambda\in\mathbb{R}^n} \quad & c_0^\top x - \lambda^\top (Ac_0 - b) + \varepsilon \left\|A^\top\lambda - x\right\|_2 \\ \text{s.t.} \quad & \lambda \succeq 0. \end{aligned}$$

Таким образом, исходный набор ограничений $c^\top x \le \alpha$ может быть переписан в виде конического ограничения:

$$\begin{aligned} &\left\|A^\top\lambda - x\right\|_2 \le \frac{1}{\varepsilon}\left(Ac_0 - b\right)^\top \lambda + \frac{1}{\varepsilon}\left(\alpha - c_0^\top x\right), \\ &\lambda \succeq 0. \end{aligned}$$

Если требуется учитывать несколько значений параметров задачи одновременно, то говорят, что имеют дело с задачей *робастного программмирования.*

Пример С10.10. Рассмотрим робастную задачу наименьших квадратов:

$$\min_{x\in\mathbb{R}^d} \left[f(x) := \sup_{(A,b)\in\mathcal{A}} \left\|Ax - b\right\|_2 \right],$$

где

$$\mathcal{A}(A_0, b_0) = \left\{ (A,b) \in \mathbb{R}^{n\times d} \times \mathbb{R}^n \;\middle|\; \left\|\begin{pmatrix} A - A_0 & b - b_0 \end{pmatrix}\right\|_F \le \rho \right\},\ A_0 \in \mathbb{R}^{n\times d},\ b_0 \in \mathbb{R}^n.$$

Покажем, что эта задача лежит в классе SOCP. Введём обозначение $\Delta A = A - A_0$, $\Delta B = B - B_0$. Пользуясь неравенством треугольника и неравенством Коши-Буняковского-Шварца 0.3, получим следующую цепочку неравенств:

$$\begin{aligned} f(x) &\leq \|A_0x - b_0\|_2 + \sup_{\|(\Delta A \quad \Delta b)\|_F \leq \rho} \|\Delta A x - \Delta b\|_2 \\ &\leq \|A_0x - b_0\|_2 + \sup_{\|(\Delta A \quad \Delta b)\|_F \leq \rho} \|(\Delta A \quad \Delta b)\|_F \left\| \begin{pmatrix} x \\ 1 \end{pmatrix} \right\|_2 \\ &= \|A_0x - b_0\|_2 + \rho \left\| \begin{pmatrix} x \\ 1 \end{pmatrix} \right\|_2. \end{aligned}$$

Покажем, что все неравенства выше реализуются как равенства.

1. Пусть $A_0x - b_0 = 0$. Тогда первое неравенство реализуется как равенство. Выберем дополнительно

$$(\Delta A \quad \Delta b) = \rho \frac{\mathbf{1} x_1^\top}{\sqrt{n}\|x_1\|_2}, \; x_1 = \begin{pmatrix} x \\ 1 \end{pmatrix}.$$

   Тогда второе неравенство также обращается в равенство.

2. Теперь пусть $A_0x - b_0 \neq 0$. Тогда возьмём

$$(\Delta A \quad \Delta b) = \rho \frac{(A_0x - b_0)x_1^\top}{\|A_0x - b_0\|_2 \|x_1\|_2}, \; x_1 = \begin{pmatrix} x \\ -1 \end{pmatrix}.$$

   Для такой матрицы оба неравенства реализуются как равенства.

Таким образом, исходная задача может быть переписана в виде

$$\min_{x \in \mathbb{R}^d} \|A_0x - b_0\|_2 + \rho \left\| \begin{pmatrix} x \\ 1 \end{pmatrix} \right\|_2,$$

которая сводится SOCP за счет переформулировки через надграфик.

C10.4 Полуопределенное программирование

Класс SOCP можем расширить за счет рассмотрения более общего случая конических ограничений.

Определение C10.4. Задача *полуопределенного программирования* (SDP) записывается следующим образом:

$$\begin{aligned} \min_{x \in \mathbb{R}^d} \quad & c^\top x \\ \text{s.t.} \quad & P_0 + \sum_{i=1}^d P_i x_i \succeq 0 \\ & Gx = h, \end{aligned}$$

где $G \in \mathbb{R}^{n \times d}$, $P_i \in \mathbb{S}^d$.

Задачи этого класса используются в областях исследования операций и комбинаторной

оптимизации, машинном обучении, теории автоматического управления, оптимизации сигналов и теории игр. Преобразование ограничений к виду, соответствующему SDP, как правило опирается на лемму Шура.

Утверждение C10.3. Рассмотрим блочную матрицу

$$M = \begin{pmatrix} A & B^\top \\ B & C \end{pmatrix},$$

где $A \in \mathbb{S}^n_{++}$, $B \in \mathbb{R}^{n\times d}$, $C \in \mathbb{S}^d$. Тогда M положительно полуопределена тогда и только тогда, когда

$$BA^{-1}B^\top \preceq C.$$

Доказательство. Положительная полуопределенность матрицы M эквивалентна выполнению неравенства

$$x^\top Ax + 2y^\top Bx + y^\top Cy \geq 0$$

для любых $x \in \mathbb{R}^d$, $y \in \mathbb{R}^n$. Эквивалентно перепишем это в виде

$$\inf_{x\in\mathbb{R}^d} \left[x^\top Ax + 2y^\top Bx + y^\top Cy\right] \geq 0$$

для любого $y \in \mathbb{R}^n$. Функция

$$g(x) = x^\top Ax + 2y^\top Bx + y^\top Cy$$

является выпуклой по x с единственным минимумом в точке

$$x^* = -A^{-1}B^\top y.$$

Таким образом, имеем

$$\begin{aligned} g(x^*) &= y^\top BA^{-1}B^\top y - 2y^\top BA^{-1}B^\top y + y^\top Cy \\ &= y^\top \left(C - BA^{-1}B^\top\right) y \geq 0, \end{aligned}$$

что в свою очередь эквивалентно

$$BA^{-1}B^\top \preceq C.$$

■

Пользуясь Утверждением C10.3, докажем следующий факт.

Утверждение C10.4. Выполнено строгое вложение SOCP $\subset$ SDP.

Доказательство. Рассмотрим коническое ограничение:

$$\left\|Px - q\right\| \leq e^\top x + f.$$

Оно может быть эквивалентно переписано в виде системы неравенств

$$\begin{cases} \left\|Px - q\right\|^2 \leq \left(e^\top x + f\right)^2, \\ e^\top x + f \geq 0, \end{cases},$$

что в свою очередь эквивалентно положительной полуопределенности матрицы

$$M = \begin{pmatrix} e^\top x + f & (Px - q)^\top \\ Px - q & e^\top x + f \end{pmatrix}.$$

Таким образом, коническое ограничение в задаче из класса SOCP сводится к виду SDP. ■

Пример С10.11. Рассмотрим задачу оптимального дизайна эксперимента. Пусть даны измерения вида

$$y_i = a^\top x_i + \varepsilon,$$

где $\varepsilon \sim \mathcal{N}(0, \sigma^2)$ — шум, $y_i \in \mathbb{R}$ — наблюдаемая величина, $a \in \mathbb{R}^d$ — параметры модели, которые необходимо восстановить, $x_i \in \mathbb{R}^d$ — параметры измерений. Векторы x_i являются гиперпараметрами и могут быть выбраны из конечного набора $\{v_i\}_{i=1}^n$. Задача состоит в выборе K векторов x_i таким образом, чтобы дисперсия решения a была минимальна. Покажем, что эту задачу можно рассматривать как SDP. Пусть d_i — число векторов v_i в наборе. Запишем логарифм правдоподобия:

$$\ln L \propto -\sum_{i=1}^{K} \frac{\left(y_i - a^\top x_i\right)^2}{2\sigma^2}.$$

Найдем информацию Фишера:

$$I(a) = -\nabla^2 \ln L = \frac{1}{\sigma^2} \sum_{i=1}^{K} x_i x_i^\top = \frac{1}{\sigma^2} \sum_{i=1}^{n} d_i v_i v_i^\top.$$

Воспользуемся неравенством Рао-Крамера для любого $u \in \mathbb{R}^d$:

$$u^\top S u \geq u^\top I^{-1}(a) u \geq \frac{1}{\lambda_{\min}(I(a))},$$

где S — ковариационная матрица. Тогда рассматриваемая задача может быть формализована как

$$\begin{aligned} \max_{d \in \mathbb{Z}_+^n} \quad & \lambda_{\min}\left(\sum_{i=1}^{n} d_i v_i v_i^\top\right) \\ \text{s.t.} \quad & \sum_{i=1}^{n} d_i = K. \end{aligned}$$

Рассмотрим ее релаксацию:

$$\begin{aligned} \max_{\mu \in \mathbb{R}^n} \quad & \lambda_{\min}\left(\sum_{i=1}^{n} \mu_i v_i v_i^\top\right) \\ \text{s.t.} \quad & \mathbf{1}^\top \mu = 1 \\ & \mu \geq 0, \end{aligned}$$

где μ_i — доля векторов v_i во всем наборе. Отметим, что точность такой релаксации растёт с числом K в исходной задаче. Перепишем в виде SDP:

$$\begin{aligned} \max_{\substack{\mu\in\mathbb{R}^n \\ t\in\mathbb{R}}} \quad & t \\ \text{s.t.} \quad & \sum_{i=1}^{n} \mu_i v_i v_i^\top \succeq tI_n \\ \text{s.t.} \quad & \mathbf{1}^\top \mu = 1 \\ & \mu \geq 0. \end{aligned}$$

Пример C10.12. Рассмотрим задачу минимизации спектрального радиуса матрицы

$$\min_{x\in\mathbb{R}^d} \big\|A(x)\big\|_2,$$

где

$$A(x) := A_0 + \sum_{i=1}^{n} A_i x_i,\ A_i \in \mathbb{R}^{n\times d}.$$

Покажем, что она лежит в классе SDP. Целевая функция негладкая. Уберем её в ограничения через надграфик и перепишем в эквивалентном виде:

$$\begin{aligned} \min_{\substack{x\in\mathbb{R}^n \\ t\in\mathbb{R}}} \quad & t \\ \text{s.t.} \quad & \lambda_{\max}\big(A(x)^\top A(x)\big) \leq t^2 \\ & t \geq 0. \end{aligned}$$

Заметим, что ограничение, содержащее максимальное собственное число, эквивалентно ограничению

$$A(x)^\top A(x) \preceq t^2 I_d.$$

Поскольку $t \geq 0$, то

$$A(x)^\top A(x) \preceq t^2 I_d \iff A(x)^\top A(x) \preceq (t+\varepsilon)^2 I_d,\ \forall \varepsilon > 0.$$

По Утверждению C10.3 это эквивалентно

$$\begin{pmatrix} (t+\varepsilon)I_n & A(x) \\ A(x)^\top & (t+\varepsilon)I_d \end{pmatrix} = \begin{pmatrix} tI_n & A(x) \\ A(x)^\top & tI_d \end{pmatrix} + \varepsilon I_{d+n} \succeq 0.$$

Но в силу произвольности выбора $\varepsilon > 0$, имеем:

$$\begin{pmatrix} tI_n & A(x) \\ A(x)^\top & tI_d \end{pmatrix} \succeq 0.$$

Таким образом, исходная негладкая задача эквивалентна следующей задаче SDP:

$$\begin{aligned} \min_{\substack{x\in\mathbb{R}^d \\ t\in\mathbb{R}}} \quad & t \\ \text{s.t.} \quad & \begin{pmatrix} tI_n & A(x) \\ A(x)^\top & tI_d \end{pmatrix} \succeq 0. \end{aligned}$$

Замечание C10.5. Задачу SDP иногда формулируют для матриц в следующем виде:

$$\begin{aligned} \min_{X \in \mathbb{S}^d} \quad & \operatorname{Tr}\left(A_0^\top X\right) \\ \text{s.t.} \quad & \operatorname{Tr}\left(A_i^\top X\right) = b_i, \quad i = \overline{1,m} \\ & X \succeq 0, \end{aligned}$$

где $A_i \in \mathbb{R}^{d \times d}$, $b_i \in \mathbb{R}$.

Пример C10.13. Рассмотрим задачу вида

$$\begin{aligned} \min_{x \in \mathbb{R}^d} \quad & \frac{1}{2} x^\top A_0 x + b_0^\top x + c_0 \\ \text{s.t.} \quad & \frac{1}{2} x^\top A_i x + b_i^\top x + c_i \leq 0, \quad i = \overline{1,m}. \end{aligned}$$

где $A_i \in \mathbb{S}^d$, $b_i \in \mathbb{R}^d$, $c_i \in \mathbb{R}$, то есть квадратичные формы не обязаны быть неотрицательно определенными. Покажем, что эту задачу можно рассматривать как SDP. Введём новую переменную $X = xx^\top \in$ §7s_+^d и перепишем задачу в эквивалентном виде:

$$\begin{aligned} \min_{\substack{x \in \mathbb{R}^d \\ X \in \mathbb{S}^d}} \quad & \frac{1}{2} \operatorname{Tr}(A_0 X) + b_0^\top x + c_0 \\ \text{s.t.} \quad & \frac{1}{2} \operatorname{Tr}(A_i X) + b_i^\top x + c_i \leq 0, \quad i = \overline{1,m} \\ & X = xx^\top. \end{aligned}$$

Заметим, что теперь целевая функции и все ограничения-неравенства являются линейными, а следовательно выпуклыми. В таком виде вся сложность исходной задачи ушла в последнее ограничение $X = xx^\top$. Ослабим его до неравенства $X \succeq xx^\top$, которое можно переписать при помощи Утверждения C10.3:

$$\begin{aligned} \min_{\substack{x \in \mathbb{R}^d \\ X \in \mathbb{S}^d}} \quad & \frac{1}{2} \operatorname{Tr}(A_0 X) + b_0^\top x + c_0 \\ \text{s.t.} \quad & \frac{1}{2} \operatorname{Tr}(A_i X) + b_i^\top x + c_i \leq 0, \quad i = \overline{1,m} \\ & \begin{pmatrix} 1 & x^\top \\ x & X \end{pmatrix} \succeq 0. \end{aligned}$$

Отметим, что если в решении такой релаксации $\operatorname{rank} X = 1$, то x — решение исходной задачи.

C10.5 Коническое программирование

Естественное обобщение рассмотренных выше классов предоставляет *коническое программирование* (CP).

Определение C10.5. Задача *конического программирования* (CP) записыва-

ется в виде:

$$\begin{aligned} \min_{x \in K} \quad & x^\top c \\ \text{s.t.} \quad & Ax = b, \end{aligned}$$

где K — выпуклый конус в $\mathbb{R}^d$, $A \in \mathbb{R}^{n \times d}$.

В этом разделе мы рассмотрим стандартные классы конусов, под которые заточено большинство солверов.

Определение C10.6. *Экспоненциальный конус* в $\mathbb{R}^d$ определяется как

$$K_{\exp} = \operatorname{cl}\left\{ (x, y, z) \in \mathbb{R}^3 \;\middle|\; x \geq y \exp\left(\frac{z}{y}\right),\ y \geq 0 \right\}.$$

Пример C10.14. Рассмотрим функцию Ламберта. Это вогнутая функция, неявная заданная на $\mathbb{R}_+$ уравнением

$$W(x) e^{W(x)} = x.$$

Встречается в различных задачах, включающих себя уравнения с элементом ye^y. Покажем, что выпуклое ограничение

$$W(x) \geq t > 0$$

переписывается, как ограничение на экспоненциальном конусе. Заметим, что $W(x) \geq t$ тогда и только тогда, когда

$$x = W(x) e^{W(x)} \geq t e^t$$

в силу строгого возрастания функции ye^y. Таким образом, исходное неравенство может быть переписано в виде

$$x \geq t e^t = t e^{\frac{t^2}{t}}.$$

Второй практически полезный конус — степенной.

Определение C10.7. Степенной конус в $\mathbb{R}^d$ определяется как

$$\mathcal{P}^{\alpha, 1-\alpha} = \left\{ x \in \mathbb{R}^d \;\middle|\; x_1^\alpha x_2^{1-\alpha} \geq \sqrt{\sum_{i=1}^d x_i^2},\ x_1 \geq 0,\ x_2 \geq 0 \right\},$$

где $\alpha \in [0, 1]$.

Пример C10.15. Рассмотрим ограничение вида $\left\|x\right\|_p \leq t$. Оно выпукло при $p \geq 1$. Боле того, при $p = 1$ и $p = \infty$ оно сводится к LP, а при $p = 2$ — к SOCP. Рассмотрим остальные случаи. Заметим, что рассматриваемое ограничение эквивалентно

$$t^p \geq \sum_{i=1}^d |x_i|^p \implies t \geq \sum_{i=1}^d \frac{|x_i|^p}{t^{p-1}}.$$

Ограничим каждое слагаемое новой переменной r_i и получим:

$$t \geq \sum_{i=1}^{d} r_i, \;\; r_i t^{p-1} \geq |x_i|^p.$$

Пользуясь определением степенного конуса, получим:

$$(r_i, t, x_i) \in \mathcal{P}^{\frac{1}{p}, \frac{p-1}{p}}.$$

Л1 Основные понятия

Л1.1 Задача оптимизации

Задачу оптимизации, которая будет рассматриваться в рамках данного пособия, можно сформулировать следующим образом:

$$\begin{aligned} \min_{x} \quad & f(x) \\ \text{s.t.} \quad & x \in \mathcal{X} \subseteq \mathbb{R}^d \\ & g_i(x) = 0, \quad i = \overline{1,l} \\ & h_j(x) \leq 0, \quad j = \overline{1,m}. \end{aligned} \tag{Л1.1}$$

Получается, что, варьируя вектор x, нам необходимо найти значение f^* — минимум функции f такой, что («s.t.» — «subject to») дополнительно выполнен ряд условий, ограничений. Взглянув на задачу (Л1.1), можно заметить, что с формальной точки зрения нам достаточно найти f^*, что и будет являться решением. Но наряду с f^* часто необходим и вектор x^*, на котором это значение достигается $f^* = f(x^*)$. Получается, что нам скорее интересна не постановка (Л1.1), а постановка вида argmin по поиску x^*. Уже более формально мы определим её в Параграфе Л2. Понятно, что в общем случае уникальность, да и вообще существование x^* и f^* никто не гарантирует (рассмотрите, например, задачу $\min_{x \in \mathbb{R}} x$). Отметим, что задача (Л1.1) имеет общий вид, и мы часто будем ее упрощать. Например, можно рассматривать задачу безусловной оптимизации, т.е. убрать все ограничения вида равенств и неравенств и положить $\mathcal{X} = \mathbb{R}^d$.
Более подробно остановимся на объектах из формулировки (Л1.1):

- $f : \mathcal{X} \to \mathbb{R}$ — некоторая функция, заданная на множестве $\mathcal{X}$. Данная функция является целевой для задачи оптимизации (Л1.1). Мы будем накладывать дополнительные свойства на f: липшицевость, гладкость, выпуклость. Связано это с тем, что уже в рамках этого параграфа станет понятно, что без предположений на f не получится построить оптимистичную теорию поиска решения (Л1.1).

- $\mathcal{X} \subseteq \mathbb{R}^d$ — подмножество d-мерного пространства. В общем случае $\mathcal{X}$ может быть любым множеством, но часто мы будем предполагать, что $\mathcal{X}$ является «простым» (суть «простоты» в данном случае неоднозначна и будет раскрыта далее, в момент изучения методов оптимизации на «простых» множествах). Например, в качестве $\mathcal{X}$ может выступать шар радиуса R с центром в точке a в p-норме

  $$\overline{B}_d^p(R, a) = \left\{ x \in \mathbb{R}^d \;\middle|\; \|x - a\|_p \leq R \right\},$$

  вероятностный симплекс

  $$\triangle_{d-1} = \left\{ x \in \mathbb{R}^d \;\middle|\; x_i \geq 0,\ i = \overline{1,d},\ \sum_{i=1}^{d} x_i = 1 \right\}$$

  или положительный ортант

  $$\perp_d = \left\{ x \in \mathbb{R}^d \;\middle|\; x_i \geq 0 \right\}.$$

- Также в задаче (Л1.1) присутствуют функциональные ограничения вида равенств и неравенств. $g_i : \mathcal{X} \to \mathbb{R}$, $i = \overline{1,l}$ и $h_j : \mathcal{X} \to \mathbb{R}$, $j = \overline{1,m}$ — функции, задающие ограничения. Мы также будем в дальнейшем предполагать, что данные функции являются достаточно «хорошими». Отметим, что формально все функциональные ограничения можно было просто спрятать в $\mathcal{X}$ (часто возможно и обратно — $\mathcal{X}$ записывается в виде набора функциональных ограничений), главной проблемой при таком действии может стать потеря «простоты» $\mathcal{X}$.

Л1.2 Примеры задач оптимизации

Л1.2.1 Оптимизация инвестиционного портфеля

Предположим на рынке имеется d акций. Пусть мы владеем информацией об изменении цен на акции за n лет: p_i^j — цена i-ого актива в j год. Тогда рассмотрим следующую постановку — нужно распределить имеющиеся средства между активами таким образом, чтобы минимизировать итоговые инвестиционные риски.
Для начала мы можем рассчитать математическое ожидание μ_i доходности i-ой акции за n лет, применив формулу с учетом сложного процента:

$$\big(1+\mu_i\big)^n = \frac{p_i^n}{p_i^0} \implies \mu_i = \left(\frac{p_i^n}{p_i^0}\right)^{\frac{1}{n}} - 1.$$

Выпишем также выборочную ковариационную матрицу этих активов S, т.е. информацию о том, на сколько сильно варьируются цены на акции, и как эти цены связаны:

$$S = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \dots & \sigma_{1d} \\ \sigma_{21} & \sigma_{22} & \dots & \sigma_{2d} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{d1} & \sigma_{d1} & \dots & \sigma_{dd} \end{bmatrix}$$

где σ_{ii} — выборочная дисперсия i-го актива, а σ_{ij} — выборочная ковариация i-го и j-го активов, которые считаются по формуле:

$$\sigma_{ij} = \frac{\sum\limits_{k=1}^{n} (\mu_i^k - \mu_i)(\mu_j^k - \mu_j)}{n},$$

где $\mu_i^k = \frac{p_i^k}{p_i^{k-1}} - 1$ — доходность за k-ый год.
Наконец, хотим понять, какую долю портфеля стоит выделить на конкретную акцию $\omega = (\omega_1, \dots, \omega_d)^\top$, где $\sum\limits_{i=1}^{d} \omega_i = 1$ и $\omega_i \geq 0$, $i = \overline{1,d}$, чтобы при фиксированной доходности $\big\langle \mu, \omega \big\rangle = r$ иметь наименьшее отклонение. Тогда можно определить целевую функцию риска как

$$f\big(\omega\big) = \big\langle \omega, S\omega \big\rangle.$$

Итоговая задача оптимизации в таком случае принимает вид:

$$\begin{aligned} \min_{\omega} \quad & \big\langle \omega, S\omega \big\rangle \\ \text{s.t.} \quad & \big\langle \mu, \omega \big\rangle - r = 0 \\ & \sum_{i=1}^{d} \omega_i - 1 = 0 \\ & -\omega_i \leq 0, \quad i = \overline{1,d}. \end{aligned}$$

Л1.2.2 Оптимизация ценообразования

Предположим, что мы владеем небольшим бизнесом. В нашем магазине имеется d товаров. Для каждого товара i, мы знаем его себестоимость C_i, спрос Q_i^0 по начальной цене P_i^0 и функцию спроса $Q_i(P_i)$, где $x_i = \frac{P_i}{P_i^0}$ — отношение новой цены P_i к старой. В наших интересах — максимизировать оборот, не потеряв при этом текущую прибыль.

Для этого, определим целевую функцию как суммарный оборот нашего магазина:

$$f\big(x\big) = \sum_{i=1}^{d} P_i Q_i(P_i).$$

Поскольку сильное изменение цен может привести к неожиданным последствиям, например, потере постоянных клиентов, дополнительно потребуем, чтобы изменения цен на товары, которые мы внесли, лежали в допустимых интервалах: $P_i \in [P_i^l, P_i^r]$, $i = \overline{1,d}$.
Также нужно потребовать ограничения на прибыль:

$$\sum_{i=1}^{d} \big(P_i - C_i\big) Q_i(P_i) \geq \sum_{i=1}^{d} (P_i^0 - C_i) Q_i^0.$$

Наконец, можем сформулировать оптимизационную задачу:

$$\begin{aligned} \min_{P} \quad & -\sum_{i=1}^{d} P_i Q_i(P_i) \\ \text{s.t.} \quad & \sum_{i=1}^{d} (P_i - C_i) Q_i(P_i) \geq \sum_{i=1}^{d} (P_i^0 - C_i) Q_i^0 \\ & P_i \in [P_i^l, P_i^r], \quad i = \overline{1,d}. \end{aligned}$$

Л1.2.3 Оптимизация в статистике

Пусть есть параметрическое семейство вероятностных распределений $\{P(X|\theta)\}_{\theta \in \Theta}$, и дана независимая выборка $X = (X_1, \ldots, X_n) \sim P(X|\theta)$ для некоторого $\theta \in \Theta$. Наша задача — оценить параметр θ.
Предположим, что совместное распределение этой выборки задаётся функцией $p(x|\theta)$. Тогда для фиксированной реализации выборки $X = x$ функция правдоподобия:

$$L(\theta|x) = p(x|\theta).$$

Мы хотим максимизировать правдоподобие по θ, так как, чем оно больше, тем вероятнее, что наблюдаемые значения пришли из распределения P_θ.
В качестве целевой функции просто возьмем правдоподобие, а так как элементы выборки независимы, то его можно переписать:

$$L(\theta|x) = \prod_{i=1}^{n} p(x_i|\theta).$$

Зачастую удобно использовать логарифм правдоподобия:

$$\log L(\theta|x) = \sum_{i=1}^{n} \log p(x_i|\theta).$$

Итоговая задача оптимизации:

$$\min_{\theta \in \Theta} -\sum_{i=1}^{n} \log p(x_i|\theta).$$

Л1.2.4 Оптимизация обработки сигналов

Рассмотрим задачу идентификации неисправностей, произошедших в системе, на основе показаний датчиков. Пусть у нас есть d различных типов неисправностей, причем каждая из них происходит независимо с вероятностью p, также есть n ($n < d$) датчиков, которые получают данные системы с целью выявления неисправностей.
Пусть $x_i \in \{0,1\}$, $i = \overline{1,d}$ — вектор индикаторов всех неисправностей, а показания датчиков задаются уравнением:

$$y = Ax + v,$$

где $A \in \mathbb{R}^{n \times d}$ — матрица неисправностей, в которой каждый столбец a_i равен показаниям датчиков при i-ой неисправности, а $v \sim \mathcal{N}(\mathbf{0}, \sigma^2 I_n)$ — случайный вектор шума.
Цель состоит в том, чтобы выяснить, какой вид имеет вектор неисправностей x по вектору агрегированных показаний датчиков y.
Итак, распределение датчиков имеет вид $y \sim \mathcal{N}(Ax, \sigma^2 I_n)$, где $x_i \sim Bern(p)$, $i = \overline{1,d}$.
Найдем правдоподобие $L(x|y)$ с точностью до константы по формуле Байеса для апостериорной вероятности:

$$L(x|y) \propto p(y|x)p(x) = \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n \exp\left(-\frac{\|Ax-y\|_2^2}{2\sigma^2}\right) p^{\langle \mathbf{1},x\rangle}(1-p)^{d-\langle \mathbf{1},x\rangle}.$$

Мы хотим максимизировать правдоподобие, поэтому зададим функцию лосса следующим образом:

$$\ell(x) = -\log L(x|y) + \text{const} = \frac{1}{2\sigma^2}\|Ax-y\|_2^2 + \log\left(\frac{1}{p}-1\right)\langle \mathbf{1}, x\rangle.$$

Таким образом, наша задача оптимизации примет следующий вид:

$$\begin{aligned} \min_x \quad & \frac{1}{2\sigma^2}\|Ax-y\|_2^2 + \log\left(\frac{1}{p}-1\right)\langle \mathbf{1}, x\rangle \\ \text{s.t.} \quad & x_i \in \{0,1\}, \quad i = \overline{1,d}. \end{aligned}$$

Л1.2.5 Оптимизация в машинном обучении

Рассмотрим стандартную постановку задачи обучения с учителем. Пусть у нас есть обучающая выборка $\mathcal{L} = \{x_i, y_i\}_{i=1}^n$, где $x_i \in \mathbb{R}^d$ — наши наблюдения, y_i — метки, их вид зависит от рассматриваемой задачи, например:

- $y_i \in \mathbb{R}$ для регрессии,
- $y_i \in \{0,1\}$ для бинарной классификации,
- $y_i \in \{c_1, \ldots, c_m\}$ для многоклассовой классификации.

Пусть у нас есть параметризованная модель $f(x|\theta)$, которая делает предсказания по наблюдениям. Задача ставится следующим образом — нужно найти оптимальные параметры модели, такие, чтобы предсказания были наиболее близки к истинным меткам. Для этого вводится функция потерь $\ell\big(f(x_i|\theta), y_i\big)$, которая измеряет несоответствие предсказания модели и реального значения. Так как данные являются случайными величинами, нас интересует минимизация математического ожидания потерь, то есть *риска*:

$$R(\theta) = \mathbb{E}_{(X,Y)\sim P}\big[\ell\big(f(X|\theta), Y\big)\big].$$

Однако распределение $P(X, Y)$ нам зачастую неизвестно, и на практике мы заменяем риск на его эмпирическую аппроксимацию по выборке:

$$\hat{R}_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} \ell(f(x_i|\theta), y_i).$$

Таким образом, задача нахождения оптимальных параметров модели формулируется в парадигме минимизации эмпирического риска:

$$\min_{\theta} \quad \frac{1}{n}\sum_{i=1}^{n} \ell(f(x_i|\theta), y_i).$$

Л1.2.6 Оптимизация объёма эллипсоида

Предположим есть какая-то выборка из $\mathbb{R}^d$, мы хотим использовать MVE (Minimum Volume Ellipsoid) — метод поиска эллипсоида минимального объема, содержащего заданный набор точек, для оценивания параметра многомерного распределения из которого пришли наблюдения. Зная параметры такого эллипсоида, можно найти сдвиг b и матрицу ковариаций Σ искомого распределения.

Положительно определенное линейное преобразование $A = \Sigma^{\frac{1}{2}}$ и сдвиг на вектор b, примененные к единичной сфере, задают эллипсоид с площадью, которая в $\det\left(A^{-1}\right)$ раз больше, чем площадь единичной сферы, поэтому мы параметризуем внутреннюю часть эллипсоида следующим образом:

$$S = \left\{ x \in \mathbb{R}^d \;\middle|\; \left\|Ax + b\right\|_2^2 \leq 1 \right\}.$$

Но есть проблема: функцию $\det\left(A^{-1}\right)$ сложно минимизировать напрямую. Воспользуемся тем фактом, что $\log\det\left(A^{-1}\right) = -\log\det\left(A\right)$.
Получим следующую задачу оптимизации:

$$\begin{aligned} \min_{A,b} \quad & -\log\det\left(A\right) \\ \text{s.t.} \quad & \left\|Ax_i + b\right\|_2 \leq 1, \quad i = \overline{1,n} \\ & A \succ 0. \end{aligned}$$

Л1.2.7 Оптимизационная задача в логистике

Рассмотрим следующую постановку: мы владельцы крупного бизнеса, у нас есть склады, каким-то образом соединенные дорогами. Мы хотим расположить новые склады так, чтобы транспортировочная стоимость была минимальной.
Задачу можно рассматривать как минимизацию функции на графе. Пусть $y_j \in \mathbb{R}^2$, $j = \overline{1,m}$ — наши склады, вершины графа, а $x_i \in \mathbb{R}^2$, $i = \overline{1,n}$ — склады, которые мы хотим поставить. Формализуем целевую функцию в виде:

$$f(x) = \sum_{i=1}^{n}\sum_{j=1}^{m} \omega_{ij} g(\left\|x_i - y_j\right\|),$$

где $w_{ij} \geq 0$ — веса, которые содержат информацию о том, насколько склады важны между собой, g — штрафная функция от расстояний. Также мы можем захотеть, чтобы некоторые склады находились в определенных областях. Это условие можно записать следующим образом:

$$x_i \in S_i, \quad i = \overline{1,n},$$

где S_i — прямоугольник, в котором должен находиться i-ый склад.
Тогда итоговая задача принимает вид:

$$\begin{aligned} \min_{x} \quad & \sum_{i=1}^{n}\sum_{j=1}^{m} \omega_{ij} g\big(\|x_i - y_j\|\big) \\ \text{s.t.} \quad & x_i \in S_i, \quad i = \overline{1,n}. \end{aligned}$$

Л1.3 Общая схема методов оптимизации

Жизненный опыт подсказывает, что даже в одномерном случае задача оптимизации может не иметь аналитического решения (если вообще имеет). Если же решение существует, то на практике такую задачу обычно решают приближённо, итеративно находя лучшее решение, при этом допуская, что оно может не являться точным. Для этого применяются специальные алгоритмы, которые и называют *методами оптимизации*.
При создании методов оптимизации хочется добиться унифицированности. Потому что нет смысла искать лучший метод для решения конкретной задачи. Например, лучший алгоритм для решения задачи $\min_{x\in\mathbb{R}} x^2$ сходится за одну итерацию: этот метод просто всегда выдает ответ $x^* = 0$. Очевидно, что для других задач такой способ находить решение не пригоден. Метод должен быть пригоден для целого класса однотипных или похожих задач.
Так как метод разрабатывается для целого класса задач, то он не может иметь с самого начала полной информации о задаче (Л1.1). Вместо этого используется модель задачи, например, формулировка задачи, описание функциональных компонент, множества, на которых происходит оптимизация и т.д. Предполагается, что численный метод может накапливать специфическую информацию о задаче при помощи некоторого оракула. Под *оракулом* можно понимать некоторое устройство (программу, процедуру), которое отвечает на последовательные вопросы метода оптимизации о свойствах функции. В идеальном мире, наверное, можно было спросить у оракула всю информацию о решении задачи (Л1.1), но очевидно, что такого рода оракулы не интересны для практики. Мы будем работать с более приземленными вариантами оракулов, которые могут возвращать локальную информацию о целевой функции.

Определение Л1.1. Оракулы имеют разный порядок в зависимости от степени подробности, возвращаемой информации:

- *Оракул нулевого порядка* в запрашиваемой точке x возвращает значение целевой функции $f(x)$.
- *Оракул первого порядка* в запрашиваемой точке x возвращает значение целевой функции $f(x)$ и её градиент $\nabla f(x)$.
- *Оракул второго порядка* в запрашиваемой точке x возвращает значение целевой функции $f(x)$, её градиент $\nabla f(x)$ и её гессиан $\nabla^2 f(x)$.

Можно продолжать и до производных более высоких порядков.

Такое определение оракула имеет вполне прикладной смысл. В методах оптимизации обращение к оракулу часто реализуется посредством вызова отдельной от метода функции, программы или процедуры, возвращающей необходимую информацию, будь то значение функции, её градиента или её гессиана.
Методы оптимизации, которые мы будем рассматривать, описываются следующей схемой.

Алгоритм Л1.1 Общая итеративная схема метода оптимизации $\mathcal{M}$

Вход: начальная точка x^0, счётчик итераций $k = 0$, накапливаемая информационная модель решаемой задачи $I_{-1} = \varnothing$

while не выполнен критерий остановки $\mathcal{T}$ **do**
    Задать вопрос к оракулу $\mathcal{P}$ в точке x^k
    Пересчитать информационную модель: $I_k = I_{k-1} \cup (x^k, \mathcal{P}(x^k))$
    Применить правило метода $\mathcal{M}$ для получения новой точки x^{k+1} по модели I_k
    Положить $k = k + 1$
end while

Выход: $\hat{x}$

Алгоритм состоит из итераций, которые повторяются, пока не выполнен критерий остановки $\mathcal{T}$. В теле итерации мы обращаемся к оракулу, записываем его ответ в информационную модель, и получаем новую точку согласно правилу обновления.

Замечание Л1.1. В Алгоритме Л1.1 и далее мы будем использовать верхний индекс x^k для обозначения переменной x на k итерации.

Пример Л1.1. Рассмотрим задачу оптимизации

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

где функция $f(x)$ дифференцируема. Предположим, что в любой точке мы можем посчитать её градиент. Тогда можно воспользоваться следующим методом:

Алгоритм Л1.2 Градиентный спуск с постоянным размером шага

Вход: стартовая точка $x^0 \in \mathbb{R}^d$, размер шага $\gamma > 0$, количество итераций K

for $k = 0, 1, \ldots, K-1$ **do**
    $x^{k+1} = x^k - \gamma \nabla f(x^k)$
end for

Выход: x^K

С градиентным спуском мы будем подробнее знакомиться уже в следующих параграфах. Сейчас мы хотим лишь понять, как он ложится в общую модель. В данном случае оракул $\mathcal{P}$ — оракул первого порядка, который возвращает градиенты ∇f в запрашиваемых точках. При этом в информационную модель мы каждую итерацию кладём значения $\big(x^k, \mathcal{P}(x^k)\big) = \big(x^k, \nabla f(x^k)\big)$. Сам же метод градиентного спуска $\mathcal{M}$ использует только последнюю пару точек из информационной модели и получает новую точку по правилу: $x^{k+1} = x^k - \gamma \nabla f(x^k)$.

Замечание Л1.2. В общей итеративной схеме метода оптимизации (Алгоритм Л1.1) информационная модель растёт линейно с номером итерации k. Пример Л1.1 показывает, что для практических методов вовсе не обязательно хранить всю информационную модель — градиентному спуску нужна только текущая точка x^k и $\mathcal{P}(x^k) = \nabla f(x^k)$.

Замечание Л1.3. В Примере Л1.1 в качестве критерия остановки $\mathcal{T}$ используется ограничение на количество итераций, но может быть и требование на достижение точности решения.

Определение Л1.2. Под утверждением, что *задача решена с точностью ε (найдено ε-решение/выполнен критерий остановки)*, можно понимать:

- По аргументу: $\left\|x^k - x^*\right\|_2 \leq \varepsilon$.
- По значению функции: $f\left(x^k\right) - f^* \leq \varepsilon$.
- По норме градиента: $\left\|\nabla f\left(x^k\right)\right\|_2 \leq \varepsilon$.

Отметим, что первые два критерия являются теоретическими. Мы их будем использовать при доказательстве сходимости методов и оценке их скорости работы. На практике же сложно, например, оценить $\left\|x^k - x^*\right\|_2$, так как мы не знаем расположение точки x^*. При этом может оказаться полезным критерий сходимости вида: $\left\|x^{k+1} - x^k\right\|_2 \leq \varepsilon$. Из такого критерия не следуют какие-либо гарантии на $\left\|x^k - x^*\right\|_2$ — можно привести массу очевидно плохих методов, которые стоят на месте вдали от решения. Но верно обратное, а именно, если $\left\|x^{k+1} - x^*\right\|_2 \leq \left\|x^k - x^*\right\|_2 \leq \varepsilon/2$, то по неравенству треугольника:

$$\left\|x^{k+1} - x^k\right\|_2 \leq \left\|x^{k+1} - x^*\right\|_2 + \left\|x^k - x^*\right\|_2 \leq \varepsilon.$$

Поэтому при имеющейся теории или интуиции о том, что $\left\|x^k - x^*\right\|_2 \to 0$, критерий вида $\left\|x^{k+1} - x^k\right\|_2 \leq \varepsilon$ может быть более чем полезен.

Похожая ситуация с критерием на основе $f\left(x^k\right) - f^*$ с той лишь разницей, что для некоторых задач мы можем знать значение f^* (например, для $\min_{x \in \mathbb{R}^d} \left\|Ax - b\right\|_2^2$, где $b \in \operatorname{Im} A$, известно $f^* = 0$).

Следить за поведением $\left\|\nabla f\left(x^k\right)\right\|_2$ кажется более практичным. Это является правдой для определенного класса функций, пока мы рассматриваем безусловные варианты задачи (Л1.1) (без ограничений и с $\mathcal{X} = \mathbb{R}^d$). В общем же случае, такого рода критерий может ничего не сказать — достаточно рассмотреть задачу $\min_{x \in [1,2]} x^2$. Существуют и другие критерии, с которыми познакомимся далее, которые подойдут и для $\mathcal{X} \neq \mathbb{R}^d$.

Л1.4 Сложность методов. Верхние и нижние оценки

Для понимания того, насколько эффективно/быстро/дёшево работают методы оптимизации, как их сравнивать между собой, необходимо ввести формальные понятия для описания *сложности* методов оптимизации:

Определение Л1.3.

- *Аналитическая/Оракульная сложность* — число обращений метода к оракулу, необходимое для решения задачи с точностью ε.
- *Арифметическая/Временная сложность* — общее число вычислений (включая работу оракула), необходимых для решения задачи с точностью ε.
- *Итерационная сложность* — общее число итераций метода, необходимых для решения задачи с точностью ε.

Арифметическая сложность даёт полное количество атомарных операций (сложений/умножений двух чисел), которое выполнил метод для решения задачи оптимизации. Время, которое затратит алгоритм, можно считать пропорциональным арифметической сложности.

Оракульная сложность учитывает только количество вызовов оракула, что пропорционально общему количеству атомарных операций, которые выполнил оракул в ходе

работы метода. На самом деле, часто оракульной сложности бывает достаточно для оценки времени работы метода, так как вычисления оракулов являются наиболее дорогой операцией во всем методе. Рассмотрим, например, алгоритм из Примера Л1.1 для целевой функции f из Примера С3.3. Вычисления ∇f требуют умножения матрицы на вектор, а все остальные операции алгоритма производятся с вектором.
Итерационная сложность несёт меньше всего информации о времени работы метода, так как разные методы могут иметь абсолютно разную арифметическую сложность вызова оракула. Поэтому из того, что один метод достигает ε-решения за меньшее число итераций, вообще говоря, не следует, что он работает быстрее по времени.
Теперь попробуем понять, а какие гарантии на сложность мы вообще можем дать при поиске ε-решения (Л1.1). Для этого рассмотрим следующую, на первый взгляд, довольно простую задачу оптимизации:

$$\min_{x \in [0,1]^d} f(x), \tag{Л1.2}$$

где $[0,1]^d = \left\{ x \in \mathbb{R}^d \;\middle|\; 0 \le x_i \le 1,\ i = \overline{1,d} \right\}$ — кубик. При этом мы дополнительно предположим, что функция $f(x)$ является *M–липшицевой* на $[0,1]^d$ относительно ℓ_∞–нормы, т.е. для любых $x, y \in [0,1]^d$ справедливо

$$\left|f(x) - f(y)\right| \le M\|x - y\|_\infty = M \max_{i=\overline{1,d}} |x_i - y_i|. \tag{Л1.3}$$

Замечание Л1.4. Множество $[0,1]^d$ является ограниченным и замкнутым, т.е. компактом, а из липшицевости функции f следует и её непрерывность. Поэтому задача (Л1.2) имеет решение, так как по теореме Вейерштрасса (см. Теорему 2 Параграфа 7 в [27]) непрерывная на компакте функция достигает своих минимального и максимального значений.

В поисках решения давайте ограничимся методами нулевого порядка (т.е. методы, которые могут вызывать оракул, возвращающий значения целевой функции). Более формально:

- **Цель:** найти $\hat{x} \in [0,1]^d$: $f(\hat{x}) - f^* \le \varepsilon$, где f удовлетворяет (Л1.3).
- **Класс методов:** методы нулевого порядка.

Рассмотрим следующий довольно незатейливый алгоритм для решения поставленной цели:

Алгоритм Л1.3 Метод равномерного перебора

Вход: целочисленный параметр перебора $p \ge 1$

1: Сформировать $(p+1)^d$ точек вида $x_{(i_1,\dots,i_d)} = \left(\frac{i_1}{p}, \frac{i_2}{p}, \dots, \frac{i_d}{p}\right)^\top$, где вектор $(i_1, \dots, i_d) \in \{\overline{0,p}\}^d$
2: Среди точек $x_{(i_1,\dots,i_d)}$ найти точку $\hat{x}$ с наименьшим значением целевой функции f.

Выход: $\hat{x}$

Попробуем получить какие-нибудь гарантии на решение, которое находит данный алгоритм.

Теорема Л1.1 (Теорема 1.1.1. из [29])**.** Пусть задача оптимизации (Л1.2) с M-липшицевой целевой функцией f решается с помощью метода равномерного перебора с параметром p (Алгоритм Л1.3). Тогда справедлива следующая оценка

сходимости:

$$f(\hat{x}) - f^* \le \frac{M}{2p},$$

откуда следует, что методу равномерного перебора нужно в худшем случае

$$\left(\left\lfloor \frac{M}{2\varepsilon} \right\rfloor + 2\right)^d$$

обращений к оракулу, чтобы гарантировать $f(\hat{x}) - f^* \le \varepsilon$.

Доказательство. Пусть x^* — решение задачи (возможно, не единственная точка минимума функции f). Тогда в построенной методом сетке из точек найдется такая точка $x_{(i_1,\dots,i_d)}$, что

$$x := x_{(i_1,\dots,i_d)} \preceq x^* \preceq x_{(i_1+1,\dots,i_d+1)} =: x',$$

где знак $\preceq$ применяется покомпонентно. Точки x и x' определяют углы кубика сетки, в котором лежит x^*, то есть $x_i^* \in [x_i, x_i']$ для всех $i = \overline{1,d}$. Отметим, что стороны данного кубика имеют длины $x_i' - x_i = \frac{1}{p}$.

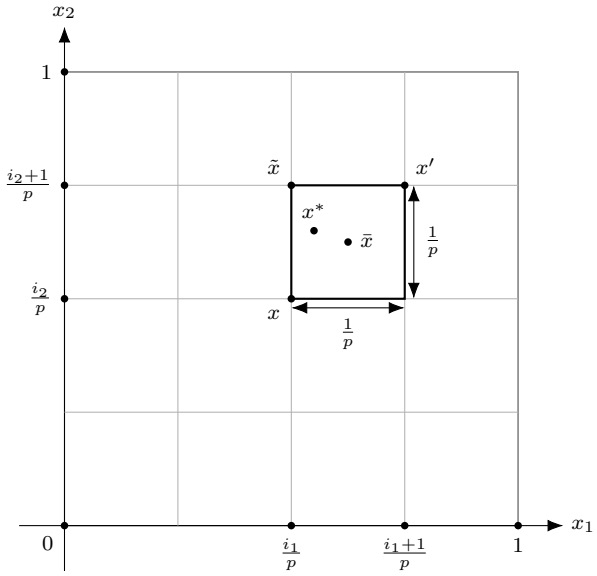


Рис. Л1.1: Построенная сетка и кубик в ней, которому принадлежит точка x^*.

Кроме того, рассмотрим точки $\bar{x}$ и $\tilde{x}$ такие, что $\bar{x} = \frac{x+x'}{2}$ и

$$\tilde{x}_i = \begin{cases} x_i', & \text{если } x_i^* \ge \bar{x}_i, \\ x_i, & \text{иначе.} \end{cases}$$

$\bar{x}$ — центр кубика, а $\tilde{x}$ определяет ближайший к x^* уголок кубика. Заметим, что $\tilde{x}$ принадлежит сетке и $\left|\tilde{x}_i - x_i^*\right| \le \frac{1}{2p}$, так как в худшем случае x^* расположен в $\bar{x}$, а

значит, равноудален от всех уголков. Тогда

$$\left\|\tilde{x} - x^*\right\|_\infty = \max_{i \in \overline{1,d}} |\tilde{x}_i - x_i^*| \leq \frac{1}{2p}.$$

Поскольку $f(\hat{x}) \leq f(\tilde{x})$ (так как $f(\hat{x})$ — минимум по всем точкам сетки), получаем

$$f(\hat{x}) - f^* \leq f(\tilde{x}) - f^* \leq M\left\|\tilde{x} - x^*\right\|_\infty \leq \frac{M}{2p}.$$

Выписанная выше оценка достигается методом равномерного перебора за $(p+1)^d$ обращений к оракулу. Ограничим сверху значением ε:

$$f(\hat{x}) - f^* \leq \frac{M}{2p} \leq \varepsilon \implies p \geq \frac{M}{2\varepsilon}.$$

Возьмем $p = \left\lfloor \frac{M}{2\varepsilon} \right\rfloor + 1$, тогда метод сделает $\left(\left\lfloor \frac{M}{2\varepsilon} \right\rfloor + 2\right)^d$ обращений к оракулу нулевого порядка и выдаст ответ с ε-точностью. ■

Интересно оценить порядок величин, которые мы получили:

- Предположим, $M = 2$, $d = 13$ и $\varepsilon = 0.01$, то есть размерность задачи сравнительно небольшая (можно сказать, что никакая для задач оптимизации в современных приложениях) и точность решения задачи не слишком высокая.
- Необходимое число обращений к оракулу: $\left(\left\lfloor \frac{M}{2\varepsilon} \right\rfloor + 2\right)^d = 102^{13} > 10^{26}$.
- Сложность одного вызова оракула не менее 1 арифметической операции.
- Производительность современной видеокарты порядка 10^{14} арифметических операций в секунду (82 TFLOPS на NVIDIA GeForce RTX 4090).
- Общее время: хотя бы 10^{12} секунд, что займёт порядка 30 тысяч лет.

Выводы получаются не очень оптимистичные. Но есть надежда, что наш теоретический анализ из Теоремы Л1.1 плох, либо метод не самый лучший, и существует какой-нибудь зубодробительный алгоритм, который будет давать куда более приятные гарантии на число оракульных вызовов.
В Теореме Л1.1 мы получили так называемые верхние оценки на гарантии нахождения решения, но существует и обратное понятие — нижние оценки.

Определение Л1.4. Пусть нам дан некоторый класс методов, а также некоторый класс задач оптимизации.

- *Верхняя оценка* — гарантия, что рассматриваемый метод из класса методов на любой задаче из класса решаемых задач имеет оракульную/арифметическую/итерационную сложность не хуже, чем утверждает верхняя оценка.
- *Нижняя оценка* — гарантия, что для любого метода из класса методов существует «плохая» задача из класса решаемых задач, такая, что метод имеет оракульную/арифметическую/итерационную сложность не лучше, чем утверждает нижняя оценка.

Напомним, что в этом параграфе мы рассматриваем класс методов нулевого порядка, т.е. все методы, которые каким-либо образом используют информацию о значениях

функции (но не градиентов). Также класс задач, которые мы сейчас рассматриваем, описывается с помощью (Л1.2) и (Л1.3).
Нижние оценки нужны, чтобы понять, насколько полученные верхние оценки хороши. В частности, если верхние и нижние оценки совпали, это означает оптимальность предложенного метода для данного класса задач. Но если нижние оценки не совпали с верхними, то это может ничего не значить — возможно, в нижних оценках подобрана недостаточно «плохая» функция, а возможно, при получении верхних оценок пришлось загрубить теоретические выкладки. Нижние оценки можно получать не только для класса методов, но и для определенного метода, чтобы доказать оптимальность и неулучшаемость теоретического анализа и полученной верхней оценки.
Вернемся к нашей задаче: (Л1.2) + (Л1.3) и методам, оперирующим информацией нулевого порядка. Следующая теорема представит нижние оценки на оракульную сложность.

Теорема Л1.2 (Теорема 1.1.2. из [29])**.** Пусть задача оптимизации (Л1.2) с M-липшицевой целевой функцией f решается с помощью методов нулевого порядка с точностью $\varepsilon < \frac{M}{2}$. Тогда справедлива следующая оценка сходимости:

$$\left(\left\lfloor \frac{M}{2\varepsilon} \right\rfloor - 1\right)^d - 1$$

обращений к оракулу, чтобы гарантировать $f(\hat{x}) - f^* \leq \varepsilon$.

Доказательство. Доказываем от противного: предположим, что существует такой метод, который решает задачу за $N < \left(\left\lfloor \frac{M}{2\varepsilon} \right\rfloor - 1\right)^d - 1$ обращений к оракулу с точностью ε (по функции). Построим такую функцию, на которой метод не сможет найти ε-решение, при помощи сопротивляющегося оракула: пусть изначально наша целевая функция $f(x)$ всюду равна 0. Запустим метод, он запросит значение f в N точках, везде получит 0 и выдаст какую-то точку (возможно, отличную от всех предыдущих N), как ответ. В итоге мы в ходе работы алгоритма заглянули в $N+1$ точку.
Заметим, что силу непрерывности функции $\frac{1}{x}$, можно подобрать число $\delta > 0$:

$$\frac{M}{2\varepsilon} - 1 < \frac{M}{(2+\delta)\varepsilon}.$$

Тогда, в силу монотонности округления:

$$\left\lfloor \frac{M}{2\varepsilon} \right\rfloor - 1 \leq \left\lfloor \frac{M}{(2+\delta)\varepsilon} \right\rfloor.$$

Теперь из этого неравенства и предположения в начале получим оценку на $N+1$:

$$N + 1 < \left(\left\lfloor \frac{M}{2\varepsilon} \right\rfloor - 1\right)^d \leq \left\lfloor \frac{M}{(2+\delta)\varepsilon} \right\rfloor^d.$$

Обозначим $p = \left\lfloor \frac{M}{(2+\delta)\varepsilon} \right\rfloor$. Рассмотрим равномерную сетку из $(p+1)^d$ точек, по аналогии с тем, как сделано в методе равномерного перебора. Из этих точек сформируем p^d кубиков с ребрами длины $\frac{1}{p}$. Воспользуемся тем, что смогли заглянуть в $N+1$ точку и $N+1 < p^d$. Тогда по принципу Дирихле (точки — кролики, кубики с углами в соседних вершинах сетки — клетки) найдётся такой кубик

$$C = \left\{ x \;\middle|\; \tilde{x} \preceq x \preceq \tilde{x} + \frac{1}{p}\mathbf{1} \right\},$$

который не содержит внутри себя ни одной из $N+1$ точки (в том числе и выхода метода). Здесь $\tilde{x}$ и $\tilde{x}+\frac{1}{p}\mathbf{1}$ — точки из сетки с шагом $\frac{1}{p}$, а $\mathbf{1}$ — вектор из единиц.

Пусть x^* — это центр кубика C, т.е. $x^* = \tilde{x} + \frac{1}{2p}\mathbf{1}$. Теперь немного модифицируем целевую функцию f:

$$\bar{f}(x) = \min\left\{0,\ M\|x - x^*\|_\infty - \left(1+\frac{\delta}{2}\right)\varepsilon\right\}.$$

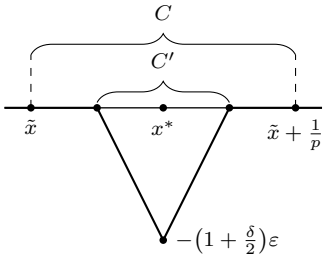


Рис. Л1.2: График функции $\bar{f}(x)$ в окрестности x^*.

Функция $\bar{f}(x)$ липшицева с константой M относительно ℓ_∞-нормы и принимает своё минимальное значение $-\varepsilon\left(1+\frac{\delta}{2}\right)$ в точке x^*. Более того, по определению функция $\bar{f}(x)$ отлична от нуля только внутри куба

$$C' = \left\{ x \;\middle|\; \|x - x^*\|_\infty < \frac{\varepsilon}{M}\left(1+\frac{\delta}{2}\right) \right\},$$

который лежит внутри куба C. Это так, поскольку согласно выбору $p = \left\lfloor \frac{M}{(2+\delta)\varepsilon} \right\rfloor$ верно

$$2p = 2\left\lfloor \frac{M}{(2+\delta)\varepsilon} \right\rfloor \le \frac{2M}{(2+\delta)\varepsilon},$$

откуда получаем

$$\frac{\varepsilon}{M}\left(1+\frac{\delta}{2}\right) \le \frac{1}{2p}.$$

Вложение $C' \subset C$ позволяет нам сказать, что алгоритм заглянул только в точки с нулевым значением, поскольку вне C' функция тождественно равна нулю.

Итак, мы привели пример функции, минимум которой достигается в точке x^* и равен $-\varepsilon\left(1+\frac{\delta}{2}\right)$, при этом рассматриваемый метод выдал точку $\hat{x}$ со значением $f(\hat{x}) = 0$. Следовательно, рассмотренный метод на данной функции нашел решение с точностью $f(\hat{x}) - f^* = \varepsilon\left(1+\frac{\delta}{2}\right) > \varepsilon$. Противоречие. ■

Получается, что нижняя оценка из Теоремы Л1.2 не особо лучше Теоремы Л1.1. Зависимость $\left(\frac{M}{\varepsilon}\right)^d$ присутствует в обоих результатах. А это значит, что для класса липшицевых функций в общем случае все довольно печально. Нужны дополнительные предположения на задачу (Л1.1), чтобы гарантировать более оптимистичные гарантии. Но это вопрос уже следующих параграфов.

Л1.5 Скорость сходимости методов

Поймём, с каким характером верхних оценок на скорость сходимости/скорость приближения к решению мы хотим и будем иметь дело в дальнейшем.

Определение Л1.5. Выделим основные типы сходимостей:

- *Сублинейная*: $\left\|x^k - x^*\right\|_2 \leq \frac{C}{k^\alpha},\ C > 0,\ \alpha > 0.$
- *Линейная*: $\left\|x^k - x^*\right\|_2 \leq Cq^k,\ C > 0,\ 0 < q < 1.$
- *Сверхлинейная*: $\left\|x^k - x^*\right\|_2 \leq Cq^{k^p},\ C > 0,\ 0 < q < 1,\ p > 1.$
- *Квадратичная*: $\left\|x^k - x^*\right\|_2 \leq Cq^{2^k},\ C > 0,\ 0 < q < 1.$
  Квадратичная локальная: $\left\|x^{k+1} - x^*\right\|_2 \leq C\left\|x^k - x^*\right\|_2^2, C > 0.$

Из определения очевидно, что типы следуют в порядке увеличения скоростей сходимости. Приведённые в Определении Л1.5 типы сходимостей могут относиться как к арифметической, так и к оракульной или итерационной сложностям. Постоим графики представителей каждого из типов сходимостей на Рисунке Л1.3.

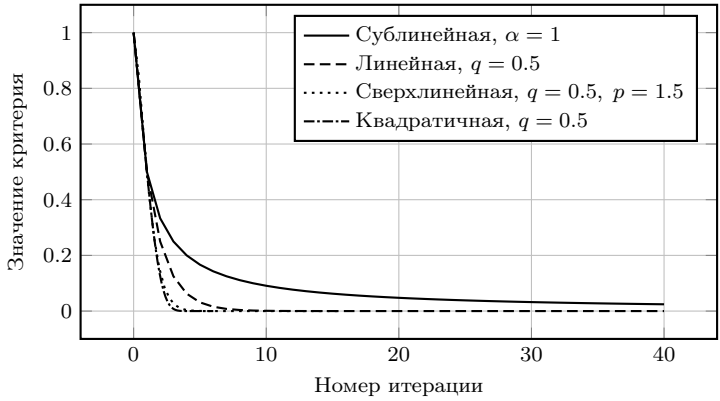


Рис. Л1.3: Основные типы сходимостей, линейный масштаб по оси критерия.

Можно лишь примерно оценить, как данные виды сходимостей визуально соотносятся между собой. Однако, спустя 5 итераций сверхлинейная и квадратичная сходимости становятся неразличимы, а после 10 итерации все, кроме сублинейной сходимости, сливаются с нулем. Естественным образом возникает вопрос: как визуально сравнивать скорости сходимостей, когда значения критерия близки к нулю? Решением этой проблемы является логарифмический масштаб оси критерия. Построим новый график с предложенным подходом (Рисунок Л1.4).
Теперь каждый график визуально отделяется от других. Также из рисунка становится понятно, почему линейная скорость получила такое название.
Отдельно рассмотрим последний тип сходимости и поймём, почему он так называется. Запишем неравенство из определения:

$$\left\|x^{k+1} - x^*\right\|_2 \leq C\left\|x^k - x^*\right\|_2^2.$$

Запустим рекурсию: продолжим неравенства, записав оценку сначала для $\left\|x^k - x^*\right\|_2$

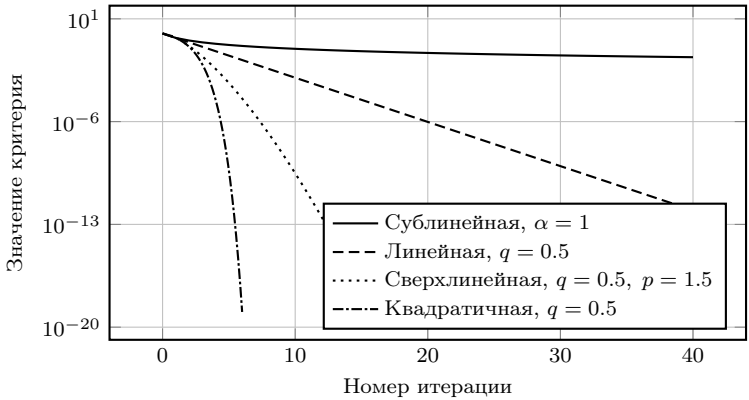


Рис. Л1.4: Основные типы сходимостей, логарифмический масштаб по оси критерия.

через $\left\|x^{k-1} - x^*\right\|_2$, потом для $\left\|x^{k-1} - x^*\right\|_2$ и так далее, пока не дойдем до $\left\|x^0 - x^*\right\|_2$.

$$\left\|x^{k+1} - x^*\right\|_2 \leq C\left\|x^k - x^*\right\|_2^2 \leq C \cdot C^2\left\|x^{k-1} - x^*\right\|_2^4 \leq \ldots \leq \left(\prod_{m=0}^{k} C^{2^m}\right)\left\|x^0 - x^*\right\|_2^{2^{k+1}}$$

$$= \left(C^{\sum_{m=0}^{k} 2^m}\right)\left\|x^0 - x^*\right\|_2^{2^{k+1}}.$$

Если учесть, что $\sum_{m=0}^{k} 2^m = 2^{k+1} - 1$, то можно записать

$$\left\|x^{k+1} - x^*\right\|_2 \leq \frac{1}{C}\left(C\left\|x^0 - x^*\right\|_2\right)^{2^{k+1}}.$$

Получили определение квадратичной скорости сходимости, где $C \to \frac{1}{C}$, $q \to C\left\|x^0 - x^*\right\|_2$. Вспомним требование на q:

$$0 < q = C\left\|x^0 - x^*\right\|_2 < 1,$$

что можно понимать как требование на достаточно хорошую стартовую точку:

$$\left\|x^0 - x^*\right\|_2 < \frac{1}{C}.$$

Поэтому эта сходимость называется локальной: она начинает работать, только если начальное приближение попало в окрестность решения.

Замечание Л1.5. Из Определения Л1.5, где даны скорости приближения к решению через k арифметических операций/оракульных вызовов/итераций, можно легко получить оценки на соответствующие сложности.
Пусть для некоторого метода мы доказали оценку на сходимость: $\left\|x^k - x^*\right\|_2 \leq \frac{C}{k^\alpha}$, где k — номер арифметических операции/оракульного вызова/итерации. Тогда, чтобы гарантировать точность ε по аргументу, нам необходимо сделать $k \geq \left(\frac{C}{\varepsilon}\right)^{1/\alpha}$:

$$\left\|x^k - x^*\right\|_2 \leq \frac{C}{k^\alpha} \leq \varepsilon \implies k^\alpha \geq \frac{C}{\varepsilon}.$$

Аналогичные манипуляции можно проделать со всеми типами сходимости из Определения Л1.5.

Л2 Условия оптимальности. Выпуклость и гладкость

Напомним, что ключевой задачей курса является (Л1.1). Начнем изучение с задачи без ограничений (безусловной задачи оптимизации):

$$\min_{x \in \mathbb{R}^d} f(x). \tag{Л2.1}$$

Формализуем понятия решения данной задачи.

Л2.1 Условия оптимальности

Определение Л2.1. Точка x^* называется *локальным минимумом* функции f на $\mathbb{R}^d$ (локальным решением задачи минимизации f на $\mathbb{R}^d$), если существует $r > 0$ такое, что для любого $y \in B_2^d(r, x^*) = \left\{ y \in \mathbb{R}^d \mid \|y - x^*\|_2 \leq r \right\}$ следует, что $f(x^*) \leq f(y)$.

Определение Л2.2. Точка x^* называется *глобальным минимумом* функции f на $\mathbb{R}^d$ (глобальным решением задачи минимизации f на $\mathbb{R}^d$), если для любого $x \in \mathbb{R}^d$ следует, что $f(x^*) \leq f(x)$.

Понятно, что глобальный минимум является одновременно и локальным. Попробуем понять, какие есть свойства локального минимума. В частности, следующая теорема приводит необходимое условие локального минимума безусловной задачи оптимизации (Л2.1).

Теорема Л2.1 (Теорема 1.2.1. из [29])**.** Пусть x^* — локальный минимум функции f на $\mathbb{R}^d$. Тогда если f дифференцируема, то $\nabla f(x^*) = 0$.

Доказательство. Пойдем от противного и предположим, что x^* — локальный минимум, но $\nabla f(x^*) \neq 0$. Разложим функцию f в ряд в окрестности локального минимума:

$$f(x) = f(x^*) + \langle \nabla f(x^*), x - x^* \rangle + o(\|x - x^*\|_2), \tag{Л2.2}$$

где $\lim_{x \to x^*} \frac{o(\|x - x^*\|_2)}{\|x - x^*\|_2} = 0$.

Рассмотрим $x_\lambda = x^* - \lambda \nabla f(x^*)$. Найдем $\lambda_1 > 0$ такое, что для любого $0 < \lambda \leq \lambda_1$ можно гарантировать, что $\|x_\lambda - x^*\|_2 \leq r$, т.е. x_λ попадает в нужную окрестность из определения локального минимума (Определение Л2.1). Понятно, что такое λ_1 можно найти в силу $r > 0$, а $\nabla f(x^*)$ конечно. Тогда для любого $0 < \lambda \leq \lambda_1$ справедливо

$$f(x_\lambda) \geq f(x^*).$$

При этом разложение в ряд (Л2.2) для точек x_λ имеет вид:

$$\begin{aligned} f(x_\lambda) &= f(x^*) + \langle \nabla f(x^*), x_\lambda - x^* \rangle + o(\|x_\lambda - x^*\|_2) \\ &= f(x^*) - \lambda \|\nabla f(x^*)\|_2^2 + o(\lambda \|\nabla f(x^*)\|_2) \end{aligned}$$

Набросим еще одно ограничение на «малость» λ. А именно, найдем $\lambda_2 > 0$ такое, что для любого $0 < \lambda \leq \min\{\lambda_1, \lambda_2\}$ выполнено

$$|o(\lambda \|\nabla f(x^*)\|_2)| \leq \frac{\lambda}{2} \|\nabla f(x^*)\|_2^2.$$

Тогда для любого $\lambda > 0$ такого, что $\lambda \leq \min\{\lambda_1, \lambda_2\}$, следует

$$f(x_\lambda) \leq f(x^*) - \frac{\lambda}{2}\|\nabla f(x^*)\|_2^2.$$

Пришли к противоречию, что x^* — локальный минимум. А значит $\nabla f(x^*) = 0$. ■

Наша цель — находить глобальный минимум, а локальных хотелось бы наоборот избегать (зависит от конкретной задачи, но обычно цель именно такая). Как мы поняли в Параграфе Л1, без дополнительных предположений на задачу (Л1.1) в худшем случае полный равномерный перебор является оптимальным алгоритмом. Поэтому пора ввести новые понятия, которые помогут сузить класс изучаемых задач и построить оптимистичную теорию поиска глобального минимума.

Л2.2 Выпуклость

Первое понятие — это выпуклость целевой функции f в задаче (Л2.1).

Определение Л2.3. Пусть дана непрерывно дифференцируемая на $\mathbb{R}^d$ функция $f : \mathbb{R}^d \to \mathbb{R}$. Будем говорить, что она является *выпуклой*, если для любых $x, y \in \mathbb{R}^d$ выполнено

$$f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle.$$

Наряду с выпуклостью также вводят еще одно, более сильное понятие.

Определение Л2.4. Пусть дана непрерывно дифференцируемая на $\mathbb{R}^d$ функция $f : \mathbb{R}^d \to \mathbb{R}$. Будем говорить, что она является *μ-сильно выпуклой* ($\mu > 0$), если для любых $x, y \in \mathbb{R}^d$ выполнено

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

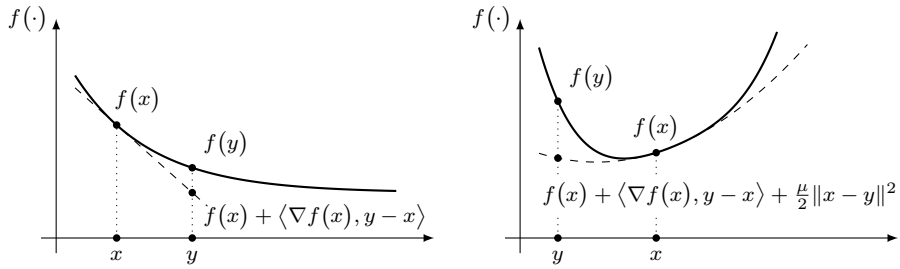
Из определений видно, что выпуклость — это сильная выпуклость с $\mu = 0$.
В Параграфе С4 также даны другие определения выпуклости и сильной выпуклости, которые не требуют дифференцируемости. В случае дифференцируемой функции данные выше определения эквивалентны определениям из Параграфа С4.
Физический смысл Определений Л2.3 и Л2.4 проиллюстрирован на Рисунке Л2.1: выпуклая функция в любой точке «подперта» снизу линейно аппроксимацией, а сильно выпуклая — квадратичной функцией.
Введя новые классы функций попробуем понять, что теперь можно сказать про точки минимума/решения задач оптимизации с такими целевыми функциями.

Теорема Л2.2. Пусть дана выпуклая на $\mathbb{R}^d$ функция $f : \mathbb{R}^d \to \mathbb{R}$. Если для некоторой точки $x^* \in \mathbb{R}^d$ верно, что $\nabla f(x^*) = 0$, то x^* — глобальный минимум f на всем $\mathbb{R}^d$.

Доказательство. Достаточно записать определение выпуклости:

$$f(x) \geq f(x^*) + \langle \nabla f(x^*), x - x^* \rangle = f(x^*).$$

■



(a) Выпуклость

(b) Сильная выпуклость

Рис. Л2.1: Иллюстрация понятий выпуклости и μ-сильной выпуклости.

Получается, что в случае выпуклой функции локальный минимум совпадает с глобальным. А значит $\nabla f(x^*) = 0$ является необходимым и достаточным условием.
Докажем несколько полезных фактов о минимумах выпуклых безусловных задач оптимизации.

Теорема Л2.3. Пусть дана выпуклая на $\mathbb{R}^d$ функция f. Тогда

- всякий локальный минимум f на $\mathbb{R}^d$ является и глобальным,
- если дополнительно f сильно выпуклая, то минимум существует и единственен.

Доказательство. Докажем последовательно пункты теоремы.

- Пусть x^* — локальный минимум. Согласно необходимому условию минимума функции (Теорема Л2.1)
$$\nabla f(x^*) = 0.$$
Пусть также x — произвольная точка из $\mathbb{R}^d$. Воспользуемся выпуклостью f:
$$f(x) \geq f(x^*) + \langle \nabla f(x^*), x - x^* \rangle = f(x^*).$$
Получаем, что $f(x^*) \leq f(x)$ для любого $x \in \mathbb{R}^d$. Значит, x^* — глобальный минимум.

- Покажем существование глобального минимума сильно выпуклой функции. Выберем произвольную точку $x^0 \in \mathbb{R}^d$ и положим $\beta = f(x^0)$. Если $\nabla f(x^0) = 0$, то x^0 — глобальный минимум (Теорема Л2.2). Далее рассуждаем в предположении, что $\nabla f(x^0) \neq 0$. Рассмотрим множество подуровня функции f:
$$L_\beta = \left\{ x \in \mathbb{R}^d \;\middle|\; f(x) \leq \beta \right\}.$$
По сильной выпуклости и неравенству Коши-Буняковского-Шварца (0.3) имеем:
$$\begin{aligned} f(x) &\geq f(x^0) + \langle \nabla f(x^0), x - x^0 \rangle + \frac{\mu}{2}\|x - x^0\|_2^2 \\ &\geq \beta - \|\nabla f(x^0)\|_2 \|x - x^0\|_2 + \frac{\mu}{2}\|x - x^0\|_2^2. \end{aligned}$$

Из условия $f(x) \leq \beta$ вытекает:

$$\frac{\mu}{2}\|x - x^0\|_2^2 - \|\nabla f(x^0)\|_2\|x - x^0\|_2 \leq 0 \implies \|x - x^0\|_2 \leq \frac{2\|\nabla f(x^0)\|_2}{\mu}.$$

Значит

$$L_\beta \subseteq B_R(x^0) = \left\{ x \in \mathbb{R}^d \;\middle|\; \|x - x^0\|_2 \leq R \right\}, \text{ где } R = \frac{2\|\nabla f(x^0)\|_2}{\mu}.$$

Из этого вложения получаем, что вне шара $B_R(x^0)$ значения функции f строго больше β:

$$f(x) > \beta, \ \forall x \notin B_R(x^0).$$

При этом, шар $B_R(x^0)$ замкнутый и ограниченный в конечномерном $\mathbb{R}^d$, следовательно, компактен. Тогда по теореме Вейерштрасса (см. Теорему 2 Параграфа 7 Главы 2 в [27]) f достигает на $B_R(x^0)$ своего минимума, который автоматически является глобальным (вне $B_R(x^0)$ значения f больше β).

Теперь покажем единственность. Пусть x^* — глобальный минимум, $x \in \mathbb{R}^d$. Поскольку f является μ-сильно выпуклой:

$$\begin{aligned} f(x) &\geq f^* + \langle \nabla f(x^*), x - x^* \rangle + \frac{\mu}{2}\|x - x^*\|_2^2 \\ &= f^* + \frac{\mu}{2}\|x - x^*\|_2^2. \end{aligned}$$

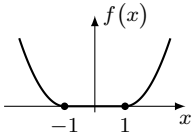
Слагаемое с градиентом занулилось, поскольку в точке оптимума $\nabla f(x^*) = 0$. Из свойств нормы получаем, что $f(x)$ достигает минимума только в точке x^*. Значит, если решение существует, то оно единственно.

■

В теореме не сказано про существование или единственность минимума выпуклой функции. Приведем два примера выпуклых функций, где эти свойства могут не выполняться.

Пример Л2.1. Покажем, что у выпуклой функции может быть больше одного минимума. Рассмотрим кусочно-заданную функцию f:

$$f(x) = \begin{cases} (x-1)^2, & x \in (1, +\infty) \\ 0, & x \in [-1, 1] \\ (x+1)^2, & x \in (-\infty, -1). \end{cases}$$



Она является выпуклой, однако, все точки отрезка $[-1, 1]$ доставляют минимум f, то есть, у f несчетное число точек минимума.

Пример Л2.2. Теперь приведем пример, когда выпуклая функция не имеет минимума на $\mathbb{R}$. Для этого подойдет линейная функция

$$f(x) = x.$$

Действительно, она выпукла, поскольку в каждой точке совпадает со своей линейной аппроксимацией, при этом функция не ограничена снизу, поэтому ни одна из точек $\mathbb{R}$ не является точкой минимума f.

Л2.3 Гладкость

Введем еще одно свойство, которое также пригодится для того, чтобы строить теорию сходимости оптимизационных методов.

Определение Л2.5. Пусть дана непрерывно дифференцируемая на $\mathbb{R}^d$ функция $f : \mathbb{R}^d \to \mathbb{R}$. Будем говорить, что данная функция имеет *L-Липшицев градиент* (говорить, что она является L-гладкой), если для любых $x, y \in \mathbb{R}^d$ выполнено

$$\big\|\nabla f\big(x\big) - \nabla f\big(y\big)\big\|_2 \leq L\big\|x - y\big\|_2.$$

Определение L–гладкости можно задавать и в не евклидовой норме. Обобщение понятия на произвольную норму мы введем в Параграфе Л9.

Теорема Л2.4 (Лемма 1.2.3. из [29])**.** Пусть дана L–гладкая непрерывно дифференцируемая на $\mathbb{R}^d$ функция $f : \mathbb{R}^d \to \mathbb{R}$. Тогда для любых $x, y \in \mathbb{R}^d$ выполнено

$$\big|f\big(y\big) - f\big(x\big) - \big\langle \nabla f\big(x\big), y - x\big\rangle\big| \leq \frac{L}{2}\big\|x - y\big\|_2^2.$$

Доказательство. Используем (см. страницу 84 из [28]) формулу Ньютона-Лейбница для криволинейного интеграла второго рода по кривой, заданной вектор функцией $r(\tau)$:

$$\int\limits_a^b \big\langle \nabla f\big(r(\tau)\big), \mathrm{d}r(\tau)\big\rangle = f\big(r(b)\big) - f\big(r(a)\big).$$

В нашем случае выберем кривую следующим образом $r(\tau) = x + \tau(y - x)$, где $\tau \in [0, 1]$. Тогда

$$\begin{aligned} f\big(y\big) - f\big(x\big) &= \int\limits_0^1 \big\langle \nabla f\big(x + \tau(y - x)\big), y - x\big\rangle \,\mathrm{d}\tau \\ &= \big\langle \nabla f\big(x\big), y - x\big\rangle + \int\limits_0^1 \big\langle \nabla f\big(x + \tau(y - x)\big) - \nabla f\big(x\big), y - x\big\rangle \,\mathrm{d}\tau\,. \end{aligned}$$

Переместив скалярное произведение влево и взяв модуль от обоих частей, получим:

$$\begin{aligned}\left|f(y)-f(x)-\langle\nabla f(x),y-x\rangle\right| &= \left|\int_0^1 \langle\nabla f(x+\tau(y-x))-\nabla f(x),y-x\rangle\,\mathrm{d}\tau\right| \\ &\le \int_0^1 \left|\langle\nabla f(x+\tau(y-x))-\nabla f(x),y-x\rangle\right|\mathrm{d}\tau\,.\end{aligned}$$

В последнем переходе мы использовали факт, что модуль суммы не превосходит сумму моделей слагаемых. Далее воспользуемся неравенством Коши-Буняковского-Шварца (0.3), а затем L-гладкостью (Определение Л2.5):

$$\begin{aligned}\left|f(y)-f(x)-\langle\nabla f(x),y-x\rangle\right| &\le \int_0^1 \left\|\nabla f(x+\tau(y-x))-\nabla f(x)\right\|_2\|y-x\|_2\,\mathrm{d}\tau \\ &\le L\|y-x\|_2^2\int_0^1\tau\,\mathrm{d}\tau = \frac{L}{2}\|x-y\|_2^2.\end{aligned}$$

■

Отметим, что Теорема Л2.4 требует только L-гладкости функции f. Посмотрим, что можно получить, если дополнительно предположить еще и выпуклость функции f.

Теорема Л2.5 (Теорема 2.1.5. из [29])**.** Непрерывно дифференцируемая функция $f:\mathbb{R}^d\to\mathbb{R}$ является выпуклой и L-гладкой тогда и только тогда, когда для любых $x,y\in\mathbb{R}^d$ выполнены следующие неравенства:

$$0\le f(y)-f(x)-\langle\nabla f(x),y-x\rangle\le\frac{L}{2}\|x-y\|_2^2, \tag{Л2.3}$$

$$\frac{1}{2L}\left\|\nabla f(x)-\nabla f(y)\right\|_2^2\le f(y)-f(x)-\langle\nabla f(x),y-x\rangle, \tag{Л2.4}$$

$$\frac{1}{L}\left\|\nabla f(x)-\nabla f(y)\right\|_2^2\le\langle\nabla f(x)-\nabla f(y),x-y\rangle. \tag{Л2.5}$$

Доказательство. Докажем следующее:

$$\begin{cases}\text{выпуклость}\\ \text{гладкость}\end{cases}\implies (\text{Л2.3})\implies(\text{Л2.4})\implies(\text{Л2.5})\implies\begin{cases}\text{выпуклость}\\ \text{гладкость}\end{cases}$$

Выпуклость + гладкость $\implies$ (Л2.3). Первое неравенство есть просто определение выпуклости, а второе является следствием из Теоремы Л2.4.
(Л2.3) $\implies$ (Л2.4). Рассмотрим $\phi(y)=f(y)-\langle\nabla f(x),y\rangle$ для некоторого фиксированного $x\in\mathbb{R}^d$. Удостоверимся, что $\nabla\phi(y)=\nabla f(y)-\nabla f(x)$ является L-гладкой:

$$\begin{aligned}\left\|\nabla\phi(y_1)-\nabla\phi(y_2)\right\|_2 &= \left\|\nabla f(y_1)-\nabla f(x)-\nabla f(y_2)+\nabla f(x)\right\|_2 \\ &= \left\|\nabla f(y_1)-\nabla f(y_2)\right\|_2\le L\|y_1-y_2\|_2.\end{aligned}$$

Проверим также, что $\phi(y)$ выпуклая (по определению). Так как f выпуклая, для произвольных y_1 и y_2 имеем:

$$f(y_1) \geq f(y_2) + \langle \nabla f(y_2), y_1 - y_2 \rangle$$
$$\Updownarrow$$
$$f(y_1) - \langle \nabla f(x), y_1 \rangle \geq f(y_2) - \langle \nabla f(x), y_2 \rangle + \langle \nabla f(y_2) - \nabla f(x), y_1 - y_2 \rangle$$
$$\Updownarrow$$
$$\phi(y_1) \geq \phi(y_2) + \langle \nabla \phi(y_2), y_1 - y_2 \rangle.$$

А это и есть выпуклость $\phi(y)$. Заметим, что $\nabla \phi(x) = 0$, тогда в силу того, что функция ϕ выпуклая, то $y^* = x$ — точка глобального минимума (Теорему Л2.2). Откуда

$$\phi(x) = \phi(y^*) \leq \phi\left(y - \frac{1}{L}\nabla \phi(y)\right). \tag{Л2.6}$$

Теперь применим первый пункт теоремы для $f \to \phi$, $y \to y - \frac{1}{L}\nabla \phi(y)$, $x \to y$:

$$\phi\left(y - \frac{1}{L}\nabla \phi(y)\right) - \phi(y) - \left\langle \nabla \phi(y), -\frac{1}{L}\nabla \phi(y) \right\rangle \leq \frac{1}{2L}\left\|\nabla \phi(y)\right\|_2^2,$$

а значит после небольшой перестановки получим:

$$\phi\left(y - \frac{1}{L}\nabla \phi(y)\right) \leq \phi(y) - \frac{1}{2L}\left\|\nabla \phi(y)\right\|_2^2. \tag{Л2.7}$$

Осталось объединить (Л2.6) и (Л2.7):

$$\phi(x) \leq \phi(y) - \frac{1}{2L}\left\|\nabla \phi(y)\right\|_2^2,$$

и подставить $\phi(y)$:

$$f(x) - \langle \nabla f(x), x \rangle \leq f(y) - \langle \nabla f(x), y \rangle - \frac{1}{2L}\left\|\nabla f(x) - \nabla f(y)\right\|_2^2.$$

Из этого легко получить то, что и хотели доказать

$$\frac{1}{2L}\left\|\nabla f(x) - \nabla f(y)\right\|_2^2 \leq f(y) - f(x) - \langle \nabla f(x), y - x \rangle.$$

(Л2.4) $\implies$ (Л2.5). Запишем два раза (Л2.4):

$$\frac{1}{2L}\left\|\nabla f(x) - \nabla f(y)\right\|_2^2 \leq f(y) - f(x) - \langle \nabla f(x), y - x \rangle,$$
$$\frac{1}{2L}\left\|\nabla f(y) - \nabla f(x)\right\|_2^2 \leq f(x) - f(y) - \langle \nabla f(y), x - y \rangle.$$

Сложим эти два неравенства:

$$\frac{1}{L}\left\|\nabla f(x) - \nabla f(y)\right\|_2^2 \leq \langle \nabla f(x) - \nabla f(y), x - y \rangle.$$

(Л2.5) $\implies$ выпуклость + гладкость. Из (Л2.5) имеем, что:

$$0 \leq \langle \nabla f(x) - \nabla f(y), x - y \rangle.$$

Снова применим (см. страницу 84 из [28]) формулу Ньютона-Лейбница для криволинейного интеграла второго рода по кривой, заданной вектор функцией $r(\tau)$:

$$\int_a^b \langle \nabla f\big(r(\tau)\big), \mathrm{d}r(\tau)\rangle = f\big(r(b)\big) - f\big(r(a)\big).$$

В нашем случае выберем кривую следующим образом $r(\tau) = x + \tau(y-x)$, где $\tau \in [0,1]$:

$$\begin{aligned} f(y) - f(x) &= \int_0^1 \langle \nabla f\big(x + \tau(y-x)\big), y - x\rangle \,\mathrm{d}\tau \\ &= \langle \nabla f(x), y - x\rangle + \int_0^1 \langle \nabla f\big(x + \tau(y-x)\big) - \nabla f(x), y - x\rangle \,\mathrm{d}\tau\,. \end{aligned}$$

Используя, что $\langle \nabla f(x) - \nabla f(x), x - y\rangle \geq 0$, получим:

$$\begin{aligned} f(y) - f(x) &= \langle \nabla f(x), y - x\rangle + \int_0^1 \frac{1}{\tau} \cdot \langle \nabla f(x + \tau(y-x)) - \nabla f(x), \tau(y-x)\rangle \,\mathrm{d}\tau \\ &\geq \langle \nabla f(x), y - x\rangle. \end{aligned}$$

А это и есть эквивалентное определение выпуклости для непрерывно дифференцируемой функции. Также (Л2.5) вместе с неравенством Коши-Буняковского-Шварца (0.3) дает:

$$\begin{aligned} \frac{1}{L}\big\|\nabla f(x) - \nabla f(y)\big\|_2^2 &\leq \langle \nabla f(x) - \nabla f(y), x - y\rangle \\ &\leq \big\|\nabla f(x) - \nabla f(y)\big\|_2 \cdot \big\|x - y\big\|_2. \end{aligned}$$

Откуда

$$\big\|\nabla f(x) - \nabla f(y)\big\|_2 \leq L\big\|x - y\big\|_2,$$

а это и есть Определение Л2.5. ■

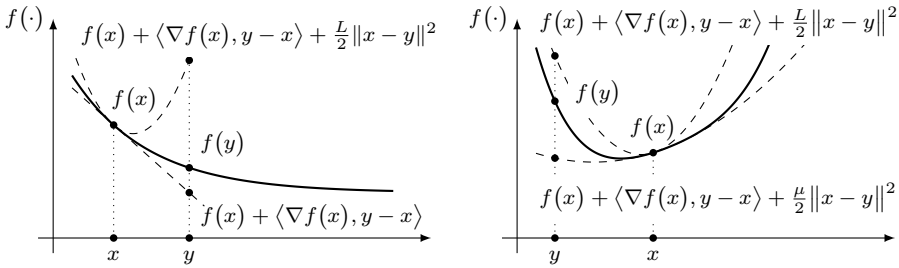
Первое неравенство из Теоремы Л2.5 может быть полезно в понимании физического смысла L-гладкости (Рисунок Л2.2): функция «подперта» снизу линейной аппроксимацией, а сверху квадратичной функцией. Похожая ситуация с L-гладкой и μ-сильно выпуклой функцией. Из Теоремы Л2.5 и Определения Л2.4 легко заметить, что $L \geq \mu$.

Упражнение 12.1. Рассмотрите квадратичную функцию:

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

где $b, x \in \mathbb{R}^d$, $A \in \mathbb{S}_+^d$. Покажите, что константы сильной выпуклости и гладкости можно оценить:

$$\mu \leq \frac{1}{2}\lambda_{\min}\big(A + A^\top\big) \quad L \geq \frac{1}{2}\lambda_{\max}\big(A + A^\top\big).$$



(a) Выпуклость (b) Сильная выпуклость

Рис. Л2.2: Иллюстрация понятий выпуклости и μ-сильной выпуклости.

Упражнение 12.2. Рассмотрите логистическую функцию потерь:

$$f(x) = -\frac{1}{n}\sum_{i=1}^{n} l(g(x, a_i), b_i) + \frac{\lambda}{2}\|x\|_2^2,$$

где $g : \mathbb{R}^d \times \mathbb{R}^d \to \{0, 1\}$ — линейный классификатор:

$$g(x, a_i) = \frac{1}{1 + \exp\left(-\langle x, a_i \rangle\right)},$$

а $l : \mathbb{R}^d \times \mathbb{R}^d \to [0, +\infty]$ — бинарная кросс-энтропия:

$$l(x, y) = y \ln x + (1 - y) \ln(1 - x),$$

$a_i \in \mathbb{R}^d$, $b_i \in \{0, 1\}$, $i = \overline{1, n}$ — данные, $x \in \mathbb{R}^d$ — целевая переменная, $\lambda \in \mathbb{R}_+$ — параметр регуляризации. Покажите, что константы сильной выпуклости и гладкости можно оценить:

$$\mu \leq \frac{1}{n}\lambda_{\min}\left(AA^\top\right) + \lambda \quad L \geq \frac{1}{n}\lambda_{\max}\left(AA^\top\right) + \lambda.$$

Л3 Градиентный спуск

Рассмотрим задачу безусловной оптимизации

$$\min_{x \in \mathbb{R}^d} f(x), \tag{Л3.1}$$

где функция $f(x)$ дифференцируема.

Алгоритм Л3.1 Градиентный спуск

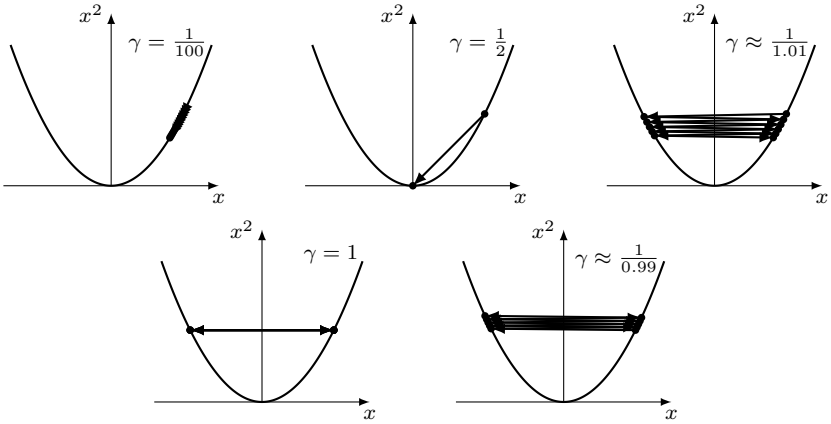
Вход: стартовая точка $x^0 \in \mathbb{R}^d$, размеры шагов $\{\gamma_k\}_{k=0} > 0$, количество итераций K

1: **for** $k = 0, 1, \ldots, K-1$ **do**
2: $\quad x^{k+1} = x^k - \gamma_k \nabla f(x^k)$
3: **end for**

Выход: x^K

Метод был придуман в середине 19го века Огюстеном Коши [4] для решения систем уравнений. Идея метода очень естественна и понятна: градиент показывает направление роста функции, тогда антиградиент — направление убывания. Поэтому давайте пойдем вдоль текущего антиградиента. Сразу же возникает вопрос: а насколько далеко вообще стоит идти по направлению, указанном антиградиентом? Кажется, что не особо далеко, потому что градиент дает только локальную информацию, поэтому совсем непонятно, что же происходит вдали от текущей точки.

Пример Л3.1. Рассмотрим градиентный спуск с постоянным шагом $\gamma_k = \gamma$ для простейшей задачи оптимизации $\min_{x \in \mathbb{R}} x^2$. Константа Липшица градиента $L = 2$. Пусть наш метод стартует из точки $x^0 = 1$. Попробуем различные шаги γ.



1. $\gamma = \frac{1}{100}$: медленная, но монотонная сходимость.
2. $\gamma = \frac{1}{2}$: оптимальный шаг — метод достигает минимума за одну итерацию.
3. $\gamma \approx \frac{1}{1.01}$: наблюдаются сильные колебания.
4. $\gamma = 1$: критический шаг — метод циклично возвращается в исходную точку.

5. $\gamma \approx \frac{1}{0.99}$: метод расходится.

Попробуем построить теорию, доказать сходимость метода градиентного спуска, а также понять допустимые диапазоны шагов γ_k.

Л3.1 Градиентный спуск для гладких сильно выпуклых задач

Отметим сразу, что мы не оформляем дальнейшие рассуждения в виде теоремы, так как они приведут к «топорному» результату. Это пример того, что анализ даже самых простых методов может быть выполнен не лучшим образом. Из Теоремы Л2.3 мы знаем, что для сильно выпуклых функций решение уникально. Попытаемся оценить, как меняется расстояние до него. Используя итерацию градиентного спуска, получаем

$$\begin{aligned}\left\|x^{k+1} - x^*\right\|_2^2 &= \left\|x^k - \gamma_k \nabla f\left(x^k\right) - x^*\right\|_2^2 \\ &= \left\|x^k - x^*\right\|_2^2 - 2\gamma_k \left\langle \nabla f\left(x^k\right), x^k - x^* \right\rangle + \gamma_k^2 \left\|\nabla f\left(x^k\right)\right\|_2^2.\end{aligned}$$

Нужно как-то оценить полученное выражение. Вспоминаем, что у нас есть гладкость (Определение Л2.5):

$$\left\|\nabla f\left(x^k\right) - \nabla f\left(x^*\right)\right\|_2^2 \leq L^2 \left\|x^k - x^*\right\|_2^2$$

и сильная выпуклость (Определение Л2.4):

$$f^* \geq f(x^k) + \left\langle \nabla f(x^k), x^* - x^k \right\rangle + \frac{\mu}{2}\left\|x^k - x^*\right\|_2^2.$$

Чтобы применить гладкость, достаточно вспомнить условие оптимальности $\nabla f\left(x^*\right) = 0$, тогда получаем:

$$\begin{aligned}\left\|x^{k+1} - x^*\right\|_2^2 &= \left\|x^k - x^*\right\|_2^2 - 2\gamma_k \left\langle \nabla f\left(x^k\right), x^k - x^* \right\rangle \\ &\quad + \gamma_k^2 \left\|\nabla f\left(x^k\right) - \nabla f\left(x^*\right)\right\|_2^2 \\ &\leq \left\|x^k - x^*\right\|_2^2 + 2\gamma_k \left(\frac{\mu}{2}\left\|x^k - x^*\right\|_2^2 + f^* - f\left(x^k\right)\right) \\ &\quad + \gamma_k^2 L^2 \left\|x^k - x^*\right\|_2^2 \\ &= \left(1 - \gamma_k \mu + \gamma_k^2 L^2\right)\left\|x^k - x^*\right\|_2^2 + 2\gamma_k \left(f^* - f\left(x^k\right)\right).\end{aligned} \tag{Л3.2}$$

Отбрасываем последнее слагаемое в силу того, что $f(x^k) \geq f^*$:

$$\left\|x^{k+1} - x^*\right\|_2^2 \leq \left(1 - \gamma_k \mu + \gamma_k^2 L^2\right)\left\|x^k - x^*\right\|_2^2.$$

Хочется, чтобы расстояние до решения уменьшалось, а именно, $\left(1 - \gamma_k \mu + \gamma_k^2 L^2\right) < 1$, а лучше вообще найти $\operatorname{argmin}_{\gamma_k}\left(1 - \gamma_k \mu + \gamma_k^2 L^2\right)$. Это несложно — минимизиурем одномерную параболу по γ_k и получаем: $\gamma_k = \frac{\mu}{2L^2}$ и

$$1 - \gamma_k \mu + \gamma_k^2 L^2 = 1 - \frac{\mu^2}{4L^2}.$$

А значит,

$$\left\|x^{k+1} - x^*\right\|_2^2 \leq \left(1 - \frac{\mu^2}{4L^2}\right)\left\|x^k - x^*\right\|_2^2.$$

Это неравенство выполняется для всех $k = \overline{0, K-1}$. Подставим $k = K-1$:

$$\left\|x^K - x^*\right\|_2^2 \leq \left(1 - \frac{\mu^2}{4L^2}\right)\left\|x^{K-1} - x^*\right\|_2^2.$$

Теперь продолжим неравенство, подставляя $k = K-2$:

$$\left\|x^K - x^*\right\|_2^2 \leq \left(1 - \frac{\mu^2}{4L^2}\right)\left(1 - \frac{\mu^2}{4L^2}\right)\left\|x^{K-2} - x^*\right\|_2^2 = \left(1 - \frac{\mu^2}{4L^2}\right)^2\left\|x^{K-2} - x^*\right\|_2^2.$$

И так можно продолжать, вплоть до $k = 0$. Запустим рекурсию:

$$\begin{aligned}\left\|x^K - x^*\right\|_2^2 &\leq \left(1 - \frac{\mu^2}{4L^2}\right)\left\|x^{K-1} - x^*\right\|_2^2 \leq \left(1 - \frac{\mu^2}{4L^2}\right)^2\left\|x^{K-2} - x^*\right\|_2^2 \\ &\leq \left(1 - \frac{\mu^2}{4L^2}\right)^3\left\|x^{K-3} - x^*\right\|_2^2 \leq \ldots \leq \left(1 - \frac{\mu^2}{4L^2}\right)^K\left\|x^0 - x^*\right\|_2^2.\end{aligned}$$

Получили линейную сходимость (Определение Л1.5). Попробуем найти оценку на число итераций. Для дальнейшего анализа часто применяют неравенство (0.4)

$$(1-\alpha)^k \leq \exp(-\alpha k),$$

где $\alpha \in [0, 1]$, $k \in \mathbb{N}$. Его можно получить, записав определение выпуклости $\exp(-x)$ в точке 0:

$$\exp(-x) \geq 1 - x,$$

после чего возвести неравенство в степень k, поскольку обе части неравенства неотрицательны.
Заметим, что $\mu \leq L$ (Теорема Л2.5 и Определение Л2.4), а значит

$$\left\|x^K - x^*\right\|_2^2 \leq \left(1 - \frac{\mu^2}{4L^2}\right)^K\left\|x^0 - x^*\right\|_2^2 \leq \exp\left(-\frac{\mu^2}{4L^2} \cdot K\right)\left\|x^0 - x^*\right\|_2^2.$$

Мы хотим гарантировать, что

$$\left\|x^K - x^*\right\|_2^2 \leq \exp\left(-\frac{\mu^2}{4L^2} \cdot K\right)\left\|x^0 - x^*\right\|_2^2 \leq \varepsilon^2.$$

Тогда логарифмируем и получаем

$$K \geq \frac{4L^2}{\mu^2}\log\left(\frac{\left\|x^0 - x^*\right\|_2^2}{\varepsilon^2}\right).$$

Но этот результат далеко не самый лучший. Анализ можно провести «тоньше».

Теорема Л3.1. Пусть задача безусловной оптимизации (Л3.1) с L-гладкой, μ-сильно выпуклой целевой функцией f решается с помощью градиентного спуска (Алгоритм Л3.1). Тогда при $\gamma_k \leq \frac{1}{L}$ справедлива следующая оценка сходимости:

$$\left\|x^K - x^*\right\|_2^2 \leq \left(\prod_{k=0}^{K-1}(1 - \mu\gamma_k)\right)\left\|x^0 - x^*\right\|_2^2. \tag{Л3.3}$$

Доказательство. Начинаем с применения итерации градиентного спуска и условия оптимальности $\nabla f(x^*) = 0$:

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

Для точек x^k и x^* воспользуемся сильной выпуклостью (Определение Л2.4):

$$f^* \geq f(x^k) + \langle \nabla f(x^k), x^* - x^k \rangle + \frac{\mu}{2}\|x^k - x^*\|_2^2$$

и гладкостью (Теорема Л2.5):

$$\frac{1}{2L}\|\nabla f(x^k) - \nabla f(x^*)\|_2^2 \leq f(x^k) - f^* - \langle \nabla f(x^*), x^k - x^* \rangle.$$

Тогда после подстановки неравенств, имеем:

$$\begin{aligned}\|x^{k+1} - x^*\|_2^2 &\leq \|x^k - x^*\|_2^2 - 2\gamma_k \Big(\frac{\mu}{2}\|x^k - x^*\|_2^2 + f(x^k) - f^*\Big) \\ &\quad + \gamma_k^2 \|\nabla f(x^k) - \nabla f(x^*)\|_2^2 \\ &\leq \|x^k - x^*\|_2^2 - 2\gamma_k \Big(\frac{\mu}{2}\|x^k - x^*\|_2^2 + f(x^k) - f^*\Big) \\ &\quad + 2\gamma_k^2 L\big(f(x^k) - f^* - \langle \nabla f(x^*), x^k - x^* \rangle\big) \\ &= \big(1 - \mu\gamma_k\big)\|x^k - x^*\|_2^2 + 2\gamma_k(\gamma_k L - 1)\big(f(x^k) - f^*\big).\end{aligned} \tag{Л3.4}$$

Применим, что $\gamma_k \leq \frac{1}{L}$, и откинем отрицательное слагаемое:

$$\|x^{k+1} - x^*\|_2^2 \leq \big(1 - \mu\gamma_k\big)\|x^k - x^*\|_2^2.$$

Запустим рекурсию:

$$\begin{aligned}\|x^K - x^*\|_2^2 &\leq \big(1 - \mu\gamma_{K-1}\big)\|x^{K-1} - x^*\|_2^2 \\ &\leq \big(1 - \mu\gamma_{K-1}\big)\big(1 - \mu\gamma_{K-2}\big)\|x^{K-2} - x^*\|_2^2 \\ &\leq \ldots \leq \left(\prod_{k=0}^{K-1}\big(1 - \mu\gamma_k\big)\right)\|x^0 - x^*\|_2^2.\end{aligned}$$

■

Проведем анализ полученной оценки. Исходя из неравенства (Л3.3), можем заключить, что шаг γ_k следует брать максимально допустимым. В условии теоремы было наложено ограничение $\gamma_k \leq \frac{1}{L}$, поэтому, сформулируем утверждение о скорости сходимости при $\gamma_k = \frac{1}{L}$.

Утверждение Л3.1. Пусть задача удовлетворяет условиям Теоремы Л3.1 и выбрано значение шага $\gamma_k = \frac{1}{L}$. Тогда справедлива следующая оценка сходимости:

$$\|x^K - x^*\|_2^2 \leq \Big(1 - \frac{\mu}{L}\Big)^K \|x^0 - x^*\|_2^2.$$

Более того, чтобы добиться точности ε по аргументу $\left(\left\|x^K - x^*\right\|_2 \leq \varepsilon\right)$, необходимо

$$K = \mathcal{O}\left(\frac{L}{\mu}\log\frac{\left\|x^0 - x^*\right\|_2}{\varepsilon}\right) = \tilde{\mathcal{O}}\left(\frac{L}{\mu}\right) \text{ итераций.}$$

Доказательство. Запишем оценку сходимости из Теоремы Л3.1:

$$\left\|x^K - x^*\right\|_2^2 \leq \left(\prod_{k=0}^{K-1}\left(1 - \mu\gamma_k\right)\right)\left\|x^0 - x^*\right\|_2^2.$$

При $\gamma_k = \frac{1}{L}$ получаем:

$$\left\|x^K - x^*\right\|_2^2 \leq \left(1 - \frac{\mu}{L}\right)^K\left\|x^0 - x^*\right\|_2^2.$$

Получили линейную сходимость (Определение Л1.5). Найдем оценку на число итераций. Заметим, что $\mu \leq L$ (Теорема Л2.5 и Определение Л2.4), а значит

$$\left\|x^K - x^*\right\|_2^2 \leq \left(1 - \frac{\mu}{L}\right)^K\left\|x^0 - x^*\right\|_2^2 \leq \exp\left(-\frac{\mu}{L}\cdot K\right)\left\|x^0 - x^*\right\|_2^2.$$

Мы хотим гарантировать, что

$$\left\|x^K - x^*\right\|_2^2 \leq \exp\left(-\frac{\mu}{L}\cdot K\right)\left\|x^0 - x^*\right\|_2^2 \leq \varepsilon^2.$$

Тогда логарифмируем и получаем

$$K \geq \frac{L}{\mu}\log\left(\frac{\left\|x^0 - x^*\right\|_2^2}{\varepsilon^2}\right).$$

■

Замечание Л3.1. Мы будем использовать $\mathcal{O}$-нотацию, чтобы «убирать» численные факторы и $\tilde{\mathcal{O}}$-нотацию, чтобы «убирать» еще и log-факторы (более формально все факторы вида $\log(\text{poly}(L, \mu, \varepsilon, \ldots))$, где $\text{poly}(L, \mu, \varepsilon, \ldots)$ — полиномы).

Но и этот анализ можно еще немного улучшить. Для этого понадобится следующее утверждение.

Утверждение Л3.2. Пусть непрерывно дифференцируемая функция $f : \mathbb{R}^d \to \mathbb{R}$ является μ-сильно выпуклой и L-гладкой, тогда для любых $x, y \in \mathbb{R}^d$ выполнено следующее неравенство:

$$\left\langle \nabla f(x) - \nabla f(y), x - y\right\rangle \geq \frac{\mu L}{\mu + L}\left\|x - y\right\|_2^2 + \frac{1}{\mu + L}\left\|\nabla f(x) - \nabla f(y)\right\|_2^2.$$

Доказательство. Рассмотрим функцию $\phi(x) = f(x) - \frac{\mu}{2}\left\|x\right\|_2^2$. Докажем, что эта функция является выпуклой и $(L-\mu)$-гладкой. Выпуклость проверим по Определению Л2.3:

$$\begin{aligned}
\phi(y) - \phi(x) - \left\langle \nabla\phi(x), y - x\right\rangle &= f(y) - \frac{\mu}{2}\|y\|_2^2 - f(x) + \frac{\mu}{2}\|x\|_2^2 - \left\langle \nabla f(x) - \mu x, y - x\right\rangle \\
&= f(y) - f(x) - \left\langle \nabla f(x), y - x\right\rangle \\
&\quad + \frac{\mu}{2}\left(\left\|x\right\|_2^2 - \left\|y\right\|_2^2 + 2\langle x, y - x\rangle\right) \\
&= f(y) - f(x) - \left\langle \nabla f(x), y - x\right\rangle - \frac{\mu}{2}\left\|y - x\right\|_2^2 \geq 0.
\end{aligned}$$

Последнее в стиле сильной выпуклости f (Определение Л2.4).
Для нахождения константы гладкости воспользуемся L-гладкостью f и неравенством (Л2.3) из Теоремы Л2.5:

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

Тогда, комбинируя два неравенства выше, получаем для ϕ:

$$\phi(y) - \phi(x) - \langle \nabla \phi(x), y - x \rangle \leq \frac{L - \mu}{2}\|x - y\|_2^2.$$

Теперь для выпуклой и $(L - \mu)$-гладкой функции воспользуемся неравенством (Л2.5) из Теоремы Л2.5:

$$(L - \mu)\langle \nabla \phi(x) - \nabla \phi(y), x - y \rangle \geq \|\nabla \phi(x) - \nabla \phi(y)\|_2^2.$$

Преобразуем обе части неравенства, подставив $\phi(x) = f(x) - \frac{\mu}{2}\|x\|_2^2$. Для левой части получим:

$$\begin{aligned}(L - \mu)\langle \nabla \phi(x) - \nabla \phi(y), x - y \rangle &= (L - \mu)\langle \nabla f(x) - \mu x - \nabla f(y) - \mu y, x - y \rangle \\ &= (L - \mu)\langle \nabla f(x) - \nabla f(y), x - y \rangle - (L - \mu)\mu\|x - y\|_2^2,\end{aligned}$$

Теперь для правой части неравенства:

$$\begin{aligned}\|\nabla \phi(x) - \nabla \phi(y)\|_2^2 &= \|\nabla f(x) - \mu x - \nabla f(y) - \mu y\| \\ &= \|\nabla f(x) - \nabla f(y)\|_2^2 - 2\mu\langle \nabla f(x) - \nabla f(y), x - y \rangle + \mu^2\|x - y\|_2^2.\end{aligned}$$

Подставляем обе части:

$$\begin{aligned}&(L - \mu)\langle \nabla f(x) - \nabla f(y), x - y \rangle - (L - \mu)\mu\|x - y\|_2^2 \\ &\quad \geq \|\nabla f(x) - \nabla f(y)\|_2^2 - 2\mu\langle \nabla f(x) - \nabla f(y), x - y \rangle + \mu^2\|x - y\|_2^2.\end{aligned}$$

После небольшого упрощения:

$$(L + \mu)\langle \nabla f(x) - \nabla f(y), x - y \rangle - L\mu\|x - y\|_2^2 \geq \|\nabla f(x) - \nabla f(y)\|_2^2.$$

Приведём к требуемому виду:

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

■

Используя доказанное утверждение, можно получить следующий результат.

Теорема Л3.2. [Теорема 2.1.15. из [29]] Пусть задача безусловной оптимизации (Л3.1) с L-гладкой, μ-сильно выпуклой целевой функцией f решается с помощью градиентного спуска (Алгоритм Л3.1). Тогда при $\gamma_k \leq \frac{2}{\mu + L}$ справедлива следующая оценка сходимости:

$$\|x^K - x^*\|_2^2 \leq \left(\prod_{k=0}^{K-1}\left(1 - \frac{2\gamma_k \mu L}{\mu + L}\right)\right)\|x^0 - x^*\|_2^2.$$

Доказательство. Пойдем по уже известному пути (Л3.2):

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

Теперь используем результат Утверждения Л3.2, чтобы оценить скалярное произведение:

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

Выбрав $\gamma_k \le \frac{2}{\mu+L}$, имеем

$$\left\|x^{k+1}-x^*\right\|_2^2 \le \left(1-\frac{2\gamma_k\mu L}{\mu+L}\right)\left\|x^k-x^*\right\|_2^2.$$

Запустив рекурсию, получаем

$$\left\|x^K-x^*\right\|_2^2 \le \left(\prod_{k=0}^{K-1}\left(1-\frac{2\gamma_k\mu L}{\mu+L}\right)\right)\left\|x^0-x^*\right\|_2^2.$$

■

Как и в случае с предыдущей оценкой сходимости с произвольным достаточно малым шагом, сформулируем утверждение с теоретически максимально допустимым значением шага $\gamma_k = \frac{2}{\mu+L}$.

Утверждение Л3.3. Пусть задача удовлетворяет условиям Теоремы Л3.2 и выбрано значение шага $\gamma_k = \frac{2}{\mu+L}$. Тогда справедлива следующая оценка сходимости:

$$\left\|x^K-x^*\right\|_2^2 \le \left(1-\frac{4\mu L}{(\mu+L)^2}\right)^K\left\|x^0-x^*\right\|_2^2.$$

Более того, чтобы добиться точности ε по аргументу $\left(\left\|x^k-x^*\right\|_2 \le \varepsilon\right)$, необходимо

$$K = \mathcal{O}\left(\frac{L}{\mu}\log\frac{\left\|x^0-x^*\right\|_2}{\varepsilon}\right) = \tilde{\mathcal{O}}\left(\frac{L}{\mu}\right) \text{ итераций.}$$

Доказательство. Запишем оценку сходимости из Теоремы Л3.2:

$$\left\|x^K-x^*\right\|_2^2 \le \left(\prod_{k=0}^{K-1}\left(1-\frac{2\gamma_k\mu L}{\mu+L}\right)\right)\left\|x^0-x^*\right\|_2^2.$$

Возьмем постоянный шаг $\gamma_k = \frac{2}{\mu+L}$ и применим неравенство $(1-\alpha)^x \le \exp(-\alpha x)$:

$$\begin{aligned}\left\|x^K - x^*\right\|_2^2 &\le \left(1 - \frac{4\mu L}{(\mu+L)^2}\right)^K \left\|x^0 - x^*\right\|_2^2 \\ &\le \exp\left(-\frac{4\mu L}{(\mu+L)^2} K\right) \left\|x^0 - x^*\right\|_2^2.\end{aligned}$$

Хотим потребовать точность ε по аргументу ($\left\|x^K - x^*\right\|_2 \le \varepsilon$), для этого оцениваем все сверху:

$$\left\|x^K - x^*\right\|_2^2 \le \exp\left(-\frac{4\mu L}{(\mu+L)^2} K\right) \left\|x^0 - x^*\right\|_2^2 \le \varepsilon^2.$$

Откуда

$$K \ge \frac{(L+\mu)^2}{4\mu L} \log\left(\frac{\left\|x^0 - x^*\right\|_2^2}{\varepsilon^2}\right).$$

■

Сравним результаты Теорем Л3.1 и Л3.2. С точки зрения $\mathcal{O}$-оценок, результаты одинаковы, но можно сравнить $\left(1 - \frac{\mu}{L}\right)$ и $\left(1 - \frac{4\mu L}{(\mu+L)^2}\right)$. Легко заметить, что

$$\left(1 - \frac{4\mu L}{(\mu+L)^2}\right) < \left(1 - \frac{\mu}{L}\right)$$

при $\mu \ne L$, а значит, результаты Теоремы Л3.2 лучше. Однако, более классическим является первый результат. Отчасти это из-за того, что оптимальное значение шага для той оценки $\gamma_k = \frac{1}{L}$ не требует от нас знания константы сильной выпуклости.

Л3.2 Градиентный спуск для гладких выпуклых задач

В данном разделе x^* — некоторая (необязательно единственная, но существующая) точка глобального минимума.
Докажем сперва вспомогательное утверждение о том, что значение функции убывает от итерации к итерации.

Утверждение Л3.4. Пусть задача безусловной оптимизации (Л3.1) с L-гладкой, выпуклой целевой функцией f решается с помощью градиентного спуска (Алгоритм Л3.1). Тогда при $\gamma_k \in \left(0, \frac{2}{L}\right)$ справедливо, что $f(x^{k+1}) \le f(x^k)$.

Доказательство. Воспользуемся L-гладкостью (Теорема Л2.5):

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

Подставим итерационный шаг градиентного спуска:

$$\begin{aligned}f(x^{k+1}) &\le f(x^k) - \gamma_k \left\|\nabla f(x^k)\right\|_2^2 + \frac{L\gamma_k^2}{2}\left\|\nabla f(x^k)\right\|_2^2 \\ &= f(x^k) - \gamma_k\left(1 - \frac{L\gamma_k}{2}\right)\left\|\nabla f(x^k)\right\|_2^2.\end{aligned} \tag{Л3.5}$$

Если $\gamma_k \in \left(0, \frac{2}{L}\right)$, то $f(x^{k+1}) \le f(x^k)$. ■

Это предложение не просто является вспомогательным фактом, оно отражает и физику градиентного спуска. Шаг метода подбирается исходя из верхней ограничивающей квадратичной функции/параболы, возникающей из-за гладкости. Как и в Примере Л3.1, нам нельзя делать большие шаги, иначе мы перескочим на противоположную ветвь параболы относительно оптимума и можем оказаться выше той точки, где находились до этого. Поэтому и подбирается шаг «осторожно», исходя из худшего варианта скорости/резкости роста функции (верхней оценки). При этом можно заметить, что если $\gamma_k = \frac{1}{L}$, то мы минимизируем выражение

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

т.е. каждый раз мы минимизируем локальную верхнюю ограничивающую параболу на поведение нашей функции.

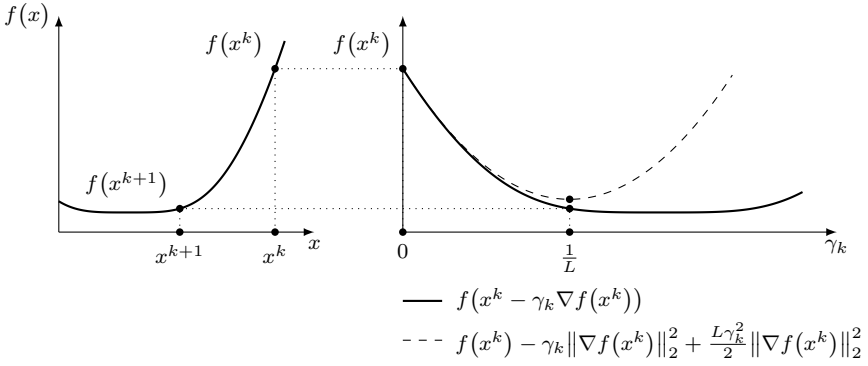


Рис. Л3.1: Иллюстрация шага градиентного спуска с оптимальным $\gamma_k = \frac{1}{L}$.

Теорема Л3.3. Пусть задача безусловной оптимизации (Л3.1) с L-гладкой, выпуклой целевой функцией f решается с помощью градиентного спуска (Алгоритм Л3.1). Тогда при $\gamma_k \le \frac{1}{2L}$ справедлива следующая оценка сходимости:

$$f(x^K) - f^* \le \frac{\|x^0 - x^*\|_2^2}{\sum\limits_{k=0}^{K-1} \gamma_k}.$$

В данной теореме есть сходимость только по функции, то есть, решение может быть не уникально (Пример Л2.1). А также предполагается существование решения (для конечности $\|x^0 - x^*\|_2^2$).

Доказательство. Начнем доказательство с оценки (Л3.4) с $\mu = 0$:

$$\|x^{k+1} - x^*\|_2^2 \le \|x^k - x^*\|_2^2 + 2\gamma_k(\gamma_k L - 1)(f(x^k) - f^*).$$

Учтем, что $\gamma_k L - 1 \le -\frac{1}{2}$ из-за ограничения на размер шага $\gamma_k \le \frac{1}{2L}$:

$$\begin{aligned}\|x^{k+1} - x^*\|_2^2 &\le \|x^k - x^*\|_2^2 + 2\gamma_k(\gamma_k L - 1)(f(x^k) - f^*)\\ &\le \|x^k - x^*\|_2^2 - \gamma_k(f(x^k) - f^*).\end{aligned}$$

Перетасуем неравенство:

$$\gamma_k\big(f(x^k) - f^*\big) \le \big\|x^k - x^*\big\|_2^2 - \big\|x^{k+1} - x^*\big\|_2^2.$$

Суммируя по всем итерациям от 0 до $K-1$:

$$\begin{aligned}\sum_{k=0}^{K-1} \gamma_k\big(f(x^k) - f^*\big) &\le \sum_{k=0}^{K-1} \Big(\big\|x^k - x^*\big\|_2^2 - \big\|x^{k+1} - x^*\big\|_2^2\Big) \\ &= \big\|x^0 - x^*\big\|_2^2 - \big\|x^K - x^*\big\|_2^2 \le \big\|x^0 - x^*\big\|_2^2.\end{aligned}$$

Воспользуемся Утверждением Л3.4 и заметим, что $f\big(x^K\big) \le f\big(x^k\big),\ k = \overline{0, K}$:

$$\big(f\big(x^K\big) - f^*\big) \sum_{k=0}^{K-1} \gamma_k \le \big\|x^0 - x^*\big\|_2^2.$$

Откуда получаем, что

$$f\big(x^K\big) - f^* \le \frac{\big\|x^0 - x^*\big\|_2^2}{\sum\limits_{k=0}^{K-1} \gamma_k}.$$

■

Утверждение Л3.5. Пусть задача удовлетворяет условиям Теоремы Л3.3 и выбрано значение шага $\gamma_k = \frac{1}{2L}$. Тогда справедлива следующая оценка сходимости:

$$f\big(x^K\big) - f^* \le \frac{2L\big\|x^0 - x^*\big\|_2^2}{K}.$$

Более того, чтобы добиться точности ε по функции $\big(f\big(x^K\big) - f^* \le \varepsilon\big)$, необходимо

$$K = \mathcal{O}\left(\frac{L\big\|x^0 - x^*\big\|_2^2}{\varepsilon}\right) \text{ итераций.}$$

Доказательство. Начнем с оценки из Теоремы Л3.3:

$$f\big(x^K\big) - f^* \le \frac{\big\|x^0 - x^*\big\|_2^2}{\sum\limits_{k=0}^{K-1} \gamma_k}.$$

Подставим фиксированный шаг $\gamma_k = \frac{1}{2L}$, получим требуемое:

$$f\big(x^K\big) - f^* \le \frac{\big\|x^0 - x^*\big\|_2^2}{\sum\limits_{k=0}^{K-1} \gamma_k} = \frac{2L\big\|x^0 - x^*\big\|_2^2}{K}.$$

Это сублинейная сходимость (Определение Л1.5). Чтобы получить оценку на число итераций, отметим, что мы хотим гарантировать:

$$f\big(x^K\big) - f^* \le \frac{2L\big\|x^0 - x^*\big\|_2^2}{K} \le \varepsilon.$$

Тогда получаем

$$K \geq \frac{2L\|x^0 - x^*\|_2^2}{\varepsilon}.$$

■

Как и в сильно выпуклом случае, данный анализ можно немного улучшить.

Теорема Л3.4. Пусть задача безусловной оптимизации (Л3.1) с L-гладкой, выпуклой целевой функцией f решается с помощью градиентного спуска (Алгоритм Л3.1). Тогда при $\gamma_k \in \left[0, \frac{1}{L}\right]$ справедлива следующая оценка сходимости:

$$f(x^K) - f(x^*) \leq \frac{\|x^0 - x^*\|_2^2}{2\sum\limits_{k=0}^{K-1} \gamma_k}.$$

Доказательство. Начинаем уже привычным образом:

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

Воспользуемся выпуклостью: $f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle$ (Определение Л2.3). Тогда с учетом $\nabla f(x^*) = 0$ имеем

$$\|x^{k+1} - x^*\|_2^2 \leq \|x^k - x^*\|_2^2 - 2\gamma_k (f(x^k) - f^*) + \gamma_k^2 \|\nabla f(x^k)\|_2^2. \tag{Л3.6}$$

Далее применим результат (Л3.5), а именно оценим $\|\nabla f(x^k)\|_2^2$:

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

Здесь учитывается, что $\gamma_k \in \left(0, \frac{2}{L}\right)$. Объединяя два предыдущих выражения, получаем

$$\|x^{k+1} - x^*\|_2^2 \leq \|x^k - x^*\|_2^2 - 2\gamma_k (f(x^k) - f^*) + \frac{2\gamma_k}{2 - L\gamma_k} (f(x^k) - f(x^{k+1})).$$

Продолжим цепочку неравенств, взяв во внимание $f^* \leq f(x^{k+1}) \leq f(x^k)$ (Утверждение Л3.4):

$$\begin{aligned}\|x^{k+1} - x^*\|_2^2 &\leq \|x^k - x^*\|_2^2 - 2\gamma_k (f(x^k) - f^*) + \frac{2\gamma_k}{2 - L\gamma_k} (f(x^k) - f(x^{k+1})) \\ &= \|x^k - x^*\|_2^2 - 2\gamma_k \left(1 - \frac{1}{2 - L\gamma_k}\right) f(x^k) - \frac{2\gamma_k}{2 - L\gamma_k} f(x^{k+1}) + 2\gamma_k f^* \\ &\leq \|x^k - x^*\|_2^2 - 2\gamma_k \left(1 - \frac{1}{2 - L\gamma_k}\right) f(x^{k+1}) - \frac{2\gamma_k}{2 - L\gamma_k} f(x^{k+1}) + 2\gamma_k f^* \\ &= \|x^k - x^*\|_2^2 - 2\gamma_k (f(x^{k+1}) - f^*).\end{aligned}$$

После небольшой перестановки:

$$2\gamma_k (f(x^{k+1}) - f^*) \leq \|x^k - x^*\|_2^2 - \|x^{k+1} - x^*\|_2^2.$$

Суммируем по всем k от 0 до K, имеем:

$$2\sum_{k=0}^{K-1}\gamma_k\big(f(x^{k+1})-f^*\big)\le\sum_{k=0}^{K-1}\Big(\big\|x^k-x^*\big\|_2^2-\big\|x^{k+1}-x^*\big\|_2^2\Big)$$
$$=\big\|x^0-x^*\big\|_2^2-\big\|x^K-x^*\big\|_2^2\le\big\|x^0-x^*\big\|_2^2.$$

Из Утверждения Л3.4 следует, что $f(x^K)\le f(x^{K-1})\le\ldots\le f(x^0)$. Воспользуемся этим в оценке:

$$2\big(f(x^K)-f^*\big)\sum_{k=0}^{K-1}\gamma_k\le\big\|x^0-x^*\big\|_2^2.$$

Откуда получаем, что

$$f(x^K)-f^*\le\frac{\big\|x^0-x^*\big\|_2^2}{2\sum\limits_{k=0}^{K-1}\gamma_k}.$$

■

Подберем оптимальное значение шага. Как видно из оценки, шаг стоит выбирать максимальным из возможных. Сформулируем отдельное утверждение для $\gamma_k=\frac{1}{L}$.

Теорема Л3.5. Пусть задача удовлетворяет условиям Теоремы Л3.4 и выбрано значение шага $\gamma_k=\frac{1}{L}$. Тогда справедлива следующая оценка сходимости:

$$f(x^K)-f^*\le\frac{L\big\|x^0-x^*\big\|_2^2}{2K}.$$

Более того, чтобы добиться точности ε по функции $\big(f(x^K)-f^*\le\varepsilon\big)$, необходимо

$$K=\mathcal{O}\left(\frac{L\big\|x^0-x^*\big\|_2^2}{\varepsilon}\right)\text{ итераций.}$$

Доказательство. Выпишем оценку из Теоремы Л3.4 и подставим $\gamma_k=\frac{1}{L}$:

$$f(x^K)-f^*\le\frac{\big\|x^0-x^*\big\|_2^2}{2\sum\limits_{k=0}^{K-1}\gamma_k}\le\frac{L\big\|x^0-x^*\big\|_2^2}{2K}.$$

Чтобы получить оценку на число итераций, отметим, что мы хотим гарантировать:

$$f(x^K)-f^*\le\frac{L\big\|x^0-x^*\big\|_2^2}{2K}\le\varepsilon.$$

Тогда

$$K\ge\frac{L\big\|x^0-x^*\big\|_2^2}{2\varepsilon}.$$

■

Сравним результаты Теорем Л3.3 и Л3.4. С точки зрения $\mathcal{O}$-оценок, результаты одинаковы, но численный фактор в Теореме Л3.4 лучше.

Л3.3 Градиентный спуск для гладких невыпуклых задач

Теорема Л3.6. Пусть задача безусловной оптимизации (Л3.1) с L-гладкой целевой функцией f решается с помощью градиентного спуска (Алгоритм Л3.1). Тогда при $\gamma_k = \frac{1}{L}$ справедлива следующая оценка сходимости:

$$\left\|\nabla f(\hat{x}^K)\right\|_2^2 \leq \frac{2L(f(x^0) - f^*)}{K},$$

где $\hat{x}^K$ — некоторая специально выбранная точка. Более того, чтобы добиться точности ε по норме градиента $\left(\left\|\nabla f(x^K)\right\|_2 \leq \varepsilon\right)$, необходимо

$$K = \mathcal{O}\left(\frac{2L(f(x^0) - f^*)}{\varepsilon^2}\right) \text{ итераций.}$$

Здесь f^* — глобальный минимум функции.

В данной теореме есть сходимость только к стационарной точке (для невыпуклых функций нет гарантий, что это экстремум), и предполагается существование глобального минимума f^*.

Доказательство. Воспользуемся L-гладкостью (Теорема Л2.4):

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

Подставим итерационный шаг градиентного спуска:

$$\begin{aligned} f(x^{k+1}) &\leq f(x^k) - \gamma_k \left\|\nabla f(x^k)\right\|_2^2 + \frac{L\gamma_k^2}{2}\left\|\nabla f(x^k)\right\|_2^2 \\ &\leq f(x^k) - \gamma_k\left(1 - \frac{L\gamma_k}{2}\right)\left\|\nabla f(x^k)\right\|_2^2. \end{aligned}$$

Возьмем $\gamma_k = \frac{1}{L}$. Тогда

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

или после небольших перестановок:

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

Суммируя по всем итерациям от 0 до $K-1$, получаем:

$$\sum_{k=0}^{K-1}\left\|\nabla f(x^k)\right\|_2^2 \leq 2L(f(x^0) - f(x^K)) \leq 2L(f(x^0) - f^*).$$

Здесь мы еще воспользовались существованием глобального минимума f^*. Усредняя по K:

$$\frac{1}{K}\sum_{k=0}^{K-1}\left\|\nabla f(x^k)\right\|_2^2 \leq \frac{2L(f(x^0) - f^*)}{K}.$$

Из левой части непонятно, для какой точки мы получили гарантии сходимости. С такой мы проблемой еще не сталкивались, есть несколько способов ее решить:

- Случайный выбор $\hat{x}^K$ из $x_0, \ldots, x_{K-1}$:

$$\mathbb{E}\left[\left\|\nabla f\left(\hat{x}^K\right)\right\|_2^2\right] = \frac{1}{K}\sum_{k=0}^{K-1}\left\|\nabla f\left(x^k\right)\right\|_2^2 \leq \frac{2L\left(f\left(x^0\right) - f^*\right)}{K};$$

- Выбор точки с минимальным по норме градиентом $\hat{x}^K = \underset{k\in 0,K-1}{\operatorname{argmin}}\left\|\nabla f\left(x^k\right)\right\|_2^2$:

$$\left\|\nabla f\left(\hat{x}^K\right)\right\|_2^2 = \min_{k\in 0,K-1}\left\|\nabla f\left(x^k\right)\right\|_2^2 \leq \frac{1}{K}\sum_{k=0}^{K-1}\left\|\nabla f\left(x^k\right)\right\|_2^2 \leq \frac{2L\left(f\left(x^0\right) - f^*\right)}{K}.$$

Это сублинейная сходимость (Определение Л1.5). Чтобы получить оценку на число итераций, потребуем

$$\left\|\nabla f\left(\hat{x}^K\right)\right\|_2^2 \leq \frac{2L\left(f\left(x^0\right) - f^*\right)}{K} \leq \varepsilon^2.$$

Тогда получаем

$$K \geq \frac{2L\left(f\left(x^0\right) - f^*\right)}{\varepsilon^2}.$$

■

Л3.4 Выбор шага градиентного спуска

До этого момента мы доказывали сходимость градиентного спуска для L-гладких функций при достаточно малом шаге γ_k, а затем формулировали утверждения, в которых фиксировали максимально допустимый шаг. Однако, в реальных задачах мы не всегда знаем, как подбирать шаг. Во-первых, это может происходить, поскольку нам может не быть известно, является ли функция L-гладкой, но это не означает, что градиентный спуск не будет работать. Во-вторых, нам может быть неизвестна константа липшицевости градиента L, от которой зависит теоретическое значение оптимального размера шага. Начнем с метода, который опирается на L-гладкость функции.

Л3.4.1 Линейный поиск значения L

Обратимся к результатам Утверждения Л3.4. В ходе доказательства мы получили оценку (Л3.5) на значение функции f на следующей итерации:

$$f\left(x^{k+1}\right) \leq f\left(x^k\right) + \underbrace{\left(\frac{L\gamma^2}{2} - \gamma\right)}_{\to\min}\left\|\nabla f\left(x^k\right)\right\|_2^2 = \left[\gamma = \frac{1}{L}\right] = f\left(x^k\right) - \frac{1}{2L}\left\|\nabla f\left(x^k\right)\right\|_2^2.$$

Из оценки выше сразу же появляется идея для линейного поиска L: увеличиваем L в ρ раз, пока условие не начинает выполняться.

Алгоритм Л3.2 Линейный поиск значения L

Вход: стартовая точка $x^k \in \mathbb{R}^d$, начальное значение $L > 0$, параметр $\rho > 1$
$x^{k+1} = x^k - \frac{1}{L}\nabla f\left(x^k\right)$
while $f\left(x^{k+1}\right) > f\left(x^k\right) - \frac{1}{2L}\left\|\nabla f\left(x^k\right)\right\|_2^2$ **do**
    $L = \rho \cdot L$
    $x^{k+1} = x^k - \frac{1}{L}\nabla f\left(x^k\right)$
end while
Выход: L

Полученный алгоритм можно встроить в метод градиентного спуска (Алгоритм Л3.1), добавив запуск его на каждой итерации. Более того, в процедуру поиска L можно добавить кусок с уменьшением L, чтобы на каждой итерации иметь оценку на L максимально близкую к истинной локальной константе гладкости.

Замечание Л3.2. При использовании адаптивного подбора L к итоговой оракульной и итерационной сложности метода добавляется лишь логарифмическое слагаемое.

Замечание Л3.3. Для μ-сильно выпуклых L-гладких функций можно построить критерий выбора шага, зависящий одновременно от μ и от L, но поиск пары (μ, L) становится двумерным и несколько усложняет процедуру. При этом при выборе $\gamma_k = \frac{1}{L}$ линейный характер сходимости сохраняется, хотя может наблюдаться потеря в константной эффективности.

Л3.4.2 Степенной шаг

Также можно использовать убывающий с номером итерации шаг:

$$\gamma_k := \frac{\gamma}{\delta + k^p},\ \gamma > 0,\ \delta > 0,\ p > 0.$$

Наиболее часто применяются на практике $\gamma_k = \frac{1}{k+1}$ и $\gamma_k = \frac{1}{\sqrt{k+1}}$.

Л3.4.3 Наискорейший спуск

Итерация градиентного спуска выглядит следующим образом:

$$x^{k+1} = x^k - \gamma_k \nabla f\big(x^k\big).$$

Посмотрим на значение целевой функции после сделанного шага как на функцию от одного аргумента γ_k:

$$\phi(\gamma_k) := f\big(x^k - \gamma_k \nabla f\big(x^k\big)\big). \tag{Л3.7}$$

Поскольку стоит задача минимизации f, то будем выбирать шаг γ_k так, чтобы минимизировать значение функции $\phi(\gamma_k)$:

$$\gamma_k^* = \operatorname*{argmin}_{\gamma_k > 0} \phi(\gamma_k) = \operatorname*{argmin}_{\gamma_k > 0} f\big(x^k - \gamma_k \nabla f\big(x^k\big)\big).$$

Получим важное свойство шага наискорейшего спуска γ_k^*. Выпишем необходимое условие оптимума:

$$\begin{aligned}\phi'(\gamma_k^*) = \left.\frac{\partial f\big(x^k - \gamma_k \nabla f\big(x^k\big)\big)}{\partial \gamma_k}\right|_{\gamma_k = \gamma_k^*} &= -\big\langle \nabla f\big(x^k - \gamma_k^* \nabla f\big(x^k\big)\big), \nabla f\big(x^k\big)\big\rangle \\ &= -\big\langle \nabla f(x^{k+1}), \nabla f(x^k)\big\rangle = 0.\end{aligned}$$

Итак, свойство заключается в следующем: градиент в точке, куда мы шагаем, ортогонален направлению, вдоль которого мы шагаем. Действительно, если бы это было не так, то можно было бы сделать шаг в направлении градиента, чтобы уменьшить значение функции, что противоречит определению шага наискорейшего спуска. Рисунок Л3.2 иллюстрирует это свойство. В качестве целевой функции взята квадратичная функция на $\mathbb{R}^2$, эллипсами показаны линии уровня этой функции. Градиенты в каждой точке направлены ортогонально линии уровня, проходящей через эту точку.

Поиск оптимального значения γ_k можно осуществлять методами одномерной оптимизации. Причем, зачастую достаточно приближенного решения.

Методы одномерное оптимизации не требуют липшицевости и даже выпуклости минимизируемой функции скалярного аргумента. Потребуем, чтобы ϕ была унимодальной на рассматриваемом отрезке.

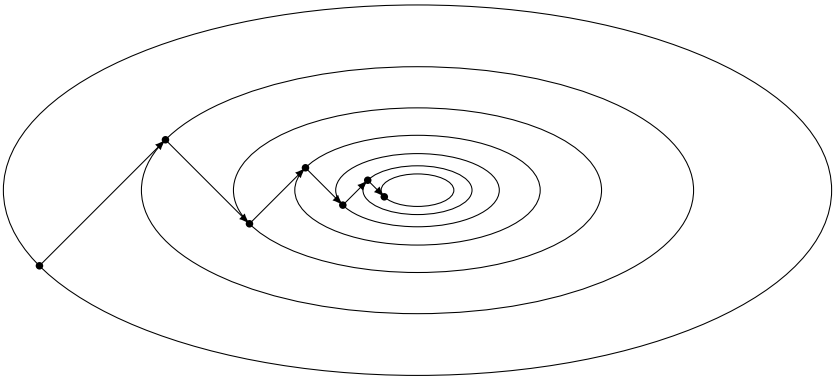


Рис. Л3.2: Наискорейший градиентный спуск для квадратичной задачи.

Определение Л3.1. Одномерная функция $\phi : [a, b] \to \mathbb{R}$ называется *унимодальной* на $[a, b]$, если существует $c^* \in [a, b]$ такое, что:

1. Для произвольных $x, y \in [a, c^*]$, таких что $x < y$, выполнено $\phi(x) > \phi(y)$;
2. Для произвольных $x, y \in [c^*, b]$, таких что $x < y$, выполнено $\phi(x) < \phi(y)$.

Замечание Л3.4. Если функция f является сильно выпуклой, то полученная в точке $x^k \neq x^*$ функция ϕ (Уравнение (Л3.7)) будет унимодальной.

Первым методом для поиска минимума унимодальных функций будет метод дихотомии. Дополнительно предположим, что нам известен отрезок $[a_0, b_0]$, на котором находится оптимум ϕ. В случае с градиентным спуском, это может быть отрезок $\left[0, \frac{2}{L}\right]$, если нам известна константа гладкости, либо $[0, b]$, где b взято достаточно большим, если L нам не известна. Будем искать $\gamma^* = \operatorname{argmin}_{\gamma \in [a,b]} \phi(\gamma)$.

Алгоритм Л3.3 Вычисление γ_k с помощью метода дихотомии

Вход: отрезок поиска $[a_0, b_0]$, количество итераций K
for $k = 0, \ldots, K-1$ **do**
  $\gamma_k = \frac{a_k + b_k}{2}$
  $\delta_k = \frac{a_k + b_k}{4}$
  $x_1 = \gamma_k - \delta_k$
  $x_2 = \gamma_k + \delta_k$
  if $\phi(x_1) < \phi(x_2)$ **then**
    $a_{k+1} = a_k$
    $b_{k+1} = x_2$
  else
    $a_{k+1} = x_1$
    $b_{k+1} = b_k$
  end if
end for
Выход: $\gamma_K = \frac{a_K + b_K}{2}$

Упражнение 13.1. Докажите корректность метода дихотомии и получите оценку

сходимости метода по аргументу.

Другой метод для поиска минимума унимодальной функции — метод золотого сечения. Постановка задачи такая же: ищем $\gamma^* = \operatorname{argmin}_{\gamma \in [a,b]} \phi(\gamma)$. Отличие от метода дихотомии заключается в том, в каких точках мы смотрим значения функции.

Алгоритм Л3.4 Вычисление γ_k с помощью золотого сечения

Вход: отрезок поиска $[a_0, b_0]$, число итераций K

1: $\varphi = \frac{\sqrt{5}-1}{2}$
2: $x_1 = b_0 - \varphi(b_0 - a_0)$
3: $x_2 = a_0 + \varphi(b_0 - a_0)$
4: **for** $i = 1, \ldots, K$ **do**
5: **if** $\phi(x_1) < \phi(x_2)$ **then**
6: $a_{k+1} = a_k$
7: $b_{k+1} = x_2$
8: $x_1 = b_{k+1} - \varphi(b_{k+1} - a_{k+1})$
9: $x_2 = x_1$
10: **else**
11: $a_{k+1} = x_1$
12: $b_{k+1} = b_k$
13: $x_1 = x_2$
14: $x_2 = a_{k+1} + \varphi(b_{k+1} - a_{k+1})$
15: **end if**
16: **end for**

Выход: $\gamma_k = \frac{a_K + b_K}{2}$

Разбиение с помощью золотого сечения позволяет нам переиспользовать точки с предыдущей итерации: это заметно по строкам 7 и 11 алгоритма. Это приводит к уменьшению количества обращений к функции.

Упражнение 13.2. Докажите корректность алгоритма золотого сечения и получите оценку сходимости метода по аргументу.

Упражнение 13.3. Сравните алгоритмы метода дихотомии и золотого сечения. Какой из них быстрее сходится с точки зрения количества итераций? А с точки зрения обращений к функции?

Л3.4.4 Шаг Поляка–Шора

Рассмотрим другой подход к выбору шага, основанный на минимизации верхней оценки на расстояние до решения, полученной ранее. В ходе доказательства Теоремы Л3.4 о сходимости градиентного спуска для гладких выпуклых функций мы получили неравенство (Л3.6):

$$\left\|x^{k+1} - x^*\right\|_2^2 \leq \left\|x^k - x^*\right\|_2^2 - 2\gamma_k(f(x^k) - f^*) + \gamma_k^2 \left\|\nabla f(x^k)\right\|_2^2.$$

Правая часть неравенства является квадратичной функцией относительно шага γ_k. Попробуем выбрать γ_k так, чтобы минимизировать эту верхнюю оценку. Для этого найдем производную по γ_k и приравняем ее к нулю:

$$-2(f\left(x^k\right) - f^*) + 2\gamma_k \left\|\nabla f\left(x^k\right)\right\|_2^2 = 0.$$

Отсюда получаем оптимальное значение шага, минимизирующее данную оценку:

$$\gamma_k^* = \frac{f(x^k) - f^*}{\|\nabla f(x^k)\|_2^2}.$$

Этот выбор шага известен как *шаг Поляка–Шора*. В общем виде шаг Поляка–Шора записывается с параметром $\alpha > 0$:

$$\gamma_k^* := \frac{f(x^k) - f^*}{\alpha\|\nabla f(x^k)\|_2^2}.$$

Значение функции в оптимуме часто неизвестно, поэтому обычно используется нижняя оценка. Например, для задачи наименьших квадратов целевая функция неотрицательна $f(x) = \|Ax - b\|_2^2 \geq 0$.

Л3.4.5 Стратегия бэктрекинга для подбора размера шага

Введем в рассмотрение $\varphi(\gamma_k) = f(x_k + \gamma_k h_k)$ и некоторые константы β_1, β_2, удовлетворяющие условию $0 < \beta_1 < \beta_2 < 1$ (Обычно $0 < \beta_1 < 0.3$, $0.9 < \beta_1 < 1$). Сформулируем несколько условий, которые будут полезными при выборе размера шага:

- **Условие достаточного убывания** гарантирует, что функция в точке x^{k+1} не превосходит линейной аппроксимации с коэффициентом наклона β_1:

  $$f(x^{k+1}) \leq f(x^k) + \beta_1\gamma_k\langle\nabla f(x^k), h^k\rangle,$$

  что эквивалентно

  $$\varphi(\gamma_k) \leq \varphi(0) + \beta_1\gamma_k\varphi'(0).$$

- **Условие существенного убывания** гарантирует, что функция в точке x^{k+1} не меньше линейной аппроксимации с коэффициентом наклона β_2:

  $$f(x^{k+1}) \geq f(x^k) + \beta_2\gamma_k\langle\nabla f(x^k), h^k\rangle,$$

  что эквивалентно

  $$\varphi(\gamma_k) \geq \varphi(0) + \beta_2\gamma_k\varphi'(0).$$

- **Условие кривизны** гарантирует, что угол наклона касательной в точке x^{k+1} не меньше, чем угол наклона касательной в точке x^k, умноженный на β_2:

  $$\langle\nabla f(x^{k+1}), h^k\rangle \geq \beta_2\langle\nabla f(x^k), h^k\rangle,$$

  что эквивалентно

  $$\varphi'(\gamma_k) \geq \beta_2\varphi'(0).$$

Различным наборам условий обычно дают собственные имена:

- Армихо (достаточное убывание).

Алгоритм Л3.5 Вычисление γ_k с помощью правила Армихо

Вход: стартовая точка $x^0 \in \mathbb{R}^d$, начальное значение $\gamma_k > 0$, $0 < \beta_1 < 1$, $0 < \rho < 1$, $\langle \nabla f(x^k), h^k \rangle < 0$

$x^{k+1} = x^k + \gamma_k h^k$
while $f(x^{k+1}) > f(x^k) + \beta_1 \gamma_k \langle \nabla f(x^k), h^k \rangle$ **do**
    $\gamma_k = \rho \cdot \gamma_k$
    $x^{k+1} = x^k + \gamma_k h^k$
end while
Выход: $\frac{\gamma_k}{\rho}$

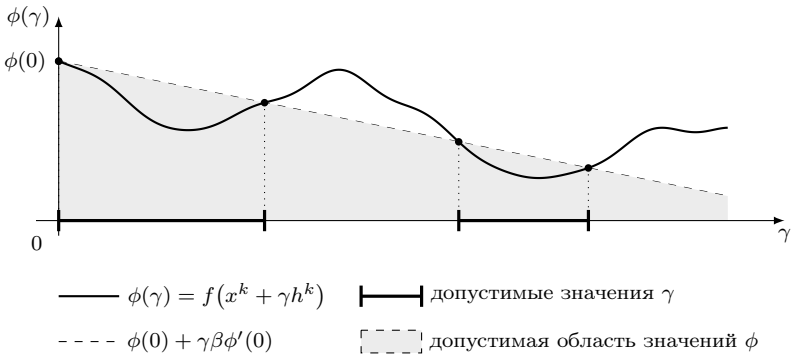


Рис. Л3.3: Иллюстрация условий Армихо.

- Вольфа (достаточное убывание + кривизна).

Алгоритм Л3.6 Вычисление γ_k с помощью правила Вольфа

Вход: стартовая точка $x^0 \in \mathbb{R}^d$, начальное значение $\gamma_k > 0$, $l = 0$, $u = \infty$, $\varepsilon > 0$, $0 < \beta_1 < \beta_2 < 1$, $\langle \nabla f(x^k), h^k \rangle < 0$

while $u - l > \varepsilon$ **do**
    if $f(x^{k+1}) > f(x^k) + \beta_1 \gamma_k \langle \nabla f(x^k), h^k \rangle$ **then**
        $u = \gamma_k$
        $\gamma_k = \frac{l+u}{2}$
    else
        if $\langle \nabla f(x^{k+1}), h^k \rangle < \beta_2 \langle \nabla f(x^k), h^k \rangle$ **then**
            $l = \gamma_k$
            if $u = \infty$ **then**
                $\gamma_k = 2 \cdot l$
            else
                $\gamma_k = \frac{l+u}{2}$
            end if
        end if
    end if
end while

В этом случае условие на строке 6 в Алгоритме Л3.6 заменяется на

$$\left|\left\langle \nabla f\left(x^{k+1}\right), h^k\right\rangle\right| \leq \beta_2 \left|\left\langle \nabla f\left(x^k\right), h^k\right\rangle\right|,\ \beta_2 \in (0,1).$$

Иллюстрация условий на Рисунке Л3.4.

- Усиленные Вольфа (достаточное убывание + кривизна с двух сторон).

  Иллюстрация условий на Рисунке Л3.5.

- Гольдштейна (достаточное убывание + существенное убывание):

  Рассматриваются $\beta_2 = 1 - \beta_1$ и $\beta_1 \in \left(0, \frac{1}{2}\right)$. В этом случае условие на строке 6 в Алгоритме Л3.6 заменяется на

$$f\left(x^{k+1}\right) < f\left(x^k\right) + (1-\beta_1)\gamma_k \left\langle \nabla f\left(x^k\right), h^k\right\rangle.$$

  Иллюстрация условий на Рисунке Л3.6.

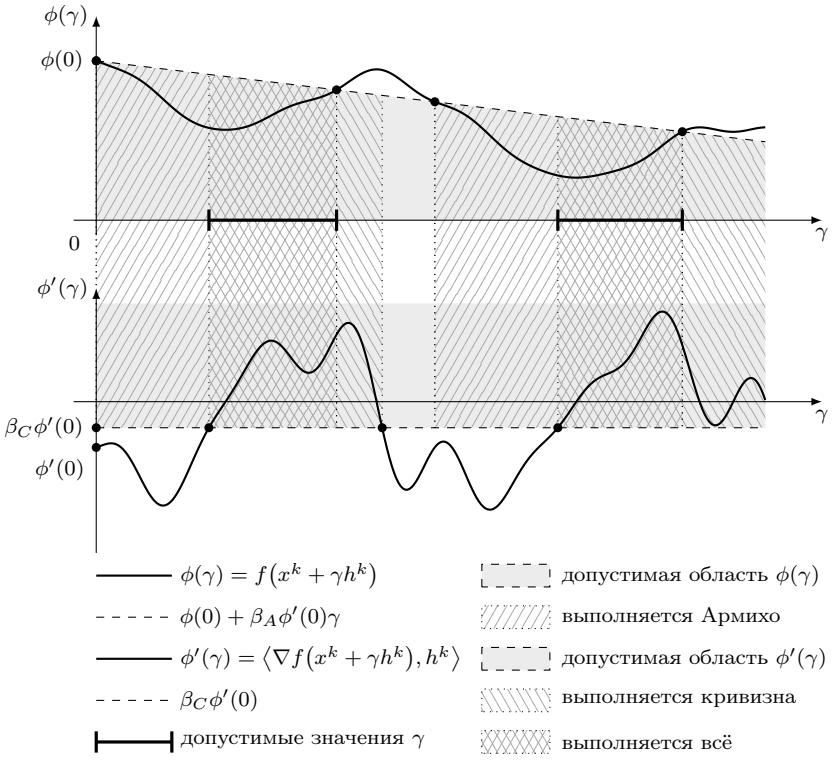


Рис. Л3.4: Иллюстрация условий Вольфа: Армихо (верхний график) и кривизна (нижний график).

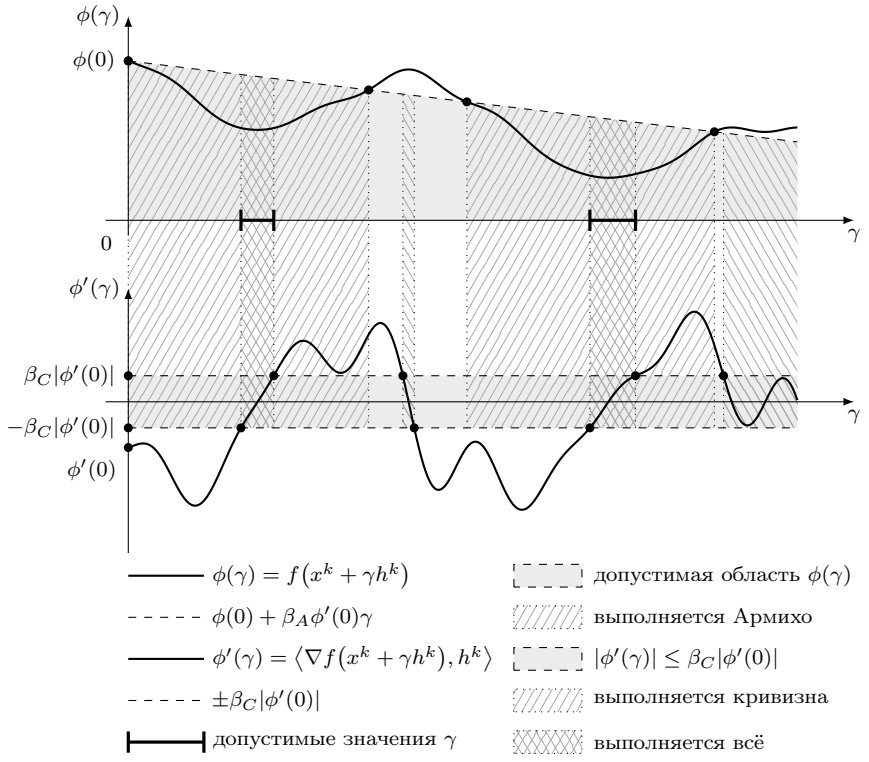


Рис. Л3.5: Иллюстрация усиленных условий Вольфа (Гольдштейн): Армихо (верхний график) и $|\phi'(\gamma)| \le \beta_C|\phi'(0)|$ (нижний график).

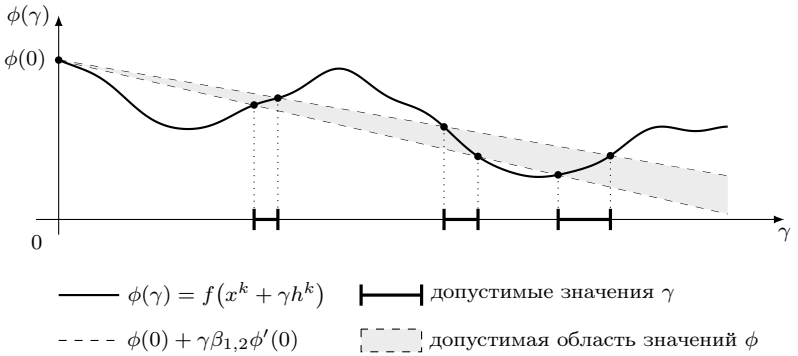


Рис. Л3.6: Иллюстрация условий Гольдштейна.

Л4 Ускоренные и оптимальные методы

В Параграфе Л3 была доказана сходимость градиентного спуска для L-гладких и μ-сильно выпуклых задач. В ней было показано, что для получения ε-решения необходимо совершить $\mathcal{O}\left(\frac{L}{\mu}\log\frac{\left\|x^0-x^*\right\|_2}{\varepsilon}\right)$ итераций. Логичный вопрос — хорошая ли это оценка, или её можно улучшить? Обсудим это в этом параграфе.

Л4.1 Метод тяжёлого шарика

Первое предложение по усовершенствованию градиентного спуска — использовать память с предыдущих итераций, добавить к алгоритму «инерцию». С такой эвристикой Б.Т. Поляк в 1964 году предложил метод тяжёлого шарика [19]:

Алгоритм Л4.1 Метод тяжёлого шарика

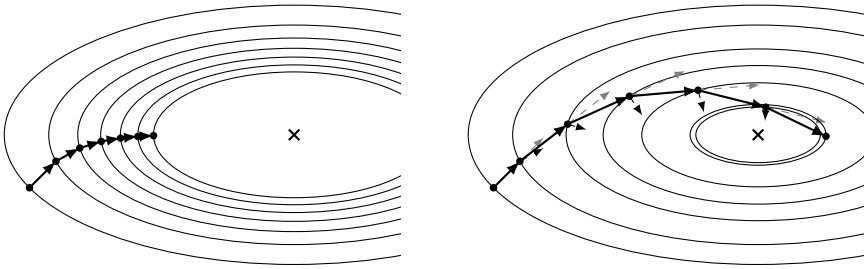
Вход: стартовая точка $x^0 = x^{-1} \in \mathbb{R}^d$, размеры шагов $\{\gamma_k\}_{k=0} > 0$, моментумы $\{\tau_k\}_{k=0} \in [0,1]$, количество итераций K

1: **for** $k = 0, 1, \ldots, K-1$ **do**

2: $\quad x^{k+1} = x^k - \gamma_k \nabla f\left(x^k\right) + \tau_k\left(x^k - x^{k-1}\right)$

3: **end for**

Выход: x^K



(a) Градиентный спуск.

(b) Метод тяжёлого шарика с $\tau = 0.9$.

Рис. Л4.1: Сравнение 6 шагов траекторий градиентного спуска и метода тяжёлого шарика с постоянным шагом $\gamma = 0.1$ при минимизации квадратичной функции $f(x) = \frac{1}{2}\langle x, Ax\rangle$, где $A = \mathrm{diag}(1,5)$. Эллипсы соответствуют уровням функции. На рисунке справа черным пунктиром показано направление градиента, серым пунктиром — инерция с прошлого шага.

У такой модели есть несколько плюсов. Во-первых, у метода понятная физика и интуиция: направления шага на предыдущей итерации и на минимум функции не должны сильно отличаться. Во-вторых, этот метод легок в имплементации: мы должны дополнительно хранить только координаты предыдущего состояния, а при вычислении нового просто добавлять слагаемое, отвечающее за «моментум». Однако у такого метода есть и ряд минусов. Помимо подбора шага, теперь необходимо подбирать параметр τ_k, который отвечает за инерцию. Обычно его берут близким к единице. Кроме того, мы должны найти ответ на исходный вопрос — лучше ли теоретические оценки сходимости? Для начала, рассмотрим оценки для метода на квадратичной задаче оптимизации.

Теорема Л4.1 (Теорема 3 из [19]). Пусть имеется квадратичная функция $f : \mathbb{R}^d \to \mathbb{R}^d$:

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

где A — симметричная положительно-определённая матрица с собственными числами на отрезке $[\mu, L]$. Пусть задача безусловной оптимизации (Л3.1) решается методом тяжёлого шарика (Алгоритм Л4.1). Тогда при

$$\gamma_k = \frac{4}{\left(\sqrt{L} + \sqrt{\mu}\right)^2},\ \tau_k = \left(\frac{\sqrt{L} - \sqrt{\mu}}{\sqrt{L} + \sqrt{\mu}}\right)^2$$

справедлива следующая оценка сходимости:

$$\left\|x^K - x^*\right\|_2 \leq C \cdot \left(\frac{\sqrt{L} - \sqrt{\mu}}{\sqrt{L} + \sqrt{\mu}}\right)^K \left(\left\|x^1 - x^*\right\|_2^2 + \left\|x^0 - x^*\right\|_2^2\right)^{\frac{1}{2}},$$

где x^* — решение задачи $Ax^* = b$, константа $C > 0$.
Более того, чтобы добиться точности ε по аргументу $\left(\left\|x^k - x^*\right\|_2 \leq \varepsilon\right)$, необходимо

$$K = \mathcal{O}\left(\sqrt{\frac{L}{\mu}} \log \frac{\left\|x^0 - x^*\right\|_2}{\varepsilon}\right) = \tilde{\mathcal{O}}\left(\sqrt{\frac{L}{\mu}}\right) \text{ итераций.}$$

Доказательство. Перепишем метод тяжёлого шарика в векторной форме:

$$\begin{pmatrix} x_{k+1} \\ x_k \end{pmatrix} = \begin{pmatrix} (1+\tau_k) I_d & -\tau_k I_d \\ I_d & 0 \end{pmatrix} \begin{pmatrix} x_k \\ x_{k-1} \end{pmatrix} + \begin{pmatrix} -\gamma_k \nabla f(x_k) \\ 0 \end{pmatrix}.$$

Для квадратичной функции $f(x) = \frac{1}{2}\langle x, Ax\rangle - \langle b, x\rangle$ градиент $\nabla f(x) = Ax - b$. В точке минимума x^* градиент обнуляется: $Ax^* = b$. Поэтому:

$$\nabla f(x_k) = A(x_k - x^*).$$

Введём вектор отклонения от решения: $e_k = x_k - x^*$. Тогда метод можно переписать как

$$\begin{pmatrix} e_{k+1} \\ e_k \end{pmatrix} = \begin{pmatrix} (1+\tau_k) I_d - \gamma_k A & -\tau_k I_d \\ I_d & 0 \end{pmatrix} \begin{pmatrix} e_k \\ e_{k-1} \end{pmatrix},$$

что задаёт рекуррентное соотношение с матрицей итераций

$$A_k = \begin{pmatrix} (1+\tau_k) I_d - \gamma_k A & -\tau_k I_d \\ I_d & 0 \end{pmatrix}.$$

Рассмотрим спектр матрицы A_k. Так как A симметрична и положительно определённа, её можно диагонализовать:

$$A = U \Lambda U^\top,\ \text{где } \Lambda = \operatorname{diag}\left(\lambda_1, \ldots, \lambda_d\right),\ U^\top U = I_d.$$

Тогда

$$A_k = \begin{pmatrix} U & 0 \\ 0 & U \end{pmatrix} \begin{pmatrix} (1+\tau_k) I_d - \gamma_k \Lambda & -\tau_k I_d \\ I_d & 0 \end{pmatrix} \begin{pmatrix} U^\top & 0 \\ 0 & U^\top \end{pmatrix}.$$

Это означает, что можно применить такое ортогональное преобразование, переставляющее строки и столбцы, что спектр A_k совпадёт со спектром блочно-диагональной матрицы

$$\widehat{A}_k = \operatorname{diag}\left(T_1, \ldots, T_d\right),\ T_i = \begin{pmatrix} 1 + \tau_k - \gamma_k \lambda_i & -\tau_k \\ 1 & 0 \end{pmatrix}.$$

Спектральный радиус A_k равен

$$\rho(A_k) = \max_{i=\overline{1,d}} \rho(T_i).$$

Для оценки $\rho(T_i)$ найдём собственные числа u оператора T_i по характеристическому уравнению:

$$\det \begin{pmatrix} 1+\tau_k-\gamma_k\lambda_i-u & -\tau \\ 1 & -u \end{pmatrix} = 0,$$

что даёт:

$$u^2 - u\big(1+\tau_k-\gamma_k\lambda_i\big) + \tau_k = 0.$$

Это квадратное уравнение имеет корни

$$u_{1,2} = \frac{\big(1+\tau_k-\gamma_k\lambda_i\big) \pm \sqrt{\big(1+\tau_k-\gamma_k\lambda_i\big)^2 - 4\tau_k}}{2}.$$

При выборе параметров:

$$\gamma_k = \frac{4}{\big(\sqrt{L}+\sqrt{\mu}\big)^2}, \quad \tau_k = \left(\frac{\sqrt{L}-\sqrt{\mu}}{\sqrt{L}+\sqrt{\mu}}\right)^2,$$

получаем, что

$$u_{1,2} = \frac{\big(L+\mu-2\lambda_i\big) \pm 2\sqrt{\big(L-\lambda_i\big)\big(\mu-\lambda_i\big)}}{\big(\sqrt{L}+\sqrt{\mu}\big)^2} = \frac{\big(L-\lambda_i \pm \big(\mu-\lambda_i\big)\big)^2}{\big(\sqrt{L}+\sqrt{\mu}\big)^2}.$$

Пусть

$$u_1 = \frac{\big(L-\mu\big)^2}{\big(\sqrt{L}+\sqrt{\mu}\big)^2},\ u_2 = \frac{\big(L+\mu-2\lambda_i\big)^2}{\big(\sqrt{L}+\sqrt{\mu}\big)^2},$$

то есть $u_1 \geq u_2$ при $\forall\ \lambda_i \in [\mu, L]$. Получаем, что

$$\rho(A_k) = \frac{\sqrt{L}-\sqrt{\mu}}{\sqrt{L}+\sqrt{\mu}}.$$

Тогда

$$\left\| \begin{pmatrix} e_K \\ e_{K-1} \end{pmatrix} \right\|_2 \leq C \cdot \left(\frac{\sqrt{L}-\sqrt{\mu}}{\sqrt{L}+\sqrt{\mu}}\right)^k \left\| \begin{pmatrix} e_1 \\ e_0 \end{pmatrix} \right\|_2,$$

откуда получаем

$$\big\|x^K - x^*\big\|_2 \leq C \cdot \left(\frac{\sqrt{L}-\sqrt{\mu}}{\sqrt{L}+\sqrt{\mu}}\right)^k \Big(\big\|x^1 - x^*\big\|_2^2 + \big\|x^0 - x^*\big\|_2^2\Big)^{\frac{1}{2}}.$$

Воспользуемся тем, что $\mu \leq L$, затем применим неравенство (0.4):

$$\begin{aligned}
\big\|x^K - x^*\big\|_2 &\leq C \cdot \left(\frac{\sqrt{L}-\sqrt{\mu}}{\sqrt{L}+\sqrt{\mu}}\right)^K \Big(\big\|x^1 - x^*\big\|_2^2 + \big\|x^0 - x^*\big\|_2^2\Big)^{\frac{1}{2}} \\
&= C \cdot \left(1 - \frac{2\sqrt{\mu}}{\sqrt{L}+\sqrt{\mu}}\right)^K \Big(\big\|x^1 - x^*\big\|_2^2 + \big\|x^0 - x^*\big\|_2^2\Big)^{\frac{1}{2}} \\
&\leq C \cdot \left(1 - \frac{\sqrt{\mu}}{\sqrt{L}}\right)^K \Big(\big\|x^1 - x^*\big\|_2^2 + \big\|x^0 - x^*\big\|_2^2\Big)^{\frac{1}{2}} \\
&\leq C \cdot \exp\left(-\sqrt{\frac{\mu}{L}}K\right) \Big(\big\|x^1 - x^*\big\|_2^2 + \big\|x^0 - x^*\big\|_2^2\Big)^{\frac{1}{2}}.
\end{aligned}$$

Ограничим правую часть неравенства ε:

$$\left\|x^K - x^*\right\|_2 \leq C \cdot \exp\left(-\sqrt{\frac{\mu}{L}}K\right)\left(\left\|x^1 - x^*\right\|_2^2 + \left\|x^0 - x^*\right\|_2^2\right)^{\frac{1}{2}} \leq \varepsilon.$$

Тогда после логарифмирования и преобразования получаем:

$$K \geq \sqrt{\frac{L}{\mu}} \log \left(\frac{C \cdot \left(\left\|x^1 - x^*\right\|_2^2 + \left\|x^0 - x^*\right\|_2^2\right)^{\frac{1}{2}}}{\varepsilon} \right)$$
$$= \mathcal{O}\left(\sqrt{\frac{L}{\mu}} \log \frac{\left\|x^0 - x^*\right\|_2}{\varepsilon}\right) = \tilde{\mathcal{O}}\left(\sqrt{\frac{L}{\mu}}\right).$$

■

Таким образом, на квадратичной задаче метод тяжёлого шарика действительно оказывается быстрее метода градиентного спуска: у градиентного спуска оценка на количество итераций $\mathcal{O}\left(\frac{L}{\mu} \log \frac{\left\|x^0 - x^*\right\|_2}{\varepsilon}\right)$, а у метода тяжёлого шарика $\mathcal{O}\left(\sqrt{\frac{L}{\mu}} \log \frac{\left\|x^0 - x^*\right\|_2}{\varepsilon}\right)$. Но будет ли метод тяжёлого шарика быстрее на более общем классе задач, а именно на μ-сильно выпуклых и L-гладких? Оказывается, что нет. Приведем пример такой μ-сильно выпуклой L-гладкой функции, что при выборе параметров метода, как на квадратичной задаче, метод тяжёлого шарика не сходится вовсе.

Пример Л4.1 (Секция 4.6 из [10])**.** Рассмотрим функцию

$$f(x) = \begin{cases} \frac{25}{2}x^2 + 12, & x < 1, \\ \frac{1}{2}x^2 + 24x, & 1 \leq x < 2, \\ \frac{25}{2}x^2 - 24x + 48, & x \geq 2, \end{cases}$$

её градиент

$$\nabla f(x) = \begin{cases} 25x, & x < 1, \\ x + 24, & 1 \leq x < 2, \\ 25x - 24, & x \geq 2. \end{cases}$$

Константы гладкости и сильной выпуклости для f: $\mu = 1$, $L = 25$. Однако, если выбрать начальное приближение $x_0 \in [3.07, 3.46]$, метод тяжёлого шарика с параметрами

$$\gamma_k = \frac{4}{(\sqrt{L} + \sqrt{\mu})^2} = \frac{1}{9}, \ \tau_k = \left(\frac{\sqrt{L} - \sqrt{\mu}}{\sqrt{L} + \sqrt{\mu}}\right)^2 = \frac{4}{9}$$

не сходится, а порождает предельный цикл, как показано на Рисунке Л4.2.

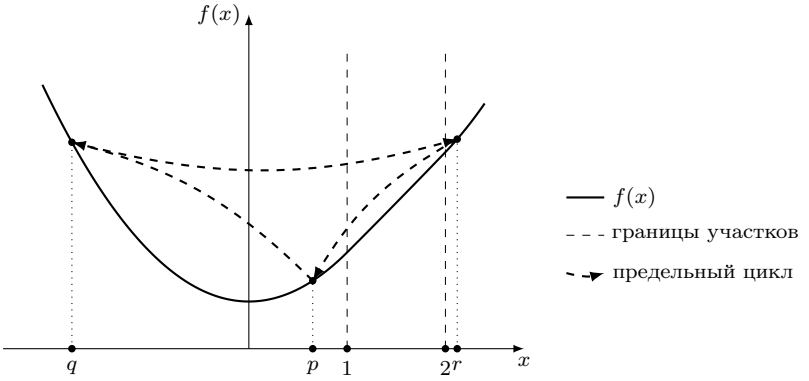


Рис. Л4.2: Иллюстрация предельного цикла в методе тяжёлого шарика на кусочно-гладкой функции $f(x)$.

Действительно, итерация метода тяжёлого шарика в точке x^k выглядит следующим образом:

$$x_{k+1} = \frac{13}{9}x_k - \frac{4}{9}x_{k-1} - \frac{1}{9}\nabla f(x_k).$$

И тогда точки $q = -\frac{2208}{1225}$, $p = \frac{792}{1225}$, $r = \frac{2592}{1225}$ образуют предельный цикл, как показано на Рисунке Л4.2. При этом, если выбрать начальное приближение $x_0 \in [3.07, 3.46]$, то траектория метода будет асимптотически стремиться к предельному циклу.

Таким образом, мы убедились, что оптимальные параметры, подобранные для квадратичной задачи, не гарантируют сходимости на более общем классе μ-сильно выпуклых и L-гладких функций. Для этого класса метод тяжёлого шарика даёт ускорение — $\mathcal{O}\left(\sqrt{\frac{L}{\mu}}\log\frac{\left\|x^0-x^*\right\|_2}{\varepsilon}\right)$ только в случае локальной сходимости [19], когда стартовая точка x^0 уже достаточно близка к решению x^*. Если же рассматривать глобальную сходимость, то есть из произвольной точки x^0, то оценки совпадают с оценками для градиентного спуска: $\mathcal{O}\left(\frac{L}{\mu}\log\frac{\left\|x^0-x^*\right\|_2}{\varepsilon}\right)$. Справиться с этой проблемой должен следующий метод.

Л4.2 Ускоренный градиентный метод

В 1983 году Ю.Е. Нестеров предложил ускоренный градиентный метод [15]:

Алгоритм Л4.2 Ускоренный градиентный метод (Метод Нестерова)

Вход: стартовые точки $x^0 = x^{-1} \in \mathbb{R}^d$, размеры шагов $\{\gamma_k\}_{k=0} > 0$, моментумы $\{\tau_k\}_{k=0} \in [0,1]$, количество итераций K

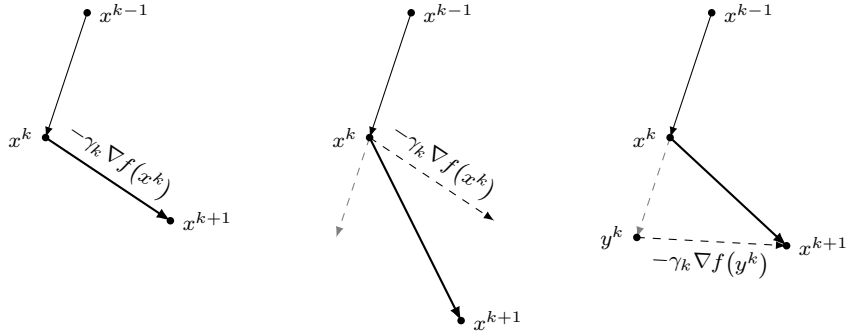
for $k = 0, 1, \ldots, K-1$ **do**
  $y^k = x^k + \tau_k\left(x^k - x^{k-1}\right)$
  $x^{k+1} = y^k - \gamma_k \nabla f\left(y^k\right)$
end for
Выход: x^K

Этот алгоритм с первого взгляда очень похож на метод тяжёлого шарика. Однако,

если подробно написать его итерацию:

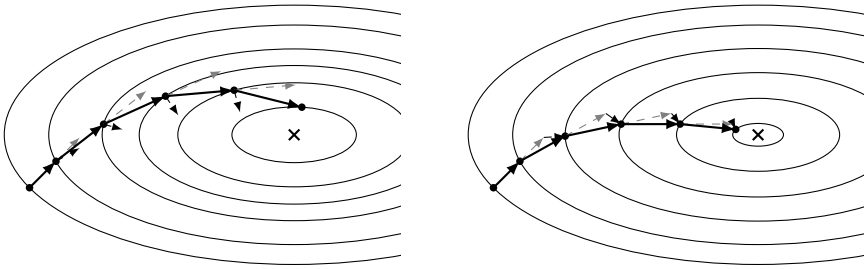
$$x^{k+1} = x^k + \tau_k\left(x^k - x^{k-1}\right) - \gamma_k \nabla f\left(x^k + \tau_k\left(x^k - x^{k-1}\right)\right),$$

то можно заметить, что градиент, как бы, считается из «будущего» состояния. Эта техника позволяет добиться $\mathcal{O}\left(\sqrt{\frac{L}{\mu}}\log\frac{f\left(x^0\right)-f\left(x^*\right)}{\varepsilon}\right)$ вызовов оракула для ε-сходимости по функции. Разницу между градиентным спуском, методом тяжёлого шарика и методом Нестерова можно увидеть на Рисунках Л4.3 и Л4.4. Как можно заметить, метод Нестерова приближается к оптимуму быстрее.



(a) Градиентный спуск (b) Метод тяжёлого шарика (c) Метод Нестерова

Рис. Л4.3: Сравнение шага методов, серым пунктиром показан $\tau_k\left(x^k - x^{k-1}\right)$.



(a) Метод тяжёлого шарика с $\tau = 0.9$. (b) Метод Нестерова с $\tau = 0.9$.

Рис. Л4.4: Сравнение траекторий: метод тяжёлого шарика и метод Нестерова с $\gamma = 0.1$.

Теорема Л4.2 (Теорема 1 из [5])**.** Пусть задача безусловной оптимизации (Л3.1) с L-гладкой, μ-сильно выпуклой целевой функцией f решается с помощью ускоренного градиентного метода (Алгоритм Л4.2). Тогда при $\gamma_k = \frac{1}{L}$, $\tau_k = \frac{\sqrt{L}-\sqrt{\mu}}{\sqrt{L}+\sqrt{\mu}}$ справедлива следующая оценка сходимости:

$$f\left(x^K\right) - f^* \leq \left(1 - \sqrt{\frac{\mu}{L}}\right)^K \cdot L\|x^0 - x^*\|_2^2.$$

Более того, чтобы добиться точности ε по функции ($f(x^K) - f^* \leq \varepsilon$), необходимо

$$K = \mathcal{O}\left(\sqrt{\frac{L}{\mu}} \log \frac{L\|x^0 - x^*\|_2^2}{\varepsilon}\right) \text{ итераций.}$$

Доказательство. Подставим в итерацию ускоренного градиентного метода γ_k и τ_k из условия:

$$y^k = x^k + \frac{\sqrt{L} - \sqrt{\mu}}{\sqrt{L} + \sqrt{\mu}}(x^k - x^{k-1}),$$
$$x^{k+1} = y^k - \frac{1}{L}\nabla f(y^k).$$

Не ограничивая общности, будем считать $x^* = 0$, сделав замену координат $x_k \to x_k - x^*$ и $y_k \to y_k - x^*$.
Введём функцию следующего вида (*функцию Ляпунова*):

$$V_k := f(x^k) - f^* + \frac{L}{2}\|x^k - \rho^2 x^{k-1}\|_2^2,$$

где $\rho^2 = 1 - \sqrt{\frac{\mu}{L}}$. Покажем, что для неё выполняется $V_{k+1} \leq \rho^2 V_k$.
Для выпуклой L-гладкой функции f запишем неравенство (Л2.3) из Теоремы Л2.5 для точек x^{k+1}, y^k:

$$\begin{aligned} V_{k+1} &= f(x^{k+1}) - f^* + \frac{L}{2}\|x^{k+1} - \rho^2 x^k\|_2^2 \\ &\leq f(y^k) + \langle \nabla f(y^k), x^{k+1} - y^k\rangle + \frac{L}{2}\|x^{k+1} - y^k\|_2^2 - f^* + \frac{L}{2}\|x^{k+1} - \rho^2 x^k\|_2^2. \end{aligned}$$

Подставляем формулу итерации $x^{k+1} - y^k = -\frac{1}{L}\nabla f(y^k)$:

$$\begin{aligned} V_{k+1} &\leq f(y^k) + \left\langle \nabla f(y^k), -\frac{1}{L}\nabla f(y^k)\right\rangle + \frac{L}{2}\left\|-\frac{1}{L}\nabla f(y^k)\right\|_2^2 - f^* + \frac{L}{2}\|x^{k+1} - \rho^2 x^k\|_2^2 \\ &= f(y^k) - \frac{1}{2L}\|\nabla f(y^k)\|_2^2 - f^* + \frac{L}{2}\|x^{k+1} - \rho^2 x^k\|_2^2. \end{aligned}$$

Прибавим и вычтем $\rho^2\big(f(y^k) - f^* + \langle \nabla f(y^k), x^k - y^k\rangle\big)$:

$$\begin{aligned} V_{k+1} &\leq f(y^k) - f^* - \frac{1}{2L}\|\nabla f(y^k)\|_2^2 + \frac{L}{2}\|x^{k+1} - \rho^2 x^k\|_2^2 \\ &\quad + \rho^2\big(f(y^k) - f^* + \langle \nabla f(y^k), x^k - y^k\rangle\big) - \rho^2\big(f(y^k) - f^* + \langle \nabla f(y^k), x^k - y^k\rangle\big) \\ &= \rho^2\big(\underbrace{f(y^k)}_{(1)} - f^* + \langle \nabla f(y^k), x^k - y^k\rangle\big) - \rho^2\langle \nabla f(y^k), x^k - y^k\rangle \\ &\quad + \big(1 - \rho^2\big)\big(f(y^k) - \underbrace{f^*}_{(2)} - \langle \nabla f(y^k), y^k\rangle\big) + \big(1 - \rho^2\big)\langle \nabla f(y^k), y^k\rangle \\ &\quad - \frac{1}{2L}\|\nabla f(y^k)\|_2^2 + \frac{L}{2}\|x^{k+1} - \rho^2 x^k\|_2^2. \end{aligned}$$

Далее, по воспользуемся μ-сильной выпуклостью f (Определение Л2.4) для точек x^k, y^k и $x^* = 0, y^k$:

$$\begin{aligned} f(y^k) &\le f(x^k) + \langle \nabla f(y^k), y^k - x^k \rangle - \frac{\mu}{2}\|y^k - x^k\|_2^2, \\ f^* &\ge f(y^k) + \langle \nabla f(y^k), x^* - y^k \rangle + \frac{\mu}{2}\|y^k - x^*\|_2^2 \\ &= f(y^k) - \langle \nabla f(y^k), y^k \rangle + \frac{\mu}{2}\|y^k\|_2^2. \end{aligned}$$

Подставляем эти два неравенства в (1) и (2) соответственно и получаем:

$$\begin{aligned} V_{k+1} &\le \rho^2\Big(f(x^k) - f^* - \frac{\mu}{2}\|x^k - y^k\|_2^2\Big) - \rho^2\langle \nabla f(y^k), x^k - y^k \rangle \\ &\quad - (1-\rho^2)\frac{\mu}{2}\|y^k\|_2^2 + (1-\rho^2)\langle \nabla f(y^k), y^k \rangle \\ &\quad - \frac{1}{2L}\|\nabla f(y^k)\|_2^2 + \frac{L}{2}\|x^{k+1} - \rho^2 x^k\|_2^2 \\ &= \rho^2 V_k + R_k, \end{aligned}$$

где введено обозначение

$$\begin{aligned} R_k = -\rho^2\frac{\mu}{2}\|x^k - y^k\|_2^2 - (1-\rho^2)\frac{\mu}{2}\|y^k\|_2^2 + \langle \nabla f(y^k), y^k - \rho^2 x^k \rangle - \frac{1}{2L}\|\nabla f(y^k)\|_2^2 \\ + \frac{L}{2}\|x^{k+1} - \rho^2 x^k\|_2^2 - \frac{\rho^2 L}{2}\|x^k - \rho^2 x^{k-1}\|_2^2. \end{aligned}$$

Преобразуем выражение выше:

$$\begin{aligned} R_k &= -\frac{\rho^2\mu}{2}\|x^k - y^k\|_2^2 - \sqrt{\frac{\mu}{L}}\frac{\mu}{2}\|y^k\|_2^2 + \langle \nabla f(y^k), y^k - \rho^2 x^k \rangle - \frac{1}{2L}\|\nabla f(y^k)\|_2^2 \\ &\quad + \frac{L}{2}\Big\|y^k - \frac{1}{L}\nabla f(y^k) - \rho^2 x^k\Big\|_2^2 - \frac{\rho^2 L}{2}\|x^k - \rho^2 x^{k-1}\|_2^2 \\ &= -\frac{\rho^2\mu}{2}\|x^k - y^k\|_2^2 - \sqrt{\frac{\mu}{L}}\frac{\mu}{2}\|y^k\|_2^2 + \frac{L}{2}\|y^k - \rho^2 x^k\|_2^2 - \frac{\rho^2 L}{2}\|x^k - \rho^2 x^{k-1}\|_2^2 \\ &= -\frac{\rho^2\mu}{2}\|x^k - y^k\|_2^2 - \sqrt{\frac{\mu}{L}}\frac{\mu}{2\big(\sqrt{L}+\sqrt{\mu}\big)^2}\|2\sqrt{L}x^k - \big(\sqrt{L}-\sqrt{\mu}\big)x^{k-1}\|_2^2 \\ &\quad + \frac{L}{2\big(\sqrt{L}+\sqrt{\mu}\big)^2}\Big\|\Big(\sqrt{L} + \frac{\mu}{\sqrt{L}}\Big)x^k - \big(\sqrt{L}-\sqrt{\mu}\big)x^{k-1}\Big\|_2^2 \\ &\quad - \frac{\sqrt{L}-\sqrt{\mu}}{2\sqrt{L}}\|\sqrt{L}x^k - \big(\sqrt{L}-\sqrt{\mu}\big)x^{k-1}\|_2^2 \\ &= -\frac{\rho^2\mu}{2}\|x^k - y^k\|_2^2 - \sqrt{\frac{\mu}{L}}\frac{\rho^2 L}{2}\left(\frac{\sqrt{L}-\sqrt{\mu}}{\sqrt{L}+\sqrt{\mu}}\right)^2\Big(\|x^k\|_2^2 - 2\langle x^k, x^{k-1}\rangle + \|x^{k-1}\|_2^2\Big) \\ &= -\frac{\rho^2\mu}{2}\|x^k - y^k\|_2^2 - \sqrt{\frac{\mu}{L}}\frac{\rho^2 L}{2}\left(\frac{\sqrt{L}-\sqrt{\mu}}{\sqrt{L}+\sqrt{\mu}}\right)^2\|x^k - x^{k-1}\|_2^2 \\ &= -\frac{\rho^2\mu}{2}\|x^k - y^k\|_2^2 - \sqrt{\frac{\mu}{L}}\frac{\rho^2 L}{2}\|x^k - y^k\|_2^2 = -\frac{\rho^2 L}{2}\left(\frac{\mu}{L} + \sqrt{\frac{\mu}{L}}\right)\|x^k - y^k\|_2^2 \le 0. \end{aligned}$$

Таким образом, $V_{k+1} \le \rho^2 V_k$. Тогда $V_K \le \rho^{2K} V_0$.

Заметим, что $f(x^K) - f^* \leq V_K$. Покажем, что $V_0 \leq L\|x^0 - x^*\|_2^2$, используя предположение $x^* = 0$, а также свойство гладкости (Л2.3):

$$V_0 = f(x^0) - f^* + \frac{L}{2}\|x^0 - \rho^2 x^{-1}\|_2^2 = f(x^0) - f^* + \frac{L}{2}(1-\rho^2)^2\|x^0\|_2^2$$
$$\leq \langle \nabla f(x^*), x^0 - x^* \rangle + \frac{L}{2}\|x^0 - x^*\|_2^2 + \frac{L}{2}\frac{\mu}{L}\|x^0 - x^*\|_2^2 \leq L\|x_0 - x^*\|_2^2.$$

Резюмируя, получаем:

$$f(x^K) - f^* \leq \left(1 - \sqrt{\frac{\mu}{L}}\right)^K \cdot L\|x^0 - x^*\|_2^2.$$

Найдем оценку на число итераций:

$$f(x^K) - f^* \leq \left(1 - \sqrt{\frac{\mu}{L}}\right)^K L\|x^0 - x^*\|_2^2 \leq \exp\left(-\sqrt{\frac{\mu}{L}} \cdot K\right) L\|x^0 - x^*\|_2^2.$$

Мы хотим гарантировать, что

$$f(x^K) - f^* \leq \exp\left(-\sqrt{\frac{\mu}{L}} \cdot K\right) L\|x^0 - x^*\|_2^2 \leq \varepsilon.$$

Тогда логарифируем и получаем

$$K \geq \sqrt{\frac{L}{\mu}} \log\left(\frac{L\|x^0 - x^*\|_2^2}{\varepsilon}\right).$$

■

Таким образом, для сильно выпуклого случая ускоренный градиентный метод действительно показывает превосходство над методом градиентного спуска. Однако,

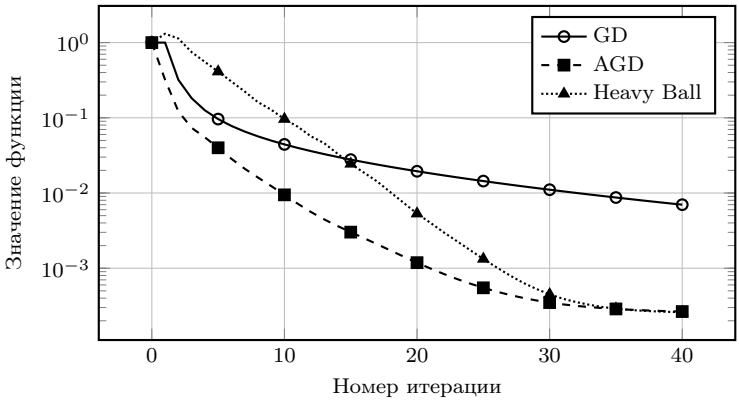


Рис. Л4.5: Сравнение методов GD (Алгоритм Л3.1), Heavy Ball (Алгоритм Л4.1) и AGD (Алгоритм Л4.2) при оптимальном подборе параметров на квадратичной задаче.

что же можно сказать про просто выпуклый случай? Для него формулируется следующая теорема о сходимости.

Теорема Л4.3 (Теорема 6 из [13])**.** Пусть задача безусловной оптимизации (Л3.1) с L-гладкой, выпуклой целевой функцией f решается с помощью ускоренного градиентного метода (Алгоритм Л4.2). Тогда при $\gamma_k = \gamma \le \frac{1}{L}$, $\tau_k = \frac{k-1}{k+2}$ справедлива следующая оценка сходимости:

$$f\big(x^K\big) - f^* \le \frac{2\big\|x^0 - x^*\big\|_2^2}{\gamma(K+1)^2}.$$

Доказательство. Обозначим $a_k = \frac{k+1}{2}, k \ge 0$ и $a_0 = 0$, тогда

$$\tau_k = \frac{k-1}{k+2} = \frac{\frac{k+1}{2} - 1}{\frac{k+2}{2}} = \frac{a_k - 1}{a_{k+1}}. \tag{Л4.1}$$

Введём вектор $p^k = \big(a_k - 1\big)\big(x^k - x^{k-1}\big)$. Из формулы обновления y_k выразим:

$$x^{k-1} = \frac{\big(1+\tau_k\big)x^k - y^k}{\tau_k}.$$

Подставим это в формулу для p^k:

$$p^k = \big(a_k - 1\big)\left(x^k - \frac{\big(1+\tau_k\big)x^k - y^k}{\tau_k}\right) = a_{k+1}\big(y^k - x^k\big).$$

Последнее верно в силу (Л4.1). Пользуясь формулой итерации метода и (Л4.1), имеем:

$$\begin{aligned} p^{k+1} + x^{k+1} &= \big(a_{k+1} - 1\big)\big(x^{k+1} - x^k\big) + x^{k+1} \\ &= a_{k+1}x^{k+1} - \big(a_{k+1} - 1\big)x^k \\ &= a_{k+1}\big(y^k - \gamma\nabla f\big(y^k\big)\big) - \big(a_{k+1} - 1\big)x^k \\ &= a_{k+1}\big(y^k - x^k\big) + x^k - a_{k+1}\gamma\nabla f\big(y^k\big) \\ &= p^k + x^k - a_{k+1}\gamma\nabla f\big(y^k\big). \end{aligned}$$

Оценим $\big\|p^{k+1} + x^{k+1} - x^*\big\|_2^2$, где x^* — одна из точек минимума:

$$\begin{aligned} \big\|p^{k+1} + x^{k+1} - x^*\big\|_2^2 &= \big\|p^k + x^k - x^* - a_{k+1}\gamma\nabla f\big(y^k\big)\big\|_2^2 \\ &= \big\|p^k + x^k - x^*\big\|_2^2 - 2a_{k+1}\gamma\big\langle p^k + x^k - x^*, \nabla f\big(y^k\big)\big\rangle \\ &\quad + a_{k+1}^2\gamma^2\big\|\nabla f\big(y^k\big)\big\|_2^2. \end{aligned}$$

Разобьём скалярное произведение:

$$\begin{aligned} \big\langle p^k + x^k - x^*, \nabla f\big(y^k\big)\big\rangle &= \big\langle a_{k+1}(y^k - x^k) + x^k - x^*, \nabla f\big(y^k\big)\big\rangle \\ &= \big(a_{k+1} - 1\big)\big\langle y^k - x^k, \nabla f\big(y^k\big)\big\rangle + \big\langle y^k - x^*, \nabla f\big(y^k\big)\big\rangle. \end{aligned}$$

Применим выпуклость в точках y^k, x^* и y^k, x^k:

$$\begin{aligned} \big\langle y^k - x^*, \nabla f\big(y^k\big)\big\rangle &\ge f\big(y^k\big) - f^*, \\ \big\langle y^k - x^k, \nabla f\big(y^k\big)\big\rangle &\ge f\big(y^k\big) - f\big(x^k\big). \end{aligned}$$

Тогда:

$$\begin{aligned}\langle p^k + x^k - x^*, \nabla f(y^k)\rangle &\geq (a_{k+1}-1)(f(y^k)-f(x^k)) + (f(y^k)-f^*)\\ &= a_{k+1}(f(y^k)-f^*) - (a_{k+1}-1)(f(x^k)-f^*)\\ &\geq a_{k+1}(f(y^k)-f^*) - \frac{a_k^2}{a_{k+1}}(f(x^k)-f^*).\end{aligned}$$

Последний переход получен благодаря неравенству $(a_{k+1}-1)a_{k+1} \leq a_k^2$, которое проверяется прямой подстановкой определения a_k. Теперь получим соотношение между $f(y^k)$ и $f(x^{k+1})$, применив свойство выпуклых гладких функций (Л2.3):

$$\begin{aligned}f(x^{k+1}) &\leq f(y^k) + \langle \nabla f(y^k), x^{k+1} - y^k\rangle + \frac{L}{2}\|x^{k+1}-y^k\|_2^2\\ &= f(y^k) + \gamma\left(-1+\frac{\gamma L}{2}\right)\|\nabla f(y^k)\|_2^2.\end{aligned}$$

Или, переставив слагаемые:

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

Подставляем в неравенство со скалярным произведением слева:

$$\begin{aligned}\langle p^k + x^k - x^*, \nabla f(y^k)\rangle \geq\ & a_{k+1}(f(x^{k+1})-f^*) - \frac{a_k^2}{a_{k+1}}(f(x^k)-f^*)\\ & - a_{k+1}\gamma\left(-1+\frac{\gamma L}{2}\right)\|\nabla f(y^k)\|_2^2.\end{aligned}$$

Подставим в оценку $\|p^{k+1} + x^{k+1} - x^*\|_2^2$:

$$\begin{aligned}\|p^{k+1} + x^{k+1} - x^*\|_2^2 &\leq \|p^k + x^k - x^*\|_2^2\\ &\quad - 2a_{k+1}^2\gamma(f(x^{k+1})-f^*) + 2a_k^2\gamma(f(x^k)-f^*)\\ &\quad + a_{k+1}^2\gamma^2(\gamma L - 2)\|\nabla f(y^k)\|_2^2 + a_{k+1}^2\gamma^2\|\nabla f(y^k)\|_2^2\\ &\leq \|p^k + x^k - x^*\|_2^2\\ &\quad + 2a_k^2\gamma(f(x^k)-f^*) - 2\gamma a_{k+1}^2(f(x^{k+1})-f^*)\\ &\quad + a_{k+1}^2\gamma^2(\gamma L - 1)\|\nabla f(y^k)\|_2^2\\ &\leq \|p^k + x^k - x^*\|_2^2\\ &\quad + 2a_k^2\gamma(f(x^k)-f^*) - 2a_{k+1}^2\gamma(f(x^{k+1})-f^*).\end{aligned}$$

Слагаемое с градиентом оценено сверху нулем, поскольку $\gamma \leq \frac{1}{L}$. Введём функцию Ляпунова:

$$V_k := \|p_k + x^k - x^*\|_2^2 + 2a_k^2\gamma(f(x^k)-f^*).$$

Будем считать, что γ постоянна, тогда получаем, что $V_{k+1} \leq V_k$ — то есть, функция убывает по k. Тогда:

$$2a_K^2\gamma(f(x^K)-f^*) \leq V_K \leq V_0 = \|x^0 - x^*\|_2^2,$$

откуда

$$f(x^K) - f^* \leq \frac{\|x^0-x^*\|_2^2}{2a_K^2\gamma} = \frac{2\|x^0-x^*\|_2^2}{\gamma(K+1)^2}.$$

■

Утверждение Л4.1. Пусть задача безусловной оптимизации (Л3.1) с L-гладкой, выпуклой функцией f решается ускоренным градиентным методом (Алгоритм Л4.2) с параметрами $\gamma_k = \frac{1}{L}$, $\tau_k = \frac{k-1}{k+2}$. Тогда выполняется следующая оценка сходимости:

$$f\left(x^K\right) - f^* \leq \frac{2L\left\|x^0 - x^*\right\|_2^2}{\left(K+1\right)^2}.$$

Более того, чтобы достичь точности ε по функции, т.е. $f\left(x^K\right) - f^* \leq \varepsilon$, достаточно выполнить

$$K = \mathcal{O}\left(\sqrt{\frac{L\left\|x^0 - x^*\right\|_2^2}{\varepsilon}}\right) \text{ итераций.}$$

Доказательство. Подставим $\gamma = \frac{1}{L}$ в оценку из Теоремы Л4.3:

$$f\left(x^K\right) - f^* \leq \frac{2\left\|x^0 - x^*\right\|_2^2}{\gamma\left(K+1\right)^2} = \frac{2L\left\|x^0 - x^*\right\|_2^2}{\left(K+1\right)^2}.$$

Чтобы получить оценку на число итераций, потребуем:

$$\frac{2L\left\|x^0 - x^*\right\|_2^2}{\left(K+1\right)^2} \leq \varepsilon.$$

Откуда:

$$(K+1)^2 \geq \frac{2L\left\|x^0 - x^*\right\|_2^2}{\varepsilon} \implies K \geq \sqrt{\frac{2L\left\|x^0 - x^*\right\|_2^2}{\varepsilon}} - 1.$$

■

Снова вернемся к классу L-гладких и μ-сильно выпуклых задач. Хочется понять, существует ли метод их решения лучше, чем ускоренный градиентный метод. Для этого прибегнем к построению нижних оценок для класса методов, где лежит рассматриваемый метод. Поскольку мы говорим об ускоренном градиентном методе, то целесообразно говорить о методах первого порядка, то есть, имеющих доступ к градиенту целевой функции. Будем рассматривать следующий класс алгоритмов:

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

где линейной оболочкой множества S называется множество всех линейных комбинаций элементов S:

$$\operatorname{span} S = \bigcup_{k \in \mathbb{N}} \left\{ \sum_{i=1}^{k} \alpha_i x_i \;\middle|\; \alpha_i \in \mathbb{R},\ x_i \in S, i = \overline{1,k} \right\}.$$

Выбор такого класса достаточно логичен: на итерации $k+1$ нам известны все предыдущие точки x^i, $i \leq k$, а также градиенты в них $\nabla f\left(x^i\right)$, $i \leq k$, а строить новую точку x^{k+1} мы будем как их линейную комбинацию. Но если мы предыдущие точки конструировали как линейные комбинации точек и градиентов, то все точки выражаются только через x^0 и градиенты $\nabla f\left(x^i\right), i \leq k$. Выходит, что и x^{k+1} выражается через них же. Отсюда получается, что градиентный спуск (Алгоритм Л3.1), метод тяжёлого шарика (Алгоритм Л4.1) и ускоренный градиентный метод (Алгоритм Л4.2) лежат в описанном классе алгоритмов.

Сформулируем теорему о нижних оценках для этого класса.

Теорема Л4.4. Пусть задача безусловной оптимизации (Л3.1) с L-гладкой, μ-сильно выпуклой функцией f решается методом первого порядка. Тогда для достижения точности ε по аргументу $\left(\left\|x^K - x^*\right\|_2 \leq \varepsilon\right)$ потребуется

$$\Omega\left(\sqrt{\frac{L}{\mu}}\log\frac{\left\|x^0 - x^*\right\|_2}{\varepsilon}\right) \text{ оракульных вызовов.}$$

Доказательство. Сконструируем «плохую» функцию и покажем, что любой метод из описанного класса решает её «достаточно долго».
Построим квадратичную функцию

$$f(x) = \frac{L-\mu}{8}\langle x, Ax\rangle - \frac{L-\mu}{4}\langle e_1, x\rangle + \frac{\mu}{2}\|x\|_2^2,$$

где матрица размерности $d = 2K$ задана следующим образом:

$$A = \begin{pmatrix} 2 & -1 & 0 & \dots & 0 \\ -1 & 2 & -1 & \dots & 0 \\ 0 & -1 & 2 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & 2 \end{pmatrix}.$$

Градиент $f(x)$ будет равен:

$$\nabla f(x) = \frac{L-\mu}{4}Ax - \frac{L-\mu}{4}e_1 + \mu x.$$

Проверим, что f будет L-гладкой:

$$\left\|\frac{L-\mu}{4}A(x-y) + \mu(x-y)\right\|_2 \leq \frac{L-\mu}{4}\|A\|_2\|x-y\|_2 + \mu\|x-y\|_2 \leq L\|x-y\|_2,$$

спектральная норма матрицы A ограничена 4, так как $A - 4I$ отрицательно определена.
Проверим, что f будет μ-сильно выпуклой:

$$\begin{aligned}\frac{L-\mu}{8}\langle y, Ay\rangle - \frac{L-\mu}{4}\langle e_1, y\rangle + \frac{\mu}{2}\|y\|_2^2 \geq \frac{L-\mu}{8}\langle x, Ax\rangle - \frac{L-\mu}{4}\langle e_1, x\rangle + \frac{\mu}{2}\|x\|_2^2 \\ + \left\langle \frac{L-\mu}{4}Ax - \frac{L-\mu}{4}e_1 + \mu x, y - x\right\rangle \\ + \frac{\mu}{2}\|y-x\|_2^2.\end{aligned}$$

Преобразуем выражение:

$$\frac{L-\mu}{8}\left(\langle y, Ay\rangle + \langle x, Ax\rangle - 2\langle y, Ax\rangle\right) \geq \frac{\mu}{2}\|y-x\|_2^2 - \frac{\mu}{2}\|y\|_2^2 - \frac{\mu}{2}\|x\|_2^2 + \mu\langle x, y\rangle = 0.$$

В итоге получаем, что $\langle (y-x), A(y-x)\rangle \geq 0$ — это верно, так как A положительно определена.

Из условия оптимальности — равенства градиента нулю, находим координаты оптимума:

$$\frac{L+\mu}{2}x_1^* - \frac{L-\mu}{4}x_2^* = \frac{L-\mu}{4},$$
$$-\frac{L-\mu}{4}x_{i-1}^* + \frac{L+\mu}{2}x_i^* - \frac{L-\mu}{4}x_{i+1}^* = 0,\ i = \overline{2, d-1},$$
$$-\frac{L-\mu}{4}x_{d-1}^* + \frac{L+\mu}{2}x_d^* = 0.$$

Введём ограничения $x_0^* = 1$, а $x_{d+1}^* = 0$, тогда:

$$-\big(L-\mu\big)x_{i-1}^* + 2\big(L+\mu\big)x_i^* - \big(L-\mu\big)x_{i+1}^* = 0,\ i = \overline{1, d}.$$

Решим рекурренту, для этого запишем характеристическое уравнение:

$$\lambda^2 - 2\frac{L+\mu}{L-\mu}\lambda + 1 = 0 \implies \lambda_{1,2} = \frac{L+\mu}{L-\mu} \pm \sqrt{\left(\frac{L+\mu}{L-\mu}\right)^2 - 1} = \frac{L \pm 2\sqrt{L\mu} + \mu}{L-\mu}.$$

Тогда решение представимо в виде:

$$x_i^* = c_1\left(\frac{\sqrt{L}+\sqrt{\mu}}{\sqrt{L}-\sqrt{\mu}}\right)^i + c_2\left(\frac{\sqrt{L}-\sqrt{\mu}}{\sqrt{L}+\sqrt{\mu}}\right)^i.$$

Учтем ограничения на x_0^* и x_{d+1}^*:

$$c_1 + c_2 = 1$$
$$c_1\left(\frac{\sqrt{L}+\sqrt{\mu}}{\sqrt{L}-\sqrt{\mu}}\right)^{d+1} + c_2\left(\frac{\sqrt{L}-\sqrt{\mu}}{\sqrt{L}+\sqrt{\mu}}\right)^{d+1} = 0$$

Введем обозначение

$$q = \frac{\sqrt{L}-\sqrt{\mu}}{\sqrt{L}+\sqrt{\mu}},$$

тогда

$$x_i^* = -\frac{q^{2d+2}}{1-q^{2d+2}}\frac{1}{q^i} + \frac{1}{1-q^{2d+2}}q^i$$

Рассмотрим начальную точку $x^0 = (0, 0, \ldots, 0)^\top$. Поскольку мы строим нижнюю оценку, то стартовую точку мы можем задавать на своё усмотрение.
Заметим, что $\nabla f\big(x^0\big) \in \text{span}(e_1)$, поэтому после первого вызова оракула только первая координата может быть ненулевой. После второго – первые две и так далее. То есть $\nabla f\big(x^k\big) \in \text{span}(e_1, e_2, \ldots, e_{k+1})$, а $x^{k+1} \in \text{span}(e_1, e_2, \ldots, e_{k+1})$. Получается,

$$\big\|x^K - x^*\big\|_2^2 \geq \sum_{i=K+1}^{d} (x_i^*)^2.$$

Таким образом:

$$\begin{aligned}\left\|x^K - x^*\right\|_2^2 &\geq \sum_{i=K+1}^{d} \left(-\frac{q^{2d+2}}{1-q^{2d+2}}\frac{1}{q^i} + \frac{1}{1-q^{2d+2}}q^i\right)^2 \geq \sum_{i=K+1}^{d}\left(q^i - \frac{q^{2d+2}}{q^i}\right)^2 \\ &\geq \sum_{i=K+1}^{d} q^{2i} = \sum_{i=K+1}^{2K} q^{2i} = q^{2K}\sum_{i=1}^{K} q^{2i} = \frac{q^{2K}}{1+q^{2K}}\left(\sum_{i=1}^{K} q^{2i} + q^{2K}\sum_{i=1}^{K} q^{2i}\right) \\ &= \frac{q^{2K}}{1+q^{2K}}\left\|x^0 - x^*\right\|_2^2 \geq \frac{q^{2K}}{2}\left\|x^0 - x^*\right\|_2^2 = \left(\frac{\sqrt{L}-\sqrt{\mu}}{\sqrt{L}+\sqrt{\mu}}\right)^{2K}\frac{\left\|x^0 - x^*\right\|_2^2}{2}.\end{aligned}$$

Воспользуемся неравенством

$$\frac{1-x}{1+x} \geq \exp(-4x),\ x \in \left[0, \frac{1}{2}\right].$$

Покажем, что оно верно, для этого прологарифмируем выражение и найдем производную:

$$-\frac{1}{1-x} - \frac{1}{1+x} + 4 = -\frac{2}{1-x^2} + 4.$$

Заметим, что производная больше 0 на отрезке $\left[0, \frac{1}{2}\right]$ и к тому же неравенство верно в 0, поэтому неравенство верно на отрезке $\left[0, \frac{1}{2}\right]$.
Тогда оценку можно продолжить:

$$\left\|x^K - x^*\right\|_2^2 \geq \exp\left(-\sqrt{\frac{\mu}{L}}\cdot K\right)\frac{\left\|x^0 - x^*\right\|_2^2}{2}$$

Тогда получаем:

$$K \geq \sqrt{\frac{L}{\mu}}\log\frac{\left\|x^0 - x^*\right\|_2^2}{2\varepsilon^2}$$

■

Мы видим, что верхние оценки ускоренного градиентного метода совпали с нижними оценками методов первого порядка на классе L-гладких μ-сильно выпуклых задач. То есть, асимптотически улучшить метод Нестерова нельзя. Отсюда заключаем, что ускоренный градиентный метод является *оптимальным* в классе методов первого порядка на L-гладких μ-сильно выпуклых задачах.

Л4.3 Линейный каплинг

К текущему моменту мы рассмотрели уже три метода первого порядка: градиентный спуск, метод тяжёлого шарика и ускоренный градиентный метод. В градиентном спуске мы двигаемся против направления роста функции. В методе тяжёлого шарика мы добавляем инерцию к нашему движению, а в методе Нестерова мы вычисляем градиент в будущей точке. Однако, класс методов, который мы описали перед теоремой с нижними оценками, позволяет нам конструировать и менее очевидные с точки зрения физики методы. Например, при обновлении точки, мы можем прибавлять градиент в другой точке, или брать выпуклую комбинацию нескольких точек в качестве третьей. Чтобы проиллюстрировать этот подход, рассмотрим один из методов, который тоже получает оптимальные оценки в сильно выпуклом случае.

Алгоритм Л4.3 Линейный каплинг: внутренний цикл

Вход: стартовые точки $x^0 = y^0 = z^0 \in \mathbb{R}^d$, размеры шагов $\{\gamma_k\}_{k=0} > 0$ и $\{\eta_k\}_{k=0} > 0$, моментумы $\{\tau_k\}_{k=0} \in [0, 1]$, количество итераций K

1: **for** $k = 0, 1, \ldots, K-1$ **do**
2: $\quad y^{k+1} = x^k - \eta_k \nabla f(x^k)$
3: $\quad z^{k+1} = z^k - \gamma_k \nabla f(x^k)$
4: $\quad x^{k+1} = \tau_k z^{k+1} + (1 - \tau_k) y^{k+1}$
5: **end for**

Выход: $\frac{1}{K} \sum_{k=0}^{K-1} x^k$

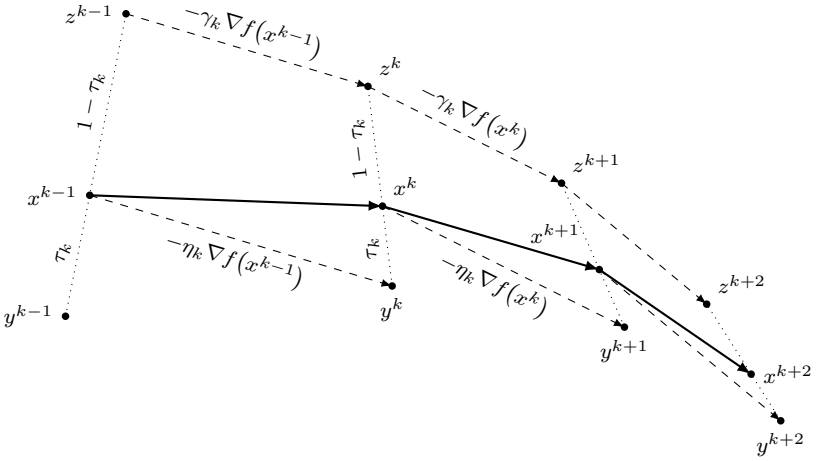


Рис. Л4.6: Три шага линейного каплинга. Пунктир — градиентные шаги для y и z вдоль градиента в точке x, сплошные линии — вычисление x^k как выпуклой комбинации.

Разберем, что происходит на итерации линейного каплинга. Точка y^{k+1} получается шагом градиентного спуска из x^k. Точка z^{k+1} обновляется через градиентный шаг из z^k вдоль направления $\nabla f(x^k)$. В каком-то смысле она живет сама по себе, однако, двигается согласно градиенту в точке x^k. Саму же точку x^{k+1} мы получаем как выпуклую комбинацию y^{k+1} и z^{k+1}. Если попытаться интерпретировать, что происходит в методе, то можно сказать, что точки y^k отвечают за классические шаги против градиента, а z^k добавляют инерцию.

Важным замечанием является, что метод линейного каплинга отличается от предыдущих еще и выходом: раньше мы всегда возвращали точку с последней итерации, теперь возвращаем среднее арифметическое всех точек. В теореме о сходимости будем смотреть за значением функции в средней точке.

Теорема Л4.5. Пусть задача безусловной оптимизации (Л3.1) с L-гладкой, выпуклой целевой функцией f решается с помощью линейного каплинга (Алгоритм Л4.3). Тогда при $\eta < \frac{2}{L}$, $\frac{1-\tau}{\tau} = \frac{\gamma}{\eta(2-L\eta)}$ справедлива следующая оценка

сходимости:

$$f\left(\frac{1}{K}\sum_{k=0}^{K-1}x^k\right)-f^*\le\frac{\|x^0-x^*\|_2^2}{2\gamma K}+\frac{\gamma\big(f(x^0)-f^*\big)}{\eta\big(2-L\eta\big)K}.$$

Доказательство. Воспользуемся линией 3 Алгоритма Л4.3:

$$\begin{aligned}\|z^{k+1}-x^*\|_2^2&=\|z^k-\gamma\nabla f(x^k)-x^*\|_2^2\\&=\|z^k-x^*\|_2^2-2\gamma\langle\nabla f(x^k),z^k-x^*\rangle+\gamma^2\|\nabla f(x^k)\|_2^2\\&=\|z^k-x^*\|_2^2-2\gamma\langle\nabla f(x^k),x^k-x^*\rangle\\&\quad-2\gamma\langle\nabla f(x^k),z^k-x^k\rangle+\gamma^2\|\nabla f(x^k)\|_2^2.\end{aligned}\tag{Л4.2}$$

Оценим $\|\nabla f(x^k)\|_2^2$, для этого применим свойство гладкой функции из Теоремы Л2.4:

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

Воспользуемся линией 2 Алгоритма Л4.3:

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

Если $\eta_k\le\frac{2}{L}$, то $\big(1-\frac{L\eta}{2}\big)>0$. Получаем оценку на норму градиента в квадрате:

$$\|\nabla f(x^k)\|_2^2\le\frac{2}{\eta(2-L\eta)}(f(x^k)-f(y^{k+1})).\tag{Л4.3}$$

Подставим (Л4.3) в (Л4.2):

$$\begin{aligned}\|z^{k+1}-x^*\|_2^2\le\|z^k-z^*\|_2^2&-2\gamma\langle\nabla f(x^k),x^k-x^*\rangle\\&+\frac{2\gamma^2}{\eta(2-L\eta)}\big(f(x^k)-f(y^{k+1})\big)+2\gamma\langle\nabla f(x^k),x^k-z^k\rangle.\end{aligned}\tag{Л4.4}$$

Теперь оценим $\langle\nabla f(x^k),z^k-x^k\rangle$, для этого задействуем линию 4 Алгоритма Л4.3:

$$\langle\nabla f(x^k),x^k-z^k\rangle=\left\langle\nabla f(x^k),x^k-\frac{1}{\tau}\big(x^k-(1-\tau)y^k\big)\right\rangle=\frac{1-\tau}{\tau}\langle\nabla f(x^k),y^k-x^k\rangle.$$

Воспользуемся выпуклостью функции f (Определение Л2.3):

$$\langle\nabla f(x^k),x^k-z^k\rangle\le\frac{1-\tau}{\tau}\big(f(y^k)-f(x^k)\big).\tag{Л4.5}$$

Объединяя оценки (Л4.4) и (Л4.5)

$$\begin{aligned}\|z^{k+1}-x^*\|_2^2\le\|z^k-x^*\|_2^2&-2\gamma\langle\nabla f(x^k),x^k-x^*\rangle\\&+\frac{2\gamma^2}{\eta(2-L\eta)}\big(f(x^k)-f(y^{k+1})\big)+2\gamma\frac{1-\tau}{\tau}\big(f(y^k)-f(x^k)\big).\end{aligned}$$

Подберем τ так, что $\frac{1-\tau}{\tau} = \frac{\gamma}{\eta\left(2-L\eta\right)}$, тогда

$$\left\|z^{k+1} - x^*\right\|_2^2 \leq \left\|z^k - x^*\right\|_2^2 - 2\gamma\left\langle \nabla f\left(x^k\right), x^k - x^*\right\rangle + \frac{2\gamma^2}{\eta\left(2-L\eta\right)}\left(f\left(y^k\right) - f\left(y^{k+1}\right)\right).$$

Сделаем перестановку:

$$2\gamma\left\langle \nabla f\left(x^k\right), x^k - x^*\right\rangle \leq \left\|z^k - x^*\right\|_2^2 - \left\|z^{k+1} - x^*\right\|_2^2 + \frac{2\gamma^2}{\eta\left(2-L\eta\right)}\left(f\left(y^k\right) - f\left(y^{k+1}\right)\right),$$

а потом снова применяем выпуклость:

$$2\gamma\left(f\left(x^k\right) - f\left(x^*\right)\right) \leq \left\|z^k - x^*\right\|_2^2 - \left\|z^{k+1} - x^*\right\|_2^2 + \frac{2\gamma^2}{\eta\left(2-L\eta_k\right)}\left(f\left(y^k\right) - f\left(y^{k+1}\right)\right).$$

Суммируем по всем k от 0 до $K-1$, а потом делим на K:

$$\begin{aligned}
\frac{2\gamma}{K}\sum_{k=0}^{K-1}\left(f\left(x^k\right) - f^*\right) &\leq \frac{1}{K}\sum_{k=0}^{K-1}\left(\left\|z^k - x^*\right\|_2^2 - \left\|z^{k+1} - x^*\right\|_2^2\right) \\
&\quad + \frac{2\gamma^2}{\eta\left(2-L\eta\right)K}\sum_{k=0}^{K-1}\left(f\left(y^k\right) - f\left(y^{k+1}\right)\right) \\
&= \frac{1}{K}\left(\left\|z^0 - x^*\right\|_2^2 - \left\|z^K - x^*\right\|_2^2\right) + \frac{2\gamma^2}{\eta\left(2-L\eta\right)K}\left(f\left(y^0\right) - f\left(y^K\right)\right) \\
&\leq \frac{\left\|z^0 - x^*\right\|_2^2}{K} + \frac{2\gamma^2\left(f\left(y^0\right) - f^*\right)}{\eta\left(2-L\eta\right)K}.
\end{aligned}$$

Учитывая инициализацию $x^0 = y^0 = z^0$ и применяя неравенство Йенсена для выпуклой функции f (0.1), имеем

$$f\left(\frac{1}{K}\sum_{k=0}^{K-1} x^k\right) - f^* \leq \frac{\left\|x^0 - x^*\right\|_2^2}{2\gamma K} + \frac{\gamma\left(f\left(x^0\right) - f^*\right)}{\eta\left(2-L\eta\right)K}.$$

■

Казалось бы, полученный результат не является оптимальным: в выпуклом случае мы получили сублинейную скорость сходимости $\mathcal{O}(1/K)$, как у обычного градиентного спуска. У метода Нестерова мы получили $\mathcal{O}\left(1/K^2\right)$ (Теорема Л4.3). Дальше мы рассмотрим технику, которая позволит получить оптимальную скорость сходимость в сильно выпуклом случае — рестарты линейного каплинга.

Алгоритм Л4.4 Линейный каплинг: рестарты

Вход: стартовая точка $x^0 \in \mathbb{R}^d$, количество итераций T
1: **for** $t = 0, 1, \ldots, T-1$ **do**
2: Запустить Алгоритм Л4.3 из x^t с параметрами γ, η, τ на K итераций
3: Получить x^{t+1} на выходе Алгоритма Л4.3
4: **end for**
Выход: x^T

Появляется логичный вопрос: зачем нам делать внешний цикл, вся суть которого перезапускать метод с прошлого выхода внутреннего метода. Чем это отличается от одного цикла на $T \cdot K$ итераций? Ответ стоит искать в том, что мы получаем на выход из внутреннего цикла, а именно, среднюю точку. И теперь уже становится понятно отличие одного цикла от двух вложенных. Более того, с техникой рестартов удается получить хорошие оценки сходимости. Стоит заметить, что результат прошлой теоремы справедлив для выпуклых функций, а следующую теорему мы формулируем для сильно выпуклых задач.

Теорема Л4.6. Пусть задача безусловной оптимизации (Л3.1) с L-гладкой, μ-сильно выпуклой целевой функцией f решается с помощью рестартов линейного каплинга (Алгоритм Л4.4). Тогда при $\gamma = \frac{1}{\sqrt{\mu L}}$, $\eta = \frac{1}{L}$, $\frac{1-\tau}{\tau} = \frac{\gamma}{\eta\big(2-L\eta\big)}$, $K = \sqrt{\frac{16L}{\mu}}$, чтобы добиться точности ε по функции $\big(f\big(x^T\big) - f^* \leq \varepsilon\big)$, необходимо

$$\mathcal{O}\left(\sqrt{\frac{L}{\mu}} \log \frac{f\big(x^0\big) - f^*}{\varepsilon}\right) \text{ оракульных вызовов.}$$

Доказательство. Применим к результату Теоремы Л4.5 сильную выпуклость (Определение Л2.4), учитывая, что $\nabla f\big(x^*\big) = 0$:

$$\begin{aligned} f\left(\frac{1}{K}\sum_{k=0}^{K-1} x^k\right) - f^* &\leq \frac{f\big(x^0\big) - f^*}{\mu\gamma K} + \frac{\gamma\big(f\big(x^0\big) - f^*\big)}{\eta\big(2-L\eta\big)K} \\ &= \left(\frac{1}{\mu\gamma K} + \frac{\gamma}{\eta\big(2-L\eta\big)K}\right)\big(f\big(x^0\big) - f^*\big). \end{aligned}$$

Подберем η оптимально, чтобы минимизировать правую часть оценки выше: $\eta = \frac{1}{L}$. Тогда

$$f\left(\frac{1}{K}\sum_{k=0}^{K-1} x^k\right) - f^* \leq \left(\frac{1}{\mu\gamma K} + \frac{\gamma L}{K}\right)\big(f\big(x^0\big) - f^*\big).$$

Подберем теперь оптимальное $\gamma = \frac{1}{\sqrt{\mu L}}$:

$$f\left(\frac{1}{K}\sum_{k=0}^{K-1} x^k\right) - f^* \leq \sqrt{\frac{4L}{\mu K^2}}\big(f\big(x^0\big) - f^*\big).$$

Положим $K = \sqrt{\frac{16L}{\mu}}$ и получим:

$$f\left(\frac{1}{K}\sum_{k=0}^{K-1} x^k\right) - f^* \leq \frac{1}{2}\big(f\big(x^0\big) - f^*\big).$$

Получается, что за один запуск линейного каплинга на K итераций, мы гарантированно уменьшаем расстояние до решения по функции в 2 раза. А значит для одной итерации Алгоритма Л4.4 имеем

$$f\big(x^{t+1}\big) - f^* \leq \frac{1}{2}\big(f\big(x^t\big) - f^*\big).$$

Запустим теперь рекурсию по t:

$$f(x^T) - f^* \leq \frac{1}{2^T}\left(f(x^0) - f^*\right).$$

Отсюда легко выразить число итерация Алгоритма Л4.4 для достижения точности ε по функции:

$$T \geq \log\left(\frac{f(x^0) - f^*}{\varepsilon}\right).$$

В данном случае оракульная сложность Алгоритма Л4.4 не совпадает с итерационной. Оракульную сложность можно получить следующим образом:

$$K \cdot T = \mathcal{O}\left(\sqrt{\frac{L}{\mu}}\log\frac{f(x^0) - f^*}{\varepsilon}\right).$$

■

Л4.4 Catalyst

В методе Нестерова мы получили ускорение благодаря добавлению инерции и вычислению градиента в будущей точке. Однако, не все методы первого порядка имеют ускоренные модификации. Давайте модифицируем саму целевую функцию, чтобы получить ускорение для всех методов, обращающихся к градиенту функции. Это позволяет сделать общая схема ускорения Catalyst, предложенная в 2015 году [11].

Алгоритм Л4.5 Catalyst

Вход: стартовые точки $x^0 = y^0 \in \mathbb{R}^d$, коэффициент регуляризацииκ, параметры α_0 и $q = \frac{\mu}{\mu+\kappa}$, последовательность $\{\varepsilon_k\}_{k\geq 0}$, метод оптимизации $\mathcal{M}$

1: **while** не выполнен критерий остановки $\mathcal{T}$ **do**

2: Найти приближенное решение следующей задачи, используя $\mathcal{M}$:

$$x^k \approx \underset{x\in\mathbb{R}^d}{\operatorname{argmin}} \quad g_k(x) = f(x) + \frac{\kappa}{2}\|x - y^{k-1}\|_2^2$$
$$\text{s.t.} \quad g_k(x^k) - g_k^* \leq \varepsilon_k.$$

3: Выбрать $\alpha_k \in (0,1)$ из уравнения $\alpha_k^2 = (1-\alpha_k)\alpha_{k-1}^2 + q\alpha_k$

4: $\beta_k = \frac{\alpha_{k-1}(1-\alpha_{k-1})}{\alpha_{k-1}^2+\alpha_k}$

5: $y^k = x^k + \beta_k(x^k - x^{k-1})$

6: **end while**

Выход: x_k

Проанализируем этот метод. Эта схема напоминает метод рестартов: есть внешний цикл, который запускает внутренний метод $\mathcal{M}$. Но дальше идут различия. На каждой итерации внешнего цикла отрешиваются разные задачи. К исходной целевой функции $f(x)$ добавляется регуляризатор $\frac{\kappa}{2}\|x - y^{k-1}\|_2^2$, который «тянет» точку x к y^{k-1}. А затем точка y^k обновляется наподобие тому, как это было в ускоренном градиентом методе, то есть, с моментумом. Можно сказать, что моментум мы вынесли в виде регуляризатора в модифицированную целевую функцию. Эта идея позволяет, например, ускорить метод градиентного спуска до скорости метода Нестерова. Далее сформулируем конкретные теоремы из оригинальной статьи.

Теорема Л4.7. Пусть задача безусловной оптимизации (Л3.1) с L-гладкой, μ-сильно выпуклой целевой функцией f решается с помощью схемы Catalyst (Алгоритм Л4.5). Тогда при $\alpha_0 = \sqrt{q}$ и последовательности

$$\varepsilon_k = \frac{2}{9}\big(f(x^0) - f^*\big)\big(1-\rho\big)^k,\ \rho \le \sqrt{\kappa}$$

справедлива следующая оценка сходимости:

$$f(x^k) - f^* \le C\big(1-\rho\big)^{k+1}\big(f(x^0) - f^*\big),\ C = \frac{8}{\big(\sqrt{q}-\rho\big)^2}.$$

В таком виде теорема не делает никаких предположений о том методе, который используется для решения подзадач. Мы лишь можем наблюдать за последовательностью x^k, генерируемой алгоритмом. Для того, чтобы понять, насколько дорого обходится работа метода целиком, необходимо сделать предположение о скорости сходимости внутреннего метода $\mathcal{M}$.

Предположение Л4.1. Пусть выполняются условия теоремы Л4.7, а метод $\mathcal{M}$ генерирует последовательность $\{z^t\}_{t\ge 0}$ при минимизации функции g_k с линейной скоростью сходимости:

$$g_k\big(z^t\big) - g_k^* \le A\big(1-\tau_{\mathcal{M}}\big)^t\big(g_k\big(z^0\big) - g_k^*\big).$$

При этом предположении и при $z_0 = x^{k-1}$, точность решения ε_k достигается за $T_{\mathcal{M}} = \tilde{\mathcal{O}}\big(1/\tau_{\mathcal{M}}\big)$ итераций внутреннего метода. Теперь мы можем оценить точность решения и константу сходимости всей схемы Catalyst. Итак, после общего количества итераций метода $\mathcal{M}$, а именно $s = kT_{\mathcal{M}}$, получаем

$$f\Big(x^{\frac{s}{T_{\mathcal{M}}}}\Big) - f^* \le C\big(1-\rho\big)^{\frac{s}{T_{\mathcal{M}}}}\big(f(x^0) - f^*\big) \le C\left(1 - \frac{\rho}{T_{\mathcal{M}}}\right)^s\big(f(x^0) - f^*\big).$$

Таким образом, константа $\tau_{\mathcal{A}}$ итогового метода равна $\tau_{\mathcal{A}} = \frac{\rho}{T_{\mathcal{M}}} = \tilde{\mathcal{O}}\Big(\tau_{\mathcal{M}}\sqrt{\frac{\mu}{\mu+\kappa}}\Big)$. При этом $\tau_{\mathcal{M}}$ обычно также зависит от κ, так что κ подбирается так, чтобы максимизировать $\tau_{\mathcal{M}}/\sqrt{\mu+\kappa}$.

Сформулируем аналогичную теорему и предположение для выпуклого случая, а затем рассмотрим, какое ускорение удается получить для изученных методов.

Теорема Л4.8. Пусть задача безусловной оптимизации (Л3.1) с L-гладкой выпуклой целевой функцией f решается с помощью схемы Catalyst (Алгоритм Л4.5). Тогда при $\alpha_0 = \frac{\sqrt{5}-1}{2}$ и последовательности

$$\varepsilon_k = \frac{2}{9}\big(f(x^0) - f^*\big)(k+2)^{-4-\eta},\ \eta > 0$$

справедлива следующая оценка сходимости:

$$f(x^k) - f^* \le \frac{8}{\big(k+2\big)^2}\left(\left(1+\frac{2}{\eta}\right)^2\big(f(x^0) - f^*\big) + \frac{\kappa}{2}\big\|x^0 - x^*\big\|^2\right).$$

Снова потребуется сделать предположение о сходимости внутреннего алгоритма для решения подзадачи.

Предположение Л4.2. Пусть выполняются условия теоремы Л4.8, а метод $\mathcal{M}$ генерирует последовательность $\{z^t\}_{t\ge 0}$ при минимизации функции g_k с линейной скоростью сходимости:

$$g_k\big(z^t\big) - g_k^* \le A\big(1-\tau_{\mathcal{M}}\big)^t\big(g_k\big(z^0\big) - g_k^*\big).$$

Мы потребовали линейной скорости сходимости, но рассматриваем выпуклый случай. Почему такое возможно? На самом деле, выпуклой является функция f, а когда мы к ней добавили регуляризатор $\frac{\kappa}{2}\|x - y^{k-1}\|_2^2$, мы сделали задачу сильно выпуклой с константой $\mu = \kappa$. А для такой задачи уже многие градиентные методы будут иметь линейную скорость сходимости.
При этом предположении и при $z_0 = x^{k-1}$, точность решения ε_k достигается за $T_{\mathcal{M}} = \tilde{\mathcal{O}}(1/\tau_{\mathcal{M}})$ итераций внутреннего метода. Если мы захотим более точную оценку, то для k-й подзадачи число итераций метода $\mathcal{M}$ составляет $T_{\mathcal{M}} \log(k+2)$ из-за полиномиального убывания точности ε_k.
Теперь мы можем оценить точность решения и константу сходимости всей схемы Catalyst. После общего количества итераций внутреннего метода $s = kT_{\mathcal{M}} \log(k+2)$, получаем

$$f\left(x^{\frac{s}{T_{\mathcal{M}} \log(k+2)}}\right) - f^* \leq \frac{8T_{\mathcal{M}}^2 \log^2(s)}{s^2}\left(\left(1 + \frac{2}{\eta}\right)^2 (f(x^0) - f^*) + \frac{\kappa}{2}\|x^0 - x^*\|^2\right).$$

Если метод $\mathcal{M}$ является методом первого порядка, то данная оценка является почти оптимальной (near-optimal), с точностью до логарифмического множителя, по сравнению с оптимальной скоростью $\mathcal{O}(1/s^2)$, что может быть разумной платой за универсальность схемы Catalyst.
Рассмотрим как метод Catalyst позволяет ускорить метод градиентного спуска, сравним со скоростью метода Нестерова. Все асимптотики написаны для количества итераций, необходимых для достижения ε-точности решения по функции ($f(x) - f^* \leq \varepsilon$). Параметры были подобраны оптимально. Как мы видим, Catalyst ускоряет метод гра-

	Без Catalyst		**Catalyst**	
	$\mu > 0$	$\mu = 0$	$\mu > 0$	$\mu = 0$
GD	$\mathcal{O}\left(\frac{L}{\mu}\log\left(\frac{1}{\varepsilon}\right)\right)$	$\mathcal{O}\left(\frac{L}{\varepsilon}\right)$	$\tilde{\mathcal{O}}\left(\sqrt{\frac{L}{\mu}}\log\left(\frac{1}{\varepsilon}\right)\right)$	$\tilde{\mathcal{O}}\left(\frac{L}{\sqrt{\varepsilon}}\right)$
AGD	$\mathcal{O}\left(\sqrt{\frac{L}{\mu}}\log\left(\frac{1}{\varepsilon}\right)\right)$	$\mathcal{O}\left(\frac{L}{\sqrt{\varepsilon}}\right)$	нет ускорения	

Таблица 1: Сравнение методов GD (Алгоритм Л3.1), AGD (Алгоритм Л4.2) с использованием Catalyst и без.

диентного спуска так, что получаются асимптотики аналогичные методу Нестерова с точностью до логарифмических факторов. Метод Нестерова дополнительно ускорить не удается. Объясняется это тем, что даже с применением схемы ускорения Catalyst ускоренный градиентный метод остается методом первого порядка. А как мы знаем из верхних и нижних оценок, он является оптимальным среди методов первого порядка на классе гладких сильно выпуклых задач.

Л5 Стохастический градиентный спуск

Стохастическая оптимизация возникает в тех случаях, когда точный градиент функции либо невозможно вычислить, либо его вычисление слишком дорого. В таких ситуациях мы используем приближённые градиенты, построенные на основе случайной информации о функции.
Придём к этой идее через пример из Параграфа Л1, а именно через оптимизацию в машинном обучении Л1.2.5. В этом примере мы приходили к задаче минимизации эмпирического риска, которую можно записать как

$$\min_{x\in\mathbb{R}^d}\left[f(x) := \frac{1}{n}\sum_{i=1}^{n} l\big(g\big(x,\xi_{a,i}\big),\xi_{b,i}\big)\right],$$

где введены обозначения:

- $\xi_i = (\xi_{i,a},\xi_{i,b}), i = \overline{1,n}$ — элементы выборки: $\xi_{i,a}$ — объект (картинка, текст) и $\xi_{i,b}$ — метка (ответ) на i-м объекте,
- g — модель машинного обучения (линейная модель, нейросеть), принимает на вход объект и настраиваемые веса x,
- l — функция потерь (штрафует модель за несовпадения с реальной меткой ξ_b).

Такую постановку называют *оффлайн* (данные фиксированы, а не поступают в режиме реального времени).
В оффлайн постановке можно считать градиент целевой функции f, но в машинном обучении часто не используют полные честные градиенты. Основная причина — их дорого и долго считать, поэтому вместо полного градиента вызывают градиент по случайному элементу выборки: $\nabla f\big(x,\xi_i\big)$, где i генерируется независимо и равномерно из $\overline{1,n}$.
Теперь продемонстрируем чуть более общий подход. До этого мы рассматривали задачи, где все вычисления были детерминистическими:

$$\min_{x\in\mathbb{R}^d} f\big(x\big).$$

А сейчас пусть целевая функция — это математическое ожидание от случайной величины ξ из распределения $\mathcal{D}$:

$$\min_{x\in\mathbb{R}^d}\left[f\big(x\big) := \mathbb{E}_{\xi\sim\mathcal{D}}\left[f\big(x,\xi\big)\right]\right] \tag{Л5.1}$$

Цель: «подстроиться» под природу, и чтобы потери модели в среднем по всему распределению были наименьшими, то есть модель наилучшим образом аппроксимировала зависимость ξ_b от ξ_a.
Проблема: Функция f, а также градиенты и более старшие производные не считаются, так как не знаем $\mathcal{D}$, да даже если и знаем, интеграл часто не взять так просто.
Решение проблемы: Возникает потребность в методе, который может оперировать с $\nabla f\big(x,\xi\big)$ (градиентом по конкретному объекту из распределения данных). То есть хотим работать в *онлайн* режиме: поступают сэмплы, мы их обрабатываем (можем считать градиент).

Замечание Л5.1. Всякого рода рандомизированные подходы, основанные на сэмплировании, называются *Монте-Карло подходами*.
Оффлайн постановка — это Монте-Карло аппроксимация исходного интеграла (математического ожидания). С ростом количества элементов в выборке, аппроксимация через конечную сумму будет стремиться к реальному интегралу (при определенных предположениях).

Л5.1 Стохастический градиентный спуск

Идея — модифицировать градиентный спуск, заменив полный градиент на стохастический.

Алгоритм Л5.1 Стохастический градиентный спуск

Вход: стартовая точка $x^0 \in \mathbb{R}^d$, размеры шагов $\{\gamma_k\}_{k=0} > 0$, количество итераций K

1: **for** $k = 0, 1, \ldots, K-1$ **do**
2: Сгенерировать независимо $\xi^k \sim \mathcal{D}$
3: $x^{k+1} = x^k - \gamma_k \nabla f\left(x^k, \xi^k\right)$
4: **end for**

Выход: x^K

Для доказательства сходимости напомним определение условного математического ожидания и закон полного математического ожидания.

Определение Л5.1. *Условное математическое ожидание:*

$$\mathbb{E}\left[\cdot \,\middle|\, x^k\right] = \mathbb{E}\left[\cdot \,|\, \mathcal{F}_k\right],$$

где $\mathcal{F}_k$ — σ-алгебра, порождённая $x^0, \xi^0, \ldots, \xi^{k-1}$.

Суть — фиксируем всю случайность, которая произошла до k итерации и ожидаем только по случайности, которая осталась размороженной. Такое математическое ожидание дает на выходе что-то, зависящее от случайных величин $x^0, \xi^0, \ldots, \xi^{k-1}$.

Теорема Л5.1 (**Закон полного математического ожидания** (Утверждение G из части 4 Параграфа 7 из [30])).

$$\mathbb{E}\left[\mathbb{E}\left[X \,|\, Y\right]\right] = \mathbb{E}\left[X\right].$$

Перейдём к описанию предположений, которые будут использованы при доказательстве сходимости. На функцию f оставляем предположения об L-гладкости и μ-сильной выпуклости. Ранее градиент функции был детерминированным. В контексте стохастического градиентного спуска, когда стохастический градиент — это случайная величина, нам потребуются два предположения на стохастический градиент.

Предположение Л5.1 (*Несмещённость стохастического градиента*). Положим $\{\mathcal{F}\}_{k \geq 0}$ — семейство вложенных σ-алгебр. Случайная величина x^0 — $\mathcal{F}_0$-измерима, и для любых $x^k \in \mathbb{R}^d$ и $k \in \mathbb{N}$ случайная величина $\nabla f\left(x^k, \xi^k\right)$ является $\mathcal{F}_k$-измеримой и выполняется

$$\mathbb{E}\left[\nabla f\left(x^k, \xi^k\right) \,\middle|\, x^{k-1}\right] = \nabla f\left(x^k\right).$$

Предположение Л5.2 (*Ограниченность дисперсии стохастического градиента*). Существует $\sigma^2 \in \mathbb{R}_+$, такое, что для любых $x^k \in \mathbb{R}^d$ и $k \in \mathbb{N}$

$$\mathbb{E}\left[\left\|\nabla f\left(x^k, \xi^k\right) - \nabla f\left(x^k\right)\right\|_2^2 \,\middle|\, x^{k-1}\right] \leq \sigma^2.$$

Теперь можем доказывать сходимость. Начнем как обычно с сильно выпуклого случая.

Теорема Л5.2. Пусть задача безусловной стохастической оптимизации (Л5.1) с L-гладкой, μ-сильно выпуклой целевой функцией f решается с помощью стохастического градиентного спуска (Алгоритм Л5.1) с $\gamma_k = \gamma \leq \frac{1}{L}$ в предположениях несмещённости (Предположение Л5.1) и ограниченности дисперсии стохастического градиента (Предположение Л5.2). Тогда справедлива следующая оценка сходимости

$$\mathbb{E}\left[\left\|x^K - x^*\right\|_2^2\right] \leq \left(1 - \gamma\mu\right)^K \mathbb{E}\left[\left\|x^0 - x^*\right\|_2^2\right] + \frac{\gamma\sigma^2}{\mu}. \tag{Л5.2}$$

Доказательство. Начинаем, как и раньше:

$$\left\|x^{k+1} - x^*\right\|_2^2 = \left\|x^k - x^*\right\|_2^2 - 2\gamma\langle\nabla f(x^k, \xi^k), x^k - x^*\rangle + \gamma^2\left\|\nabla f(x^k, \xi^k)\right\|_2^2.$$

Вспомним, что мы знаем предположения про функцию f, а в выражении выше её нет (есть только её стохастический градиент). Значит, надо как-то перейти к этой функции в неравенствах. Для этого берем условное математическое ожидание по случайности только на итерации k (важно, что x^k — это неслучайная величина относительно условного математического ожидания по разбиению $\mathcal{F}_k$, случайными здесь будут только величины x^{k+1} и градиент $\nabla f(x^k, \xi^k)$):

$$\begin{aligned}\mathbb{E}\left[\left\|x^{k+1} - x^*\right\|_2^2 \,\middle|\, x^k\right] &= \left\|x^k - x^*\right\|_2^2 - 2\gamma\mathbb{E}\left[\left\langle\nabla f(x^k, \xi^k), x^k - x^*\right\rangle \,\middle|\, x^k\right] \\ &\quad + \gamma^2\mathbb{E}\left[\left\|\nabla f(x^k, \xi^k)\right\|_2^2 \,\middle|\, x^k\right].\end{aligned} \tag{Л5.3}$$

Рассмотрим $\mathbb{E}\left[\left\langle\nabla f(x^k, \xi^k), x^k - x^*\right\rangle \,\middle|\, x^k\right]$ и воспользуемся линейностью условного математического ожидания, а также Предположением Л5.1:

$$\begin{aligned}\mathbb{E}\left[\left\langle\nabla f(x^k, \xi^k), x^k - x^*\right\rangle \,\middle|\, x^k\right] &= \left\langle\mathbb{E}\left[\nabla f(x^k, \xi^k) \,\middle|\, x^k\right], x^k - x^*\right\rangle \\ &= \left\langle\nabla f(x^k), x^k - x^*\right\rangle.\end{aligned}$$

Рассмотрим $\mathbb{E}\left[\left\|\nabla f(x^k, \xi^k)\right\|_2^2 \,\middle|\, x^k\right]$, добавим и вычтем градиент, после чего раскроем квадрат нормы:

$$\begin{aligned}\mathbb{E}\left[\left\|\nabla f(x^k, \xi^k)\right\|_2^2 \,\middle|\, x^k\right] &= \mathbb{E}\left[\left\|\nabla f(x^k, \xi^k) - \nabla f(x^k) + \nabla f(x^k)\right\|_2^2 \,\middle|\, x^k\right] \\ &= \mathbb{E}\left[\left\|\nabla f(x^k, \xi^k) - \nabla f(x^k)\right\|_2^2 \,\middle|\, x^k\right] + \mathbb{E}\left[\left\|\nabla f(x^k)\right\|_2^2 \,\middle|\, x^k\right] \\ &\quad + 2\mathbb{E}\left[\left\langle\nabla f(x^k, \xi^k) - \nabla f(x^k), \nabla f(x^k)\right\rangle \,\middle|\, x^k\right]\end{aligned} \tag{Л5.4}$$

Теперь мы можем воспользоваться Предположением Л5.1 и Предположением Л5.2:

$$\begin{aligned}\mathbb{E}\left[\left\|\nabla f(x^k, \xi^k)\right\|_2^2 \,\middle|\, x^k\right] &= \mathbb{E}\left[\left\|\nabla f(x^k, \xi^k) - \nabla f(x^k)\right\|_2^2 \,\middle|\, x^k\right] + \left\|\nabla f(x^k)\right\|_2^2 \\ &\quad + 2\left\langle\mathbb{E}\left[\nabla f(x^k, \xi^k) \,\middle|\, x^k\right] - \nabla f(x^k), \nabla f(x^k)\right\rangle \\ &\leq \sigma^2 + \left\|\nabla f(x^k)\right\|_2^2.\end{aligned}$$

Объединим все промежуточные шаги, подставим в (Л5.3) и получим:

$$\mathbb{E}\left[\left\|x^{k+1} - x^*\right\|_2^2 \,\middle|\, x^k\right] \leq \left\|x^k - x^*\right\|_2^2 - 2\gamma\langle\nabla f(x^k), x^k - x^*\rangle + \gamma^2\left\|\nabla f(x^k)\right\|_2^2 + \gamma^2\sigma^2.$$

Выглядит похоже на то, что было в градиентном спуске, только появляется $\gamma^2\sigma^2$. Поэтому далее воспользуемся L-гладкостью (Теорема Л2.5) и μ-сильной выпуклостью (Определение Л2.4):

$$\begin{aligned}\mathbb{E}\left[\left\|x^{k+1}-x^*\right\|_2^2 \,\middle|\, x^k\right] &\le \left\|x^k-x^*\right\|_2^2 - 2\gamma\Big(f\big(x^k\big)-f^* + \frac{\mu}{2}\left\|x^k-x^*\right\|_2^2\Big) \\ &\quad + 2\gamma^2 L\big(f\big(x^k\big)-f^*\big) + \gamma^2\sigma^2 \\ &= \big(1-\gamma\mu\big)\left\|x^k-x^*\right\|_2^2 - 2\gamma\big(1-\gamma L\big)\big(f\big(x^k\big)-f^*\big) + \gamma^2\sigma^2.\end{aligned}$$

При выборе $\gamma \le \frac{1}{L}$ слагаемое $-2\gamma\big(1-\gamma L\big)\big(f\big(x^k\big)-f^*\big)$ отрицательно и от него можно избавиться:

$$\mathbb{E}\left[\left\|x^{k+1}-x^*\right\|_2^2 \,\middle|\, x^k\right] \le \big(1-\gamma\mu\big)\left\|x^k-x^*\right\|_2^2 + \gamma^2\sigma^2.$$

Сейчас запустить рекурсию мешает то, что слева и справа стоят случайные величины, причем оценок на $\left\|x^k-x^*\right\|_2$ у нас нет. Чтобы перейти к неслучайным величинам, в неравенстве выше берем полное математическое ожидание и применяем результат Теоремы Л5.1:

$$\mathbb{E}\left[\left\|x^{k+1}-x^*\right\|_2^2\right] \le \big(1-\gamma\mu\big)\mathbb{E}\left[\left\|x^k-x^*\right\|_2^2\right] + \gamma^2\sigma^2.$$

Разворачиваем рекурсию:

$$\begin{aligned}\mathbb{E}\left[\left\|x^{K}-x^*\right\|_2^2\right] &\le \big(1-\gamma\mu\big)\mathbb{E}\left[\left\|x^{K-1}-x^*\right\|_2^2\right] + \gamma^2\sigma^2 \\ &\le \big(1-\gamma\mu\big)^2\mathbb{E}\left[\left\|x^{K-2}-x^*\right\|_2^2\right] + \big(1-\gamma\mu\big)\gamma^2\sigma^2 + \gamma^2\sigma^2 \\ &\le \big(1-\gamma\mu\big)^K\mathbb{E}\left[\left\|x^0-x^*\right\|_2^2\right] + \gamma^2\sigma^2\sum_{i=0}^{K-1}\big(1-\gamma\mu\big)^i.\end{aligned}$$

Второе слагаемое — сумма первых k членов геометрической прогрессии, оценим её:

$$\sum_{i=0}^{K-1}\big(1-\gamma\mu\big)^i \le \sum_{i=0}^{+\infty}\big(1-\gamma\mu\big)^i = \frac{1}{\gamma\mu}.$$

После подстановки в выражение выше:

$$\mathbb{E}\left[\left\|x^{k}-x^*\right\|_2^2\right] \le \big(1-\gamma\mu\big)^k\mathbb{E}\left[\left\|x^0-x^*\right\|_2^2\right] + \frac{\gamma\sigma^2}{\mu}.$$

■

Этот результат похож на линейную сходимость к решению для градиентного спуска. Однако, добавляется второе слагаемое — это некоторая неточность, зависящая от γ, σ и μ. Её метод преодолеть не может и начинает осциллировать, больше не приближаясь к решению. В таком случае говорят, что имеет место сходимость к окрестности решения x^*.

Теперь рассмотрим выпуклый случай.

Теорема Л5.3. Пусть задача безусловной стохастической оптимизации (Л5.1) с L-гладкой, выпуклой целевой функцией f решается с помощью стохастического градиентного спуска (Алгоритм Л5.1) с $\gamma_k = \gamma \le \frac{1}{2L}$ в предположениях несмещённости (Предположение Л5.1) и ограниченности дисперсии стохастического градиента (Предположение Л5.2). Тогда справедлива следующая оценка сходи-

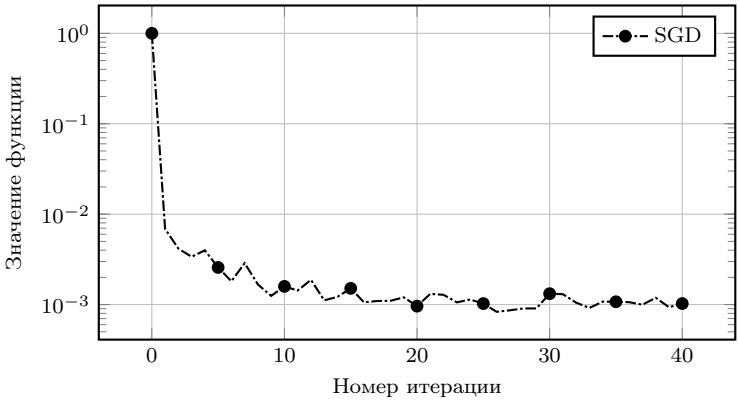


Рис. Л5.1: Характерная «дрожь» сходимости стохастического градиентного спуска.

мости

$$\mathbb{E}\left[f\left(\frac{1}{K}\sum_{k=0}^{K-1}x^k\right)-f^*\right]\leq\frac{\|x^0-x^*\|_2^2}{\gamma K}+\gamma\sigma^2.$$

Доказательство. Как и раньше, разложим квадрат нормы:

$$\|x^{k+1}-x^*\|_2^2=\|x^k-x^*\|_2^2-2\gamma\langle\nabla f(x^k,\xi^k),x^k-x^*\rangle+\gamma^2\|\nabla f(x^k,\xi^k)\|_2^2.$$

Берём условное ожидание по $\mathcal{F}_k$ и пользуемся несмещённостью стохастического градиента (Предположение Л5.1):

$$\mathbb{E}\left[\|x^{k+1}-x^*\|_2^2\,\middle|\,x^k\right]=\|x^k-x^*\|_2^2-2\gamma\langle\nabla f(x^k),x^k-x^*\rangle+\gamma^2\mathbb{E}\left[\|\nabla f(x^k,\xi^k)\|_2^2\,\middle|\,x^k\right].$$

Аналогично сильно выпуклому случаю, воспользуемся оценкой на второй момент градиента (Л5.4):

$$\mathbb{E}\left[\|\nabla f(x^k,\xi^k)\|_2^2\,\middle|\,x^k\right]\leq\sigma^2+\|\nabla f(x^k)\|_2^2.$$

Подставим в неравенство:

$$\mathbb{E}\left[\|x^{k+1}-x^*\|_2^2\,\middle|\,x^k\right]\leq\|x^k-x^*\|_2^2-2\gamma\langle\nabla f(x^k),x^k-x^*\rangle+\gamma^2\|\nabla f(x^k)\|_2^2+\gamma^2\sigma^2.$$

Вспомним определение выпуклой функции (Определение Л2.3):

$$\langle\nabla f(x^k),x^k-x^*\rangle\geq f(x^k)-f^*.$$

и неравенство для L-гладкой функции (неравенство (Л2.4) из Теоремы Л2.5):

$$\|\nabla f(x^k)\|_2^2\leq 2L(f(x^k)-f^*).$$

Продолжим основное неравенство:

$$\mathbb{E}\left[\left\|x^{k+1}-x^*\right\|_2^2 \,\middle|\, x^k\right] \le \left\|x^k-x^*\right\|_2^2 - 2\gamma\big(f(x^k)-f^*\big) + 2\gamma^2 L\big(f(x^k)-f^*\big) + \gamma^2\sigma^2$$
$$= \left\|x^k-x^*\right\|_2^2 - 2\gamma(1-\gamma L)\big(f(x^k)-f^*\big) + \gamma^2\sigma^2.$$

Возьмём $\gamma \le \frac{1}{2L}$, чтобы $1-\gamma L \ge \frac{1}{2}$, тогда:

$$\mathbb{E}\left[\left\|x^{k+1}-x^*\right\|_2^2 \,\middle|\, x^k\right] \le \left\|x^k-x^*\right\|_2^2 - \gamma\big(f(x^k)-f^*\big) + \gamma^2\sigma^2.$$

Берем полное математическое ожидание и суммируем по всем k от 0 до K:

$$\mathbb{E}\left[\left\|x^K-x^*\right\|_2^2\right] \le \left\|x^0-x^*\right\|_2^2 - \gamma\sum_{k=0}^{K-1}\mathbb{E}\left[f(x^k)-f^*\right] + \gamma^2\sigma^2 K.$$

Так как $\mathbb{E}\left[\left\|x^K-x^*\right\|_2^2\right]$ — неотрицательное, то:

$$\sum_{k=0}^{K-1}\mathbb{E}\left[f(x^k)-f^*\right] \le \frac{\left\|x^0-x^*\right\|_2^2}{\gamma} + \gamma\sigma^2 K.$$

Воспользуемся линейностью математического ожидания и неравенством Йенсена (0.1):

$$\mathbb{E}\left[f\left(\frac{1}{K}\sum_{k=0}^{K-1}x^k\right)-f^*\right] \le \mathbb{E}\left[\frac{1}{K}\sum_{k=0}^{K-1}f(x^k)-f^*\right] = \frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}\left[f(x^k)-f^*\right]$$
$$\le \frac{\left\|x^0-x^*\right\|_2^2}{\gamma K} + \gamma\sigma^2.$$

■

Результат теоремы похож на классическую оценку для градиентного спуска из Теоремы Л3.3 с точностью до добавки в виде дисперсии решения $\gamma\sigma^2$.

Утверждение Л5.1. Пусть задача удовлетворяет условиям Теоремы Л5.3 и выбрано значение шага $\gamma_k = \frac{1}{2L}$. Тогда справедлива следующая оценка:

$$\mathbb{E}\left[f\left(\frac{1}{K}\sum_{k=0}^{K-1}x^k\right)-f^*\right] \le \frac{2L\left\|x^0-x^*\right\|_2^2}{K} + \frac{\sigma^2}{2L}.$$

Чтобы добиться точности ε по функции $\left(\mathbb{E}\left[f\left(\frac{1}{K}\sum_{k=0}^{K-1}x^k\right)-f^*\right] \le \varepsilon\right)$, необходимо

$$K = \mathcal{O}\left(\frac{L\left\|x^0-x^*\right\|_2^2}{\varepsilon - \frac{\sigma^2}{2L}}\right) \text{ итераций.}$$

Доказательство. Подставим $\gamma = \frac{1}{2L}$ в оценку из Теоремы Л5.3 и потребуем, чтобы правая часть была не больше ε:

$$\mathbb{E}\left[f(\bar{x}^K)-f^*\right] \le \frac{\left\|x^0-x^*\right\|_2^2}{\gamma K} + \gamma\sigma^2 = \frac{2L\left\|x^0-x^*\right\|_2^2}{K} + \frac{\sigma^2}{2L} \le \varepsilon.$$

Выразим отсюда K:

$$K \geq \frac{2L\left\|x^0 - x^*\right\|_2^2}{\varepsilon - \frac{\sigma^2}{2L}}.$$

■

Л5.2 Модификации стохастического градиентного спуска

Стохастический градиентный спуск с постоянным шагом обладает существенным недостатком — сходимостью к окрестности решения. Как можно попробовать решить проблемы неточной сходимости? Рассмотрим простые техники для уменьшения дисперсии. Например, в работе [2] проведен анализ для уменьшающегося шага. Получим следствия из Теоремы 1 этой статьи.

Теорема Л5.4. Пусть задача безусловной стохастической оптимизации (Л5.1) с L-гладкой, μ-сильно выпуклой целевой функцией f решается с помощью стохастического градиентного спуска (Алгоритм Л5.1) с $\gamma_k = \frac{2}{\mu(k+1)}$ в предположениях несмещённости (Предположение Л5.1) и ограниченности дисперсии стохастического градиента (Предположение Л5.2). Тогда справедлива следующая оценка сходимости

$$\mathbb{E}\left[\left\|x^K - x^*\right\|_2^2\right] \leq \frac{\exp\left(8L^2/\mu^2\right)}{K^2}\left(\mathbb{E}\left[\left\|x^0 - x^*\right\|_2^2\right] + \frac{\sigma^2}{L^2}\right) + \frac{8\sigma^2 \log K}{\mu^2 K}.$$

При такой стратегии подбора шага дисперсия решения убывает сублинейно, но мы теряем в скорости сходимости (была линейная, стала сублинейная). Так происходит, потому что размер шага слишком быстро уменьшается ($\sim 1/k$). Можно получить оценку и для менее агрессивно убывающего γ_k.

Теорема Л5.5. Пусть задача безусловной стохастической оптимизации (Л5.1) с L-гладкой, μ-сильно выпуклой целевой функцией f решается с помощью стохастического градиентного спуска (Алгоритм Л5.1) с $\gamma_k = \frac{2}{\mu\sqrt{k+1}}$ в предположениях несмещённости (Предположение Л5.1) и ограниченности дисперсии стохастического градиента (Предположение Л5.2). Тогда справедлива следующая оценка сходимости

$$\mathbb{E}\left[\left\|x^K - x^*\right\|_2^2\right] \leq 2K^{16L^2/\mu^2}\exp\left(-\frac{\sqrt{K}}{2}\right)\left(\mathbb{E}\left[\left\|x^0 - x^*\right\|_2^2\right] + \frac{\sigma^2}{L^2}\right) + \frac{8\sigma^2}{\mu^2\sqrt{K}}.$$

Сравнивая оценки для шагов $\gamma_k \sim 1/\sqrt{k}$ с $\gamma_k \sim 1/k$, наблюдаем, что дисперсия решения стала убывать медленнее, но сходимость первого слагаемого стала быстрее, хотя все еще остается сублинейной. Лучшие оценки сходимости были получены в работе [22].

Теорема Л5.6. Пусть задача безусловной стохастической оптимизации (Л5.1) с L-гладкой, μ-сильно выпуклой целевой функцией f решается с помощью стохастического градиентного спуска (Алгоритм Л5.1) в предположениях несмещённости (Предположение Л5.1) и ограниченности дисперсии стохастического гра-

диента (Предположение Л5.2). Тогда при выборе

$$\gamma_k = \begin{cases} \frac{1}{2L}, & k < \lceil \frac{K}{2} \rceil, \\ \frac{2}{\mu\left(\frac{4L}{\mu} + k - \lceil \frac{K}{2} \rceil\right)}, & k \geq \lceil \frac{K}{2} \rceil, \end{cases}$$

$$w_k = \begin{cases} 0, & k < \lceil \frac{K}{2} \rceil, \\ \left(\frac{4L}{\mu} + k - \lceil \frac{K}{2} \rceil\right)^2, & k \geq \lceil \frac{K}{2} \rceil, \end{cases}$$

справедлива следующая оценка сходимости

$$\mathbb{E}\left[f(\bar{x}^K) - f^*\right] + \mu\mathbb{E}\left[\left\|x^{K+1} - x^*\right\|_2^2\right] \leq \min\left\{64LR^2 \exp\left(-\frac{\mu K}{4L}\right) + \frac{36\sigma^2}{\mu K}, \frac{2LR^2}{K} + \frac{2\sigma R}{\sqrt{K}}\right\},$$

где $R = \left\|x^0 - x^*\right\|_2, W_K = \sum_{i=1}^K w_i$, $\bar{x}^K = \frac{1}{W_K}\sum_{i=1}^K w_i x^i$.

Эта оценка сочетает в себе линейную сходимость первого слагаемого и сублинейный спад слагаемого с шумом. Стратегию выбора шагов спуска и весов для усреднения точек можно описать следующим образом: сначала делаем шаги с константным значением $\gamma = \frac{1}{2L}$ и не учитываем эти $\lceil \frac{K}{2} \rceil$ точек при усреднении, а затем берем уменьшающийся шаг, получаем сходимость к более точной окрестности, усредняем эти точки с ненулевыми возрастающими весами.

Другим подходом может быть уменьшение σ с помощью техники батчирования. Идея заключается в замене стохастической оценки градиента по одному сэмплу на среднее арифметическое по подвыборке:

$$\nabla f(x^k, \xi^k) \to \frac{1}{b}\sum_{j \in S^k} \nabla f(x, \xi_j),$$

где S^k — набор индексов из $\overline{1, n}$, $\left|S^k\right| = b$, и все индексы генерируются независимо друг от друга (возможны повторы).

Утверждение Л5.2. Пусть задача удовлетворяет условиям Теоремы Л5.2 и применена техника батчирования по b независимым сэмплам. Тогда справедлива следующая оценка сходимости:

$$\mathbb{E}\left[\left\|x^K - x^*\right\|_2^2\right] \leq (1 - \gamma\mu)^K \mathbb{E}\left[\left\|x^0 - x^*\right\|_2^2\right] + \frac{\gamma\sigma^2}{\mu b}.$$

Доказательство. Посмотрим, что становится со свойствами несмещённости и конечности дисперсии при замене стохастического градиента по 1 сэмплу на среднее значение по батчу размера b.

Несмещённость остается, так как у каждого сэмпла j есть несмещённость стохастического градиента и выполняется линейность математического ожидания:

$$\mathbb{E}\left[\frac{1}{b}\sum_{j \in S^k} \nabla f(x, \xi_j) \,\middle|\, x^k\right] = \nabla f(x).$$

А с дисперсией интереснее:

$$\mathbb{E}\left[\left\|\frac{1}{b}\sum_{j\in S^k}\big(\nabla f(x,\xi_j)-\nabla f(x)\big)\right\|_2^2\,\middle|\,x^k\right]\le\frac{1}{b^2}\mathbb{E}\left[\sum_{j\in S^k}\big\|\big(\nabla f(x,\xi_j)-\nabla f(x)\big)\big\|_2^2\,\middle|\,x^k\right]$$
$$\le\frac{\sigma^2}{b}.$$

Остается лишь подставить оценки в (Л5.2) и получить требуемое. ■

Получается, дисперсию можно уменьшить в b раз, но тогда и вычисление стохастического градиента дорожает в b раз.
Если теперь объединить батчирование и стратегию подбора шага из Теоремы Л5.6, то получается добиться следующей оценки сходимости:

$$\mathbb{E}\left[\big\|x^K-x^*\big\|_2^2\right]=\mathcal{O}\left(\frac{L}{\mu}\exp\left(-\frac{\mu}{4L}K\right)\big\|x^0-x^*\big\|_2^2+\frac{\sigma^2}{\mu^2bK}\right).$$

Получается линейная сходимость по «детерминистической» части и сублинейная по «стохастической».
Оказывается, этот результат ускоряется. По аналогии с тем, как инерция в алгоритме Нестерова дает ускорение, можно ускорить и стохастический градиентный спуск. Если объединить батчирование и результат работы [1], то получается оценка:

$$\mathbb{E}\left[f(x^{K+1})-f^*\right]=\mathcal{O}\left(\exp\left(-\frac{1}{2}\sqrt{\frac{\mu}{L}}K\right)\big(f(x^0)-f^*\big)+\frac{\sigma^2}{\mu bK}\right).$$

Л5.3 Нижние оценки

Ранее мы уже строили нижние оценки для задач оптимизации. При построении нижних оценок для задач стохастической оптимизации применим знания из математической статистики. Следующий результат был получен в работе [3].

Теорема Л5.7. Пусть задача безусловной стохастической оптимизации (Л5.1) с L-гладкой, μ-сильно выпуклой целевой функцией f решается методом с доступом к стохастическому оракулу первого порядка, для которого выполняются предположения несмещённости (Предположение Л5.1) и ограниченности дисперсии стохастического градиента (Предположение Л5.2). Тогда выполняется неравенство

$$\mathbb{E}\left[\big\|x^K-x^*\big\|_2^2\right]=\Omega\left(\frac{\sigma^2}{\mu^2K}\right).$$

Доказательство. Нам придется сконструировать не только «плохую» функцию, но и описать стохастический оракул. С одной стороны, задача стала сложнее: стало больше свободы. С другой же, можно не выдумывать сложную функцию, а всю «плохость» задачи сосредоточить в том, насколько зашумленные ответы будет выдавать оракул. Так и поступим — возьмём примитивную целевую функцию:

$$f(x)=\frac{\mu}{2}\big\|x-x^*\big\|_2^2,\ x\in\mathbb{R}.$$

Она μ-сильно выпуклая и μ-гладкая. Считаем, что x^* нам не известно, а имеется доступ к зашумленному оракулу первого порядка. На запрос в точке x оракул будет возвращать стохастический градиент

$$\nabla f(x,\xi)=\mu(x-x^*+\xi),\ \xi\sim\mathcal{N}\left(0,\frac{\sigma^2}{\mu^2}\right).$$

Проверим, что выполняются предположения о стохастическом градиенте. Несмещённость:

$$\mathbb{E}\left[\nabla f(x,\xi)\right] = \mathbb{E}\left[\mu(x - x^* + \xi)\right] = \nabla f(x).$$

Ограниченность дисперсия:

$$\mathbb{E}\left[\left\|\nabla f(x,\xi) - \nabla f(x)\right\|_2^2\right] = \mathbb{E}\left[\left\|\mu(x - x^* + \xi) - \mu(x - x^*)\right\|_2^2\right] = \mathbb{E}\left[\mu^2\|\xi\|_2^2\right] = \sigma^2.$$

Таким образом, все требования на функцию и стохастический градиент выполняются. Пусть мы обратились к оракулу в K точках $\{x^k\}_{k=0}^{K-1}$ и получили в ответ значения $y^k = \mu(x^k + \xi^k - x^*)$, где $\xi^k \sim \mathcal{N}(0, \sigma^2/\mu^2)$ независимы. Хотим получить наилучшую оценку на x^*, используя имеющиеся данные. Введём z^k:

$$z^k = x^k - y^k/\mu = x^* - \xi^k.$$

Тогда задачу можно переформулировать следующим образом: по набору значений $z^k \sim \mathcal{N}(x^*, \sigma^2/\mu^2)$ необходимо восстановить константу x^*. Это можно сделать с помощью оценки максимального правдоподобия:

$$\hat{x} = \frac{1}{K}\sum_{k=0}^{K-1} z^k \sim \mathcal{N}\left(x^*, \frac{\sigma^2}{\mu^2 K}\right).$$

Из математической статистики известно, что для квадратичной функции потерь оценка максимального правдоподобия является оптимальной в смысле минимизации максимального эмпирического риска, то есть, получить предсказание лучше оценки максимального правдоподобия невозможно. Из того, какое распределение на $\hat{x}$, мы сразу видим, что оценка является несмещенной, а ее дисперсия

$$\mathbb{E}\left[\left(\hat{x} - x^*\right)^2\right] = \frac{\sigma^2}{\mu^2 K}.$$

Остается заключить, что поскольку такое значение дисперсии достигается для оптимальной оценки, то все остальные будут не лучше, а значит, они лежат в $\Omega\left(\frac{\sigma^2}{\mu^2 K}\right)$. ■

Смысл этой нижней оценки в том, что дисперсия решения убывает не быстрее $1/K$. Следующий логичный вопрос: являются ли какие-то из ранее полученных оценок оптимальными? Обратимся к результатам Теоремы Л5.6. Из нее можно получить, что $\mathbb{E}\left[\left\|x^K - x^*\right\|_2^2\right] = \mathcal{O}\left(\frac{\sigma^2}{\mu^2 K}\right)$, то есть, стохастический градиентный спуск с описанным подбором шага является оптимальным в классе методов с доступом к зашумленному оракулу первого порядка.

Л6 Метод сопряжённых градиентов

Метод сопряжённых градиентов изначально был разработан для решения систем линейных уравнений:

$$Ax = b, \tag{Л6.1}$$

где $A \in \mathbb{S}^d_{++}$, $b \in \mathbb{R}^d$.
Эту систему можно переписать в виде задачи безусловной оптимизации:

$$\min_{x \in \mathbb{R}^d} \left\{ f(x) = \frac{1}{2}\langle x, Ax \rangle - \langle b, x \rangle \right\}. \tag{Л6.2}$$

В самом деле, функция f сильно выпукла, поскольку матрица A положительно определена, а её градиент $\nabla f(x) = Ax - b$ зануляется в точке x^* — решении исходной системы (Л6.1).
Далее мы увидим, что метод сопряжённых градиентов можно обобщить за пределы квадратичных функций.

Л6.1 Сопряжённые направления

Ключевым определением в этом параграфе станут сопряжённые направления.

Определение Л6.1. Множество ненулевых векторов $\{p^i\}_{i=0}^{n-1}$ будем называть *сопряжённым* относительно матрицы $A \in \mathbb{S}^d_{++}$, если для любых $i \neq j \in \overline{0, n-1}$ следует

$$\langle p^i, Ap^j \rangle = 0.$$

Свойство сопряжённости во многом похоже на ортогональность. В частности, для сопряжённых векторов выполняется свойство линейной независимости.

Теорема Л6.1. Множество сопряжённых векторов $\{p^i\}_{i=0}^{n-1}$ относительно матрицы $A \in \mathbb{S}^d_{++}$ является линейно независимым.

Доказательство. Пойдём от противного: пусть найдется нетривиальная линейная комбинация сопряжённых векторов, равная нулю. То есть, существует набор не равных одновременно нулю чисел $\alpha_0, \ldots, \alpha_{n-1} \in \mathbb{R} : \sum_{i=0}^{n-1} \alpha_i^2 \neq 0$, для которых:

$$\sum_{i=0}^{n-1} \alpha_i p^i = 0.$$

Домножим на матрицу A и возьмём скалярное произведение с произвольным p^j:

$$\sum_{i=0}^{n-1} \alpha_i \langle p^j, Ap^i \rangle = \alpha_j \langle p^j, Ap^j \rangle = 0.$$

При переходе от суммы к одному слагаемому мы воспользовались определением сопряжённых векторов. Так как $p^j \neq 0$, а матрица A положительно определена, то

$$\langle p^j, Ap^j \rangle > 0.$$

Следовательно, равенство нулю возможно только при $\alpha_j = 0$ для всех $j = \overline{0, n-1}$ — противоречие. ■

Следствием из Теоремы Л6.1 будет полезный факт, что d сопряжённых векторов формируют базис в $\mathbb{R}^d$. Разложим по нему решение x^*:

$$x^* = \sum_{i=0}^{d-1} \lambda_i p^i. \tag{Л6.3}$$

Для нахождения λ_i подействуем матрицей A на обе части (Л6.3) и возьмём скалярное произведение с p^j:

$$\langle p^j, Ax^* \rangle = \sum_{i=0}^{d-1} \lambda_i \langle p^j, Ap^i \rangle = \lambda_j \langle p^j, Ap^j \rangle.$$

При переходе от суммы к одному слагаемому мы опять воспользовались определением сопряжённости векторов.
Так как x^* — решение, то $Ax^* = b$ и получаем, что:

$$\langle p^j, b \rangle = \lambda_j \langle p^j, Ap^j \rangle.$$

Откуда выражается

$$\lambda_j = \frac{\langle p^j, b \rangle}{\langle p^j, Ap^j \rangle}.$$

Л6.2 Метод сопряжённых градиентов

Теперь можно придумать схему обновления x^k для поиска решения:

$$x^{k+1} = x^k + \alpha_k p^k. \tag{Л6.4}$$

Эта схема называется схемой сопряжённых градиентов.
Разложение по сопряжённым направлениям (Л6.3) совпадает с итеративной схемой при $\alpha_k = \lambda_k$ и $x^0 = 0$. А для того, чтобы метод сходился из произвольной точки x^0, потребуется модифицировать выбор коэффициентов α_k.

Теорема Л6.2 (Теорема 5.1 из [18])**.** Пусть задача безусловной оптимизации (Л6.2) решается схемой сопряжённых градиентов (Л6.4). Тогда при

$$\alpha_k = -\frac{\langle r^k, p^k \rangle}{\langle p^k, Ap^k \rangle},$$

где $r^k = Ax^k - b$ — вектор невязки, метод сходится к точному решению x^* за не более чем d итераций.

Доказательство. Как ранее уже обсуждалось, по следствию из Теоремы Л6.1 набор из d сопряжённых векторов образуют базис в $\mathbb{R}^d$. Разложим по нему вектор $x^* - x^0$:

$$x^* - x^0 = \sum_{i=0}^{d-1} \lambda_i p^i.$$

Подействуем матрицей A и возьмём скалярное произведение с p^j:

$$\lambda_j = \frac{\langle p^j, A(x^* - x^0) \rangle}{\langle p^j, Ap^j \rangle}, \ j = \overline{0, d-1}.$$

Принимая во внимание, что $A\big(x^* - x^0\big) = b - Ax^0 = -r^0$, можно записать:

$$x^* = x^0 + \sum_{i=0}^{d-1} \lambda_i p^i = x^0 - \sum_{i=0}^{d-1} \frac{\langle r^0, p^i \rangle}{\langle p^i, Ap^i \rangle} p^i.$$

Чтобы получить результат теоремы, остается показать, что $\langle r^0, p^k \rangle = \langle r^k, p^k \rangle$. Действительно:

$$\langle r^k, p^k \rangle = \langle Ax^k - b, p^k \rangle = \left\langle Ax^0 + \sum_{i=0}^{k-1} A\alpha_i p^i - b, p^k \right\rangle = \langle Ax^0 - b, p^k \rangle = \langle r^0, p^k \rangle.$$

Теперь становится видно, что с выбором коэффициентов $\alpha_k = -\frac{\langle r^k, p^k \rangle}{\langle p^k, Ap^k \rangle}$ схема сопряжённых градиентов (Л6.4) на итерации d достигнет точного решения x^*:

$$x^d = x^0 - \sum_{i=0}^{d-1} \frac{\langle r^i, p^i \rangle}{\langle p^i, Ap^i \rangle} p^i = x^*.$$

При этом, возможна ситуация, когда точное решение достигается и на более ранней итерации. Например, когда $\langle p^k, r^k \rangle = 0$ начиная с некоторого $k < d$. ■

Добавим другую интерпретацию схеме сопряжённых градиентов. Как мы отметили в самом начале, решение системы (Л6.1) эквивалентно решению безусловной задачи оптимизации (Л6.2). Это минимизация квадратичной функции на $\mathbb{R}^d$. Линии её уровня — это эллипсоиды с центром в x^*.
Рассмотрим задачу минимизации функции $f(x) = \frac{1}{2}\langle x, Ax \rangle - \langle b, x \rangle$ из точки x^k вдоль p^k. Если записать в виде задачи оптимизации, то получится

$$\min_{\alpha \in \mathbb{R}} f\big(x^k + \alpha p^k\big).$$

Приравняем производную к нулю:

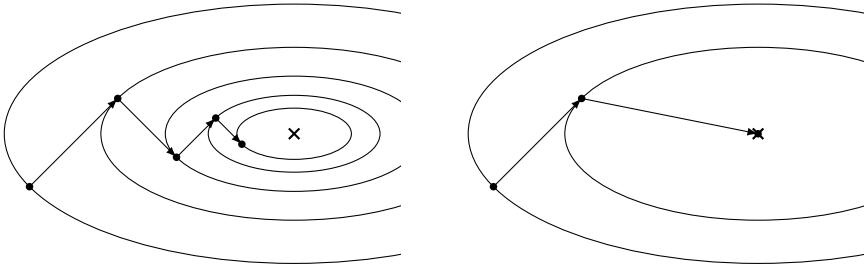
$$\big\langle \nabla f\big(x^k + \alpha^* p^k\big), p^k \big\rangle = \big\langle A\big(x^k + \alpha^* p^k\big) - b, p^k \big\rangle = 0 \implies \alpha^* = \alpha_k = -\frac{\langle Ax^k - b, p^k \rangle}{\langle p^k, Ap^k \rangle}.$$

То есть, шаг метода сопряжённых направлений — это шаг наискорейшего спуска для минимизации f вдоль направления p^k.
Наглядно сравним итерации схемы сопряжённых градиентов и наискорейшего покоординатного спуска. Из этой иллюстрации можно сделать несколько выводов.

- Схема сопряжённых градиентов находит точное решение за d итераций, наискорейший покоординатный спуск не сходится точно.
- Сопряжённые направления не ортогональны в привычном смысле этого смысла слова.

Докажем некоторые полезные факты для схемы сопряжённых градиентов.

Теорема Л6.3 (Теорема 5.2 из [18])**.** Пусть схема сопряжённых градиентов (Л6.4) решает задачу (Л6.2) с выбором коэффициентов $\alpha_k = -\frac{\langle r^k, p^k \rangle}{\langle p^k, Ap^k \rangle}$. Тогда



(a) Наискорейший спуск (несколько шагов). (b) Сопряжённые градиенты (2 шага).

Рис. Л6.1: Сравнение траекторий на квадратичной задаче с $A = \mathrm{diag}(1, 5)$.

1. $\langle r^k, p^j \rangle = 0,\ \forall j < k$;

2. $x^k = \underset{x \in P}{\operatorname{argmin}} f(x)$, где $P = x^0 + \operatorname{span}\{p^0, \ldots, p^{k-1}\}$.

Доказательство. Обратимся к доказательству Теоремы Л6.2. В нем показывалось, что коэффициенты α_k можно записать как

$$\alpha_k = -\frac{\langle r^0, p^k \rangle}{\langle p^k, Ap^k \rangle}.$$

Подставим их и формулу итерации (Л6.4) в скалярное произведение:

$$\begin{aligned}
\langle r^k, p^j \rangle = \langle Ax^k - b, p^i \rangle &= \left\langle A\left(x^0 - \sum_{i=0}^{k-1} \frac{\langle r^0, p^i \rangle}{\langle p^i, Ap^i \rangle} p^i\right) - b, p^j \right\rangle \\
&= \langle Ax^0 - b, p^j \rangle - \sum_{i=0}^{k-1} \frac{\langle r^0, p^i \rangle}{\langle p^i, Ap^i \rangle} \langle p^j, Ap^i \rangle \\
&= \langle r^0, p^j \rangle - \frac{\langle r^0, p^j \rangle}{\langle p^j, Ap^j \rangle} \langle p^j, Ap^j \rangle = 0.
\end{aligned}$$

Покажем равносильность 1 и 2. Введём функцию $\phi : \mathbb{R}^k \to \mathbb{R}$, определённую как

$$\phi(\gamma) = f\big(x^0 + \gamma_0 p^0 + \ldots + \gamma_{k-1} p^{k-1}\big).$$

По сути, это сужение функции f на множество P. Поскольку функция f — сильно выпукла, а все векторы p^i ненулевые, то и композиция f с линейным взаимно однозначным преобразованием из R^k в P, будет сильно выпуклой функцией. Отсюда следует, что у ϕ существует единственный минимум γ^* (Теорема Л2.3). Тогда, пользуясь необходимым и достаточным условием глобального минимума сильно выпуклой функции, получаем:

$$\frac{\partial \phi(\gamma^*)}{\partial \gamma_j} = \big\langle \nabla f\big(x^0 + \gamma_0^* p^0 + \ldots + \gamma_{k-1}^* p^{k-1}\big), p^j \big\rangle = 0.$$

Воспользуемся явным видом f:

$$\nabla f\big(x^0 + \gamma_0^* p^0 + \ldots + \gamma_{k-1}^* p^{k-1}\big) = \nabla f\big(x^k\big) = Ax^k - b = r^k.$$

■

Суть этой теоремы в том, что вектор невязки на k-ой итерации ортогонален всем найденным направлениям. Этот факт нам понадобится позже для ускорения вычислений и доказательства корректности выбора направлений.
Теперь перейдём к способу поиска p^i. Если вспомнить формулу обновления x^k (Л6.4), то можно заметить, что она похожа на градиентный спуск. Поэтому в качестве нулевого направления можно взять антиградиент в стартовой точке:

$$p^0 = -\nabla f(x^0) = -r^0.$$

Множество из одного вектора всегда сопряжено, поэтому ничего дополнительно требовать не надо. Последующие направления надо конструировать так, чтобы они были сопряжены всем уже найденным. Для этого предложим следующую формулу:

$$p^{k+1} = -r^{k+1} + \beta_{k+1} p^k, \tag{Л6.5}$$

где β_{k+1} будем подбирать, требуя сопряженность p^k и p^{k+1}:

$$\langle p^{k+1}, Ap^k \rangle = \langle -r^{k+1} + \beta_{k+1} p^k, Ap^k \rangle = -\langle r^{k+1}, Ap^k \rangle + \beta_{k+1} \langle p^k, Ap^k \rangle = 0.$$

Откуда получаем

$$\beta_{k+1} = \frac{\langle r^{k+1}, Ap^k \rangle}{\langle p^k, Ap^k \rangle}.$$

Применяя этот алгоритм поиска p^i, можем записать метод сопряжённых градиентов.

Алгоритм Л6.1 Метод сопряжённых градиентов (теоретическая версия)

Вход: стартовая точка $x^0 \in \mathbb{R}^d$, стартовая невязка $r^0 = Ax^0 - b$, стартовый сопряжённый вектор $p^0 = -r_0$, количество итераций K

1: **for** $k = 0, 1, \ldots, K-1$ **do**
2: $\quad \alpha_k = -\dfrac{\langle r^k, p^k \rangle}{\langle p^k, Ap^k \rangle}$
3: $\quad x^{k+1} = x^k + \alpha_k p^k$
4: $\quad r^{k+1} = Ax^{k+1} - b$
5: $\quad \beta_{k+1} = \dfrac{\langle r^{k+1}, Ap^k \rangle}{\langle p^k, Ap^k \rangle}$
6: $\quad p^{k+1} = -r^{k+1} + \beta_{k+1} p^k$
7: **end for**

Выход: x^K

Сейчас мы можем гарантировать только сопряженность p^k и p^{k+1}. А по тому, как мы строили метод, нам нужна сопряженность p^{k+1} со всеми $p^i, i \leq k$. Для этого докажем соответствующую теорему.

Теорема Л6.4 (Теорема 5.3 из [18]). Пусть на k-ой итерации метод сопряжённых градиентов (Алгоритм Л6.1) получил точку x^k, которая не является решением x^*. Тогда верны следующие свойства:

1. $\text{span}\{r^0, r^1, \ldots, r^k\} = \text{span}\{r^0, Ar^0, \ldots, A^k r^0\}$;
2. $\text{span}\{p^0, p^1, \ldots, p^k\} = \text{span}\{r^0, Ar^0, \ldots, A^k r^0\}$;
3. $\langle p^k, Ap^i \rangle = 0, \ i = \overline{0, k-1}$;
4. $\langle r^k, r^i \rangle = 0, \ i = \overline{0, k-1}$.

Более того, метод сходится к решению x^* не более чем за d шагов.

Доказательство. Предположим, что утверждения (1), (2) и (3) верны для некоторого k и докажем их для $k+1$. Утверждения тривиально верны для $k=0$, а утверждение (3) верно для $k=1$ из-за выбора β.

Доказательство (1). Сначала покажем, что левая часть включена в правую. По предположению индукции:

$$r^k \in \operatorname{span}\{r^0, Ar^0, \ldots, A^k r^0\},$$
$$p^k \in \operatorname{span}\{r^0, Ar^0, \ldots, A^k r^0\}.$$

Тогда:

$$Ap^k \in \operatorname{span}\{Ar^0, A^2 r^0, \ldots, A^{k+1} r^0\}.$$

По формуле обновления

$$r^{k+1} = Ax^{k+1} - b = A\big(x^k + \alpha_k p^k\big) - b = r^k + \alpha_k A p^k. \tag{Л6.6}$$

получаем:

$$r^{k+1} \in \operatorname{span}\{r^0, Ar^0, \ldots, A^{k+1} r^0\}.$$

Отсюда получаем вложение:

$$\operatorname{span}\{r^0, r^1, \ldots, r^k, r^{k+1}\} \subseteq \operatorname{span}\{r^0, Ar^0, \ldots, A^{k+1} r^0\}.$$

Для доказательства обратного включения воспользуемся (2) и заметим:

$$A^{k+1} r^0 = A\big(A^k r^0\big) \in A \operatorname{span}\{p^0, \ldots, p^k\}.$$

Поскольку по формуле обновления невязки:

$$Ap^i = \frac{r^{i+1} - r^i}{\alpha_i}, \; i = \overline{0,k},$$

то получаем:

$$A^{k+1} r^0 \in \operatorname{span}\{r^0, r^1, \ldots, r^{k+1}\},$$

и, следовательно:

$$\operatorname{span}\{r^0, r^1, \ldots, r^{k+1}\} \supseteq \operatorname{span}\{r^0, Ar^0, \ldots, A^{k+1} r^0\}.$$

Таким образом, утверждение (1) выполнено при $k+1$.

Доказательство (2). Используем формулу для получения нового вектора p^{k+1} (Л6.5) и утверждение (1):

$$\begin{aligned}
\operatorname{span}\{p^0, \ldots, p^k, p^{k+1}\} &= \operatorname{span}\{p^0, \ldots, p^k, r^{k+1}\} \\
&= \operatorname{span}\{r^0, Ar^0, \ldots, A^k r^0, r^{k+1}\} \\
&= \operatorname{span}\{r^0, r^1, \ldots, r^k, r^{k+1}\} \\
&= \operatorname{span}\{r^0, Ar^0, \ldots, A^{k+1} r^0\},
\end{aligned}$$

что и требовалось.

Доказательство (3). Скалярно умножим (Л6.5) на Ap^i при $i = \overline{0,k}$:

$$\big\langle p^{k+1}, Ap^i \big\rangle = -\big\langle r^{k+1}, Ap^i \big\rangle + \beta_{k+1} \big\langle p^k, Ap^i \big\rangle.$$

При $i = k$ правая часть обнуляется по определению β_{k+1}. Покажем это и при $i < k$.

По предположению индукции направления $p^0, \dots, p^k$ — сопряжённые, а значит по Теореме Л6.3:

$$\langle r^{k+1}, p^i \rangle = 0, \ i = \overline{0,k}.$$

Воспользуемся утверждением (2):

$$Ap^i \in A\,\mathrm{span}\{r^0, Ar^0, \dots, A^i r^0\} = \mathrm{span}\{Ar^0, A^2 r^0, \dots, A^{i+1} r^0\} \subset \mathrm{span}\{p^0, \dots, p^{i+1}\}.$$

Следовательно:

$$\langle r^{k+1}, Ap^i \rangle = 0, \ i = \overline{0,k-1}.$$

Кроме того, по предположению индукции для (3):

$$\langle p^k, Ap^i \rangle = 0, \ i = \overline{0,k-1}.$$

Таким образом, $\langle p^{k+1}, Ap^i \rangle = 0$ для всех $i = \overline{0,k}$, что завершает доказательство.
Доказательство (4). Покажем, что невязка r^k ортогональна предыдущим. Так как $p^0, \dots, p^{k-1}$ — сопряжены, то по Теореме Л6.3:

$$\langle r^k, p^i \rangle = 0, \ i = \overline{0,k-1}.$$

Из формулы обновления направления (Л6.5):

$$p^i = -r^i + \beta_i p^{i-1},$$

получаем:

$$r^i \in \mathrm{span}\{p^i, p^{i-1}\}, \ i = \overline{1,k-1}.$$

Следовательно:

$$\langle r^k, r^i \rangle = 0, \ i = \overline{1,k-1}.$$

Кроме того, $p^0 = -r^0 \implies \langle r^k, r^0 \rangle = 0$, то есть, утверждение также доказано.
Все четыре утверждения выполнены, и по Теореме Л6.2, метод сходится к решению x^* не более чем за d шагов. ■

Теперь, когда доказана корректность и скорость сходимости метода до точного решения, можно проанализировать результаты. Итак, для решения системы из d линейных уравнений методу требуется d итераций. На каждой итерации несколько перемножений векторов, а также два перемножения матрицы A на вектор. Последнее наиболее дорого с точки зрения вычислений, каждое умножение порядка d^2 арифметических операций. При выполнении d итераций, получаем $\mathcal{O}(d^3)$ арифметических операций суммарно. А если вспомнить классические алгоритмы решения систем линейных уравнений, например, метод Гаусса, то мы не получили никакого выигрыша по скорости. Тогда появляется вопрос: чем метод сопряжённых градиентов лучше? Ответ заключается в том, что метод сопряжённых градиентов итеративный, то есть, на каждой итерации мы получаем какое-то приближение решения. И зачастую нас устроит неточное, но достаточно хорошее решение. Тогда мы можем сделать не все d итераций, а оборвать цикл раньше.
К тому же, мы на данный момент получили много полезных фактов про сопряжённые направления, векторы невязки и их обновления. Это нам позволит ускорить медленную версию метода в константу раз. Начнем с α_k:

$$\alpha_k = -\frac{\langle r^k, p^k \rangle}{\langle p^k, Ap^k \rangle} = -\frac{\langle r^k, -r^k + \beta_k p^{k-1} \rangle}{\langle p^k, Ap^k \rangle} = \frac{\langle r^k, r^k \rangle}{\langle p^k, Ap^k \rangle}.$$

Также по формуле (Л6.6) можно записать:

$$r^{k+1} = r^k + \alpha_k Ap^k.$$

Остается поработать с β_k. Выражаем $Ap^k = \frac{1}{\alpha_k}(r^{k+1} - r^k)$ и подставляем:

$$\beta_{k+1} = \frac{\langle r^{k+1}, Ap^k \rangle}{\langle p^k, Ap^k \rangle} = \frac{\langle r^{k+1}, r^{k+1} - r^k \rangle}{\langle -r^k + \beta_k p^{k-1}, r^{k+1} - r^k \rangle} = \frac{\langle r^{k+1}, r^{k+1} \rangle}{\langle r^k, r^k \rangle}.$$

Запишем обновленную версию алгоритма.

Алгоритм Л6.2 Метод сопряжённых градиентов

Вход: стартовая точка $x^0 \in \mathbb{R}^d$, стартовая невязка $r^0 = Ax^0 - b$, стартовый сопряжённый вектор $p^0 = -r_0$, количество итераций K

1: **for** $k = 0, 1, \ldots, K-1$ **do**
2: $\quad \alpha_k = \frac{\langle r^k, r^k \rangle}{\langle p^k, Ap^k \rangle}$
3: $\quad x^{k+1} = x^k + \alpha_k p^k$
4: $\quad r^{k+1} = r^k + \alpha_k Ap^k$
5: $\quad \beta_{k+1} = \frac{\langle r^{k+1}, r^{k+1} \rangle}{\langle r^k, r^k \rangle}$
6: $\quad p^{k+1} = -r^{k+1} + \beta_{k+1} p^k$
7: **end for**

Выход: x^K

По итогу сэкономили 1 умножение матрицы на вектор и получили ускорение за счёт константы. Оказывается, что можно записать и более оптимальные верхние оценки для метода.

Определение Л6.2. Пусть даны матрица $A \in \mathbb{S}^d_{++}$ и вектор $x \in \mathbb{R}^d$. Функцию $\|\cdot\|_A$, определённую как

$$\|x\|_A = \sqrt{\langle x, Ax \rangle},$$

будем называть *A-нормой*.

Теорема Л6.5 (Формула (5.32) из [18])**.** Пусть система линейных уравнений (Л6.1) решается с помощью метода сопряжённых градиентов (Алгоритм Л6.2). Тогда справедлива следующая оценка скорости сходимости:

$$\|x^K - x^*\|_A^2 \leq \min_{P_{K-1}} \max_{i=\overline{1,d}} \left[1 + \lambda_i P_{K-1}(\lambda_i)\right]^2 \|x^0 - x^*\|_A^2,$$

где $\lambda_i,\ i = \overline{1,d}$ — собственные значения матрицы A, а минимум берется по P_K — все полиномы степени K.

Доказательство. Точку x^{k+1} мы строим как

$$x^{k+1} = x^0 + \sum_{i=0}^{k} \alpha_i p^i.$$

В Теореме Л6.4 мы показали, что

$$\text{span}\{p^0, \ldots, p^k\} = \text{span}\{r^0, Ar^0 \ldots, A^k r^0\}.$$

Тогда $x^{k+1} \in x^0 + \text{span}\{r^0, Ar^0 \ldots, A^k r^0\}$ и мы можем представить x^{k+1} как линейную комбинацию:

$$x^{k+1} = x^0 + \gamma_0 r^0 + \gamma_1 A r^0 + \ldots + \gamma_k A^k r^0.$$

Можем ввести полином P_k^* степени k. И поскольку A — квадратная матрица, можно взять её в качестве аргумента P_k^*:

$$P_k^*(A) = \gamma_0 I + \gamma_1 A + \ldots + \gamma_k A^k.$$

Тогда справедлива запись:

$$x^{k+1} = x^0 + P_k^*(A) r^0.$$

Воспользуемся A-нормой и запишем:

$$\frac{1}{2}\|x - x^*\|_A^2 = \frac{1}{2}\langle x - x^*, A(x - x^*)\rangle = f(x) - f^*,$$

где $f(x) = \frac{1}{2}\langle x, Ax\rangle - \langle b, x\rangle$.
Согласно Теореме Л6.3 x^{k+1} минимизирует f на множестве

$$P = x^0 + \text{span}\{p^0, \ldots, p^k\} = x^0 + \text{span}\{r^0, Ar^0 \ldots, A^k r^0\},$$

а следовательно, минимизирует и $\|x - x^*\|_A^2$.
Любой точке из P можно сопоставить свой полином P_k степени k. Объединяя это с предыдущим фактом, имеем, что

$$P_k^* = \underset{P_k}{\text{argmin}} \|x^0 + P_k(A) r^0 - x^*\|_A.$$

Поскольку

$$r^0 = Ax^0 - b = A(x^0 - x^*),$$

то верно

$$x^{k+1} - x^* = x^0 + P_k^*(A) r^0 - x^* = \big(I + P_k^*(A) A\big)(x^0 - x^*). \qquad (\text{Л6.7})$$

Пусть $0 < \lambda_1 \leq \lambda_2 \leq \ldots \leq \lambda_d$ — собственные числа A, а $v^1, v^2, \ldots, v^d$ — соответствующие ортонормированные собственные векторы. Так как они являются ортонормированным базисом в $\mathbb{R}^d$, то можно разложить по ним:

$$x^0 - x^* = \sum_{i=1}^{d} \xi_i v^i. \qquad (\text{Л6.8})$$

Посмотрим, как матрица $P_k^*(A)$ действует на v^i:

$$P_k^*(A) v^i = \sum_{j=0}^{k} \gamma_j A^j v^i = \sum_{j=0}^{k} \gamma_j \lambda_i^j v^i = P_k^*(\lambda_i) v^i, \ i = \overline{1, d}.$$

Получается, что $P_k^*(\lambda_1), \ldots, P_k^*(\lambda_d)$ — собственные значения $P_k^*(A)$, отвечающие собственным векторам $v^1, \ldots, v^d$.
Подставляя разложение (Л6.8) в (Л6.7), получим:

$$x^{k+1} - x^* = \sum_{i=1}^{d} \big(1 + \lambda_i P_k^*(\lambda_i)\big) \xi_i v^i,$$

откуда для A-нормы:

$$\left\|x^{k+1} - x^*\right\|_A^2 = \left\langle x^{k+1} - x^*, A\left(x^{k+1} - x^*\right)\right\rangle = \sum_{i=1}^{d} \lambda_i \left(1 + \lambda_i P_k^*\left(\lambda_i\right)\right)^2 \xi_i^2.$$

Теперь применяем оптимальность P_k^* по отношению к A-норме:

$$\left\|x^{k+1} - x^*\right\|_A^2 = \min_{P_k} \sum_{i=1}^{d} \lambda_i \left(1 + \lambda_i P_k\left(\lambda_i\right)\right)^2 \xi_i^2.$$

Вынося самое большой множитель $\left(1 + \lambda_i P_k\left(\lambda_i\right)\right)^2$ из суммы, получаем оценку:

$$\begin{aligned}\left\|x^K - x^*\right\|_A^2 &\leq \min_{P_{K-1}} \max_{i=\overline{1,d}} \left(1 + \lambda_i P_{K-1}\left(\lambda_i\right)\right)^2 \left(\sum_{i=1}^{d} \lambda_i \xi_i^2\right) \\ &= \min_{P_{K-1}} \max_{i=\overline{1,d}} \left(1 + \lambda_i P_{K-1}\left(\lambda_i\right)\right)^2 \left\|x^0 - x^*\right\|_A^2.\end{aligned}$$

■

Теорема Л6.6 (Теорема 5.4 из [18])**.** Пусть система линейных уравнений (Л6.1) решается с помощью метода сопряжённых градиентов (Алгоритм Л6.2). Тогда если матрица A имеет только r различных собственных значений, то метод сойдется к точному решению x^* не более чем за r итераций.

Доказательство. Пусть собственные значения матрицы A равны $\lambda_1, \lambda_2, \ldots, \lambda_d$ и принимают r различных значений $\tau_1 < \tau_2 < \ldots < \tau_r$. Определим полином $Q_r\left(\lambda\right)$:

$$Q_r(\lambda) = \frac{(-1)^r}{\tau_1 \tau_2 \ldots \tau_r}\left(\lambda - \tau_1\right)\left(\lambda - \tau_2\right)\ldots\left(\lambda - \tau_r\right).$$

Его корнями являются $\tau_1, \tau_2, \ldots, \tau_r$, а в нуле он принимает значение $Q_r(0) = 1$. Следовательно, многочлен $Q_r(\lambda) - 1$ имеет степень r и корень 0. Тогда, определенный как

$$\bar{P}_{r-1}(\lambda) = \frac{Q(\lambda) - 1}{\lambda}$$

многочлен имеет степень $r-1$. Применяем оценку из Теоремы Л6.5, и в качестве оценки минимума сверху возьмём $\bar{P}_{r-1}$:

$$0 \leq \min_{P_{r-1}} \max_{i=\overline{1,d}} \left(1 + \lambda_i P_{r-1}(\lambda_i)\right)^2 \leq \max_{i=\overline{1,d}} \left(1 + \lambda_i \bar{P}_{r-1}(\lambda_i)\right)^2 = \max_{i=\overline{1,d}} Q_r^2(\lambda_i) = 0.$$

Таким образом, мы подобрали многочлен $\bar{P}_{r-1}$, который дает точную оценку в неравенстве выше. Тогда согласно оценке из Теоремы Л6.5:

$$\left\|x^r - x^*\right\|_A^2 \leq \min_{P_{r-1}} \max_{i=\overline{1,d}} \left[1 + \lambda_i P_{r-1}(\lambda_i)\right]^2 \left\|x^0 - x^*\right\|_A^2 = 0.$$

■

Помимо оценок для достижения точного решения, есть оценка и для ошибки на k-ой итерации.

Теорема Л6.7 (Теорема 5.5 из [18])**.** Пусть система линейных уравнений (Л6.1) решается с помощью метода сопряжённых градиентов (Алгоритм Л6.2). Тогда справедлива следующая оценка сходимости:

$$\left\|x^K - x^*\right\|_A^2 \leq \left(\frac{\lambda_{d-K+1} - \lambda_1}{\lambda_{d-K+1} + \lambda_1}\right)^2 \left\|x^0 - x^*\right\|_A^2,$$

где $\lambda_1 \leq \lambda_2 \leq \ldots \leq \lambda_d$ — собственные числа матрицы A.

Доказательство. Запишем оценку сходимости из Теоремы Л6.5:

$$\left\|x^{K+1} - x^*\right\|_A^2 \leq \min_{P_K} \max_{i=\overline{1,d}} \left[1 + \lambda_i P_K(\lambda_i)\right]^2 \left\|x^0 - x^*\right\|_A^2.$$

Построим многочлен $\bar{P}_K(\lambda)$, который позволит оценить минимум сверху. Для этого определим многочлен $Q_{K+1}(\lambda)$:

$$Q_{K+1}(\lambda) = \frac{(-1)^{K+1} \cdot (\lambda - \lambda_d)(\lambda - \lambda_{d-1}) \ldots (\lambda - \lambda_{d-K+1})\left(\lambda - \frac{\lambda_1 + \lambda_{d-K}}{2}\right)}{\lambda_d \lambda_{d-1} \ldots \lambda_{d-K+1} \cdot \frac{\lambda_1 + \lambda_{d-K}}{2}}.$$

В нуле он принимает значение 1, поэтому многочлен $Q_{K+1}(\lambda) - 1$ степени $K+1$ и имеет корень в нуле. Полином $\bar{P}_K$ степени K определим как

$$\bar{P}_K(\lambda) = \frac{Q_{K+1}(\lambda) - 1}{\lambda}.$$

Его и подставим для нашей оценки:

$$0 \leq \min_{P_K} \max_{i=\overline{1,d}} \left[1 + \lambda_i P_K(\lambda_i)\right]^2 \leq \max_{i=\overline{1,d}} \left[1 + \lambda_i \bar{P}_K(\lambda_i)\right]^2 = \max_{i=\overline{1,d}} Q_{K+1}^2(\lambda_i).$$

Внимательнее посмотрим, как выглядит выражение справа:

$$\max_{i=\overline{1,d}} Q_{K+1}^2(\lambda_i) = \max_{i=\overline{1,d}} \frac{(\lambda_i - \lambda_d)^2(\lambda_i - \lambda_{d-1})^2 \ldots (\lambda_i - \lambda_{d-K+1})^2\left(\lambda_i - \frac{\lambda_1 + \lambda_{d-K}}{2}\right)^2}{\lambda_d^2 \lambda_{d-1}^2 \ldots \lambda_{d-K+1}^2 \cdot \left(\frac{\lambda_1 + \lambda_{d-K}}{2}\right)^2}.$$

В знаменателе константы, которые не зависят от i. Остается максимизировать только числитель. На значениях $\lambda_d, \ldots, \lambda_{d-K+1}$ числитель равен нулю. Среди оставшихся собственных чисел скобка $\left(\lambda_i - \frac{\lambda_1 + \lambda_{d-K}}{2}\right)^2$ максимальна и принимает одинаковые значения в точках λ_1 и λ_{d-K}. Однако, другие члены в произведении монотонно растут при движении от λ_{d-K} к λ_1. Отсюда делаем вывод, что в λ_1 достигается максимум каждого отдельного неотрицательного члена в произведении, поэтому и всего произведения. Продолжаем оценку значением в точке λ_1:

$$\begin{aligned}\max_{i=\overline{1,d}} Q_{K+1}^2(\lambda_i) &= \frac{(\lambda_1 - \lambda_d)^2(\lambda_1 - \lambda_{d-1})^2 \ldots (\lambda_1 - \lambda_{d-K+1})^2\left(\lambda_1 - \frac{\lambda_1 + \lambda_{d-K}}{2}\right)^2}{\lambda_d^2 \lambda_{d-1}^2 \ldots \lambda_{d-K+1}^2 \cdot \left(\frac{\lambda_1 + \lambda_{d-K}}{2}\right)^2} \\ &\leq \frac{\left(\lambda_1 - \frac{\lambda_1 + \lambda_{d-K}}{2}\right)^2}{\left(\frac{\lambda_1 + \lambda_{d-K}}{2}\right)^2} = \left(\frac{\lambda_{d-K} - \lambda_1}{\lambda_{d-K} + \lambda_1}\right)^2.\end{aligned}$$

Эта оценка завершает доказательство. ■

Таким образом, если мы сделали k итераций, и разница между λ_{d-k} и λ_1 уже достаточно мала, то можно остановить метод. На практике часто бывает, что у матрицы 5–10 больших собственных значений, а остальные сильно меньше. В таких случаях метод сопряжённых градиентов находит достаточно хорошее решение быстрее остальных методов (метода тяжелого шарика, метода Нестерова). На самом деле, имеет место и другая оценка, более напоминающая ранее полученные оценки для ускоренных методов.

Теорема Л6.8 (Формула (5.35) из [18])**.** Пусть система линейных уравнений (Л6.1) решается с помощью метода сопряжённых градиентов (Алгоритм Л6.2). Тогда справедлива следующая оценка сходимости:

$$\left\|x^K - x^*\right\|_A \leq 4\left(\frac{\sqrt{\kappa(A)}-1}{\sqrt{\kappa(A)}+1}\right)^{2K}\left\|x^0 - x^*\right\|_A,$$

где $\kappa(A) = \lambda_{\max}(A)/\lambda_{\min}(A)$ — число обусловленности матрицы A.

Доказательство. Обратимся к оценке из Теоремы Л6.5:

$$\left\|x^K - x^*\right\|_A^2 \leq \min_{P_{K-1}} \max_{i=\overline{1,d}} \left[1 + \lambda_i P_{K-1}(\lambda_i)\right]^2 \left\|x^0 - x^*\right\|_A^2.$$

Обозначим

$$Q_K(\lambda) = 1 + \lambda P_{K-1}(\lambda).$$

Теперь мы можем искать минимум по всем многочленам $Q_K(\lambda)$ степени K с ограничением $Q_K(0) = 1$:

$$\left\|x^K - x^*\right\|_A^2 \leq \min_{Q_K : Q_K(0)=1} \max_{i=\overline{1,d}} Q_K^2(\lambda_i)\left\|x^0 - x^*\right\|_A^2.$$

В желаемой оценке должны присутствовать только максимальное и минимальное собственные значения, поэтому подберём многочлен, который минимально отклоняется от нуля на отрезке $[\lambda_{\min}, \lambda_{\max}]$, при этом удовлетворяет ограничению $Q_K(0) = 1$:

$$\left\|x^K - x^*\right\|_A^2 \leq \min_{Q_K : Q_K(0)=1} \max_{\lambda \in [\lambda_{\min}, \lambda_{\max}]} Q_K^2(\lambda)\left\|x^0 - x^*\right\|_A^2.$$

Такой многочлен можно построить, используя полиномы Чебышёва. Определяются они следующим образом:

$$T_k(z) = \begin{cases} \frac{1}{2}\left(\left(z + \sqrt{z^2-1}\right)^k + \left(z - \sqrt{z^2-1}\right)^k\right), & |z| > 1, \\ \cos(k \arccos z), & |z| \leq 1. \end{cases}$$

Главное их свойство, что на отрезке $[-1, 1]$ они наименьшим образом отклоняются от 0, ровно то свойство, что нам хочется. Остается только биективно отобразить отрезок $[\lambda_{\min}, \lambda_{\max}]$ на $[-1, 1]$ и выполнить ограничение в нуле. Из этих рассуждений, получаем многочлен

$$Q_K(\lambda) = T_K\left(\frac{\lambda_{\min} + \lambda_{\max} - 2\lambda}{\lambda_{\max} - \lambda_{\min}}\right) \Big/ T_K\left(\frac{\lambda_{\max} + \lambda_{\min}}{\lambda_{\max} - \lambda_{\min}}\right).$$

Действительно, это многочлен от λ степени K, линейное преобразование под аргументом биективно отображает отрезок $[\lambda_{\min}, \lambda_{\max}]$ в $[-1, 1]$. Это обеспечивает минимальное отклонение от нуля на заданном отрезке. Деление на константу необходимо, чтобы выполнить ограничение на значение в нуле. Теперь найдём максимум. Известно,

что на $[-1, 1]$ максимальное значение модуля многочленов Чебышёва равно 1. С этого начинаем оценку:

$$\max_{\lambda\in[\lambda_{\min},\lambda_{\max}]} Q_K^2(\lambda) = \max_{\lambda\in[\lambda_{\min},\lambda_{\max}]} \left(T_K\left(\frac{\lambda_{\min}+\lambda_{\max}-2\lambda}{\lambda_{\max}-\lambda_{\min}}\right)\Big/ T_K\left(\frac{\lambda_{\max}+\lambda_{\min}}{\lambda_{\max}-\lambda_{\min}}\right)\right)^2$$

$$= T_K^{-2}\left(\frac{\lambda_{\max}+\lambda_{\min}}{\lambda_{\max}-\lambda_{\min}}\right).$$

Подставляем выражение для полинома Чебышёва при $|z| > 1$:

$$\begin{aligned}\max_{\lambda\in[\lambda_{\min},\lambda_{\max}]} Q_K^2(\lambda) &= 4\left(\left(\frac{\lambda_{\max}+\lambda_{\min}}{\lambda_{\max}-\lambda_{\min}} + \sqrt{\frac{(\lambda_{\max}+\lambda_{\min})^2}{(\lambda_{\max}-\lambda_{\min})^2}-1}\right)^K\right.\\ &\quad\left.+\left(\frac{\lambda_{\max}+\lambda_{\min}}{\lambda_{\max}-\lambda_{\min}} - \sqrt{\frac{(\lambda_{\max}+\lambda_{\min})^2}{(\lambda_{\max}-\lambda_{\min})^2}-1}\right)^K\right)^{-2}\\ &= 4\left(\left(\frac{\lambda_{\max}+\lambda_{\min}+2\sqrt{\lambda_{\max}\lambda_{\min}}}{\lambda_{\max}-\lambda_{\min}}\right)^K\right.\\ &\quad\left.+\left(\frac{\lambda_{\max}+\lambda_{\min}-2\sqrt{\lambda_{\max}\lambda_{\min}}}{\lambda_{\max}-\lambda_{\min}}\right)^K\right)^{-2}\\ &= 4\left(\left(\frac{\sqrt{\lambda_{\max}}+\sqrt{\lambda_{\min}}}{\sqrt{\lambda_{\max}}-\sqrt{\lambda_{\min}}}\right)^K + \left(\frac{\sqrt{\lambda_{\max}}-\sqrt{\lambda_{\min}}}{\sqrt{\lambda_{\max}}+\sqrt{\lambda_{\min}}}\right)^K\right)^{-2}.\end{aligned}$$

Переписывая через число обусловленности $\kappa = \lambda_{\max}/\lambda_{\min}$:

$$\begin{aligned}\max_{\lambda\in[\lambda_{\min},\lambda_{\max}]} Q_K^2(\lambda) &= 4\left(\left(\frac{\sqrt{\kappa}+1}{\sqrt{\kappa}-1}\right)^K + \left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)^K\right)^{-2}\\ &\le 4\left(\frac{\sqrt{\kappa}+1}{\sqrt{\kappa}-1}\right)^{-2K} = 4\left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)^{2K}.\end{aligned}$$

■

Эта оценка более грубая, поскольку появляется зависимость от числа обусловленности. Если внимательно посмотреть на прошлую оценку, то в ней спустя k итераций мы обрабатываем k самых больших собственных чисел, и они более не дают свой вклад в оценку. В этой же формуле зависимость от самого большого собственного числа присутствует на всех итерациях. Сама же оценка похожа на оценки для ускоренных методов. Действительно, если перегруппировать и оценить, то можно получить множитель в правой части

$$\left(1 - 2\frac{1}{\sqrt{\kappa(A)}+1}\right)^{2K} \le \left(1 - \sqrt{\frac{\mu}{L}}\right)^{2K}.$$

По этой оценке можно сделать вывод, что метод сопряжённых градиентов сходится как метод Нестерова на квадратичной задаче.

Л6.3 Нелинейные методы сопряжённых градиентов

Мы получили метод сопряжённых градиентов (Алгоритм Л6.2) для решения систем линейных уравнений (Л6.1) или, эквивалентно, минимизации квадратичной функции (Л6.2). Но что делать, если мы хотим решать более общую задачу, а именно, безусловную задачу минимизации:

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

где на f мы как и ранее будем накладывать условия L-гладкости и выпуклости.
Обобщить метод сопряжённых градиентов действительно возможно. Ранее мы подбирали шаг α_k как шаг наискорейшего спуска вдоль p^k. Для квадратичной функции он вычисляется аналитически. Для функции общего вида же потребуется осуществлять линейный поиск, например, как это было описано в Параграфе Л3 в разделе про подбор шага. Другим изменением будет замена r^k на $\nabla f\left(x^k\right)$. Это мотивировано тем, что для нелинейной функции нет понятия невязки, зато есть градиент, который в случае квадратичной функции совпадал с невязкой. Полученный модифицированный метод сопряжённых градиентов называется методом Флетчера–Ривса.

Алгоритм Л6.3 Метод сопряжённых градиентов (Флетчер–Ривс)

Вход: стартовая точка $x^0 \in \mathbb{R}^d$, стартовый сопряжённый вектор $p^0 = -\nabla f\left(x^0\right)$, количество итераций K

for $k = 0, 1, \ldots, K-1$ **do**
  $\alpha_k =$ линейный поиск
  $x^{k+1} = x^k + \alpha_k p^k$
  $\beta_{k+1} = \dfrac{\left\langle \nabla f\left(x^{k+1}\right), \nabla f\left(x^{k+1}\right) \right\rangle}{\left\langle \nabla f\left(x^k\right), \nabla f\left(x^k\right) \right\rangle}$
  $p^{k+1} = -\nabla f\left(x^{k+1}\right) + \beta_{k+1} p^k$
end for
Выход: x^K

В случае когда f — сильно выпуклая квадратичная функция и производится точный поиск α_k, итерации метода Флетчера–Ривса совпадают с классическим методом сопряжённых градиентов. Теоретические гарантии для метода достаточно слабые, и только если использовать сильные условия Вольфа для линейного поиска.
Есть также модификации метода Флетчера–Ривса, например, Полак и Рибьер предложили другой вариант для подсчёта β_{k+1}.

Алгоритм Л6.4 Метод сопряжённых градиентов (Полак–Рибьер)

Вход: стартовая точка $x^0 \in \mathbb{R}^d$, стартовый сопряжённый вектор $p^0 = -\nabla f\left(x^0\right)$, количество итераций K

for $k = 0, 1, \ldots, K-1$ **do**
  $\alpha_k =$ линейный поиск
  $x^{k+1} = x^k + \alpha_k p^k$
  $\beta_{k+1} = \dfrac{\left\langle \nabla f\left(x^{k+1}\right), \nabla f\left(x^{k+1}\right) - \nabla f\left(x^k\right) \right\rangle}{\left\langle \nabla f\left(x^k\right), \nabla f\left(x^k\right) \right\rangle}$
  $p^{k+1} = -\nabla f\left(x^{k+1}\right) + \beta_{k+1} p^k$
end for
Выход: x^K

Если бы функция f была квадратичной и сильно выпуклой, то градиенты на соседних итерациях были бы ортогональны, и итерации снова совпали с классическим вариантом (Алгоритм Л6.2). Для выпуклых гладких функций это не всегда правда, поэтому эта поправка действительно влияет на сходимость. На практике метод Полака–Рибьера сходится быстрее, чем метод Флетчера–Ривса.
Еще один вариант выбора β_{k+1} предложили Хестенс и Штифель. На практике такой метод сходится примерно как метод Полака–Рибьера.

Алгоритм Л6.5 Метод сопряжённых градиентов (Хестенс–Штифель)

Вход: стартовая точка $x^0 \in \mathbb{R}^d$, стартовый сопряжённый вектор $p^0 = -\nabla f\left(x^0\right)$, количество итераций K

1: **for** $k = 0, 1, \ldots, K-1$ **do**
2: $\quad \alpha_k =$ линейный поиск
3: $\quad x^{k+1} = x^k + \alpha_k p^k$
4: $\quad \beta_{k+1} = \dfrac{\left\langle \nabla f\left(x^{k+1}\right), \nabla f\left(x^{k+1}\right) - \nabla f(x^k) \right\rangle}{\left\langle \nabla f\left(x^{k+1}\right) - \nabla f\left(x^k\right), p^k \right\rangle}$
5: $\quad p^{k+1} = -\nabla f\left(x^{k+1}\right) + \beta_{k+1} p^k$
6: **end for**

Выход: x^K

Л7 Метод Ньютона. Квазиньютоновские методы

Ранее мы рассматривали градиентные методы, которые благодаря низкой вычислительной стоимости шага остаются основным инструментом для задач с очень большой размерностью. Их скорость сходимости обычно линейная — этого достаточно во многих практических случаях. Возникает вопрос, можно ли двигаться к решению быстрее, если мы обладаем большей информацией о целевой функции? Методы второго порядка, такие как метод Ньютона, используют не только значения функции и градиент, но и информацию о гессиане, что позволяет достигать квадратичной скорости сходимости и существенно сокращать число итераций.

Л7.1 Задача поиска нуля

Рассмотрим задачу поиска «корня» функции. Формально требуется для функции $\varphi : \mathbb{R} \to \mathbb{R}$ найти точку t^*, для которой верно

$$\varphi\big(t^*\big) = 0.$$

Пусть мы находимся в точке t^0 и хотим найти такую поправку Δt, что $t^0 + \Delta t \approx t^*$. Разложим φ по формуле Тейлора до первого порядка:

$$\varphi\big(t^0 + \Delta t\big) = \varphi\big(t^0\big) + \varphi'\big(t^0\big)\Delta t + o(\Delta t).$$

Поскольку $t^0 + \Delta t \approx t^*$, то:

$$\varphi\big(t^0 + \Delta t\big) \approx \varphi\big(t^*\big) = 0 \implies \varphi\big(t^0\big) + \varphi'\big(t^0\big)\Delta t \approx 0.$$

Таким образом, $\Delta t \approx -\dfrac{\varphi\big(t^0\big)}{\varphi'\big(t^0\big)}$. Продолжая строить новые точки, получаем выражение для итерации метода, предложенного Ньютоном в 17-м веке [17]:

$$t^{k+1} = t^k - \frac{\varphi\big(t^k\big)}{\varphi'\big(t^k\big)}.$$

Замечание Л7.1. Важно отметить ключевую особенность предложенного метода — сходимость к решению только в некоторой его окрестности. В качестве примера можно рассмотреть $\varphi(x) = \frac{x}{\sqrt{1+x^2}}$. Выберем в качестве начального приближения некоторую точку t^0. Тогда нетрудно видеть, что

- $\big|t^0\big| < 1$ — метод сходится,
- $\big|t^0\big| = 1$ — метод колеблется в точках -1 и 1,
- $\big|t^0\big| > 1$ — метод расходится.

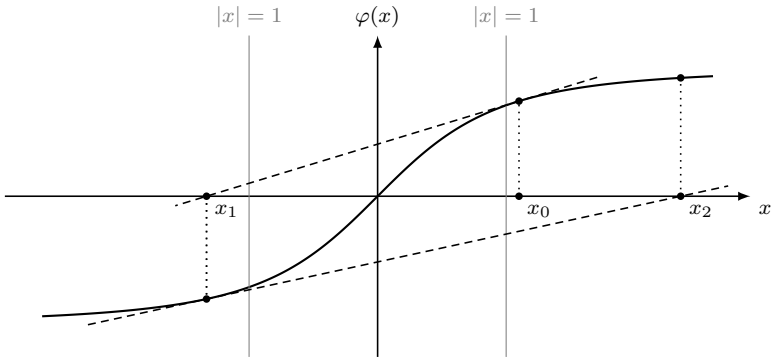
Л7.2 Метод Ньютона

Вернемся к задаче безусловной минимизации:

$$\min_{x \in \mathbb{R}^d} f(x). \tag{Л7.1}$$

Ранее уже обсуждалось, что для выпуклой целевой функции она эквивалентна задаче поиска нуля градиента:

$$\nabla f\big(x^*\big) = 0,$$



где $\nabla f : \mathbb{R}^d \to \mathbb{R}^d$.
Интуитивно понятно, что мы ожидаем получить шаг:

$$x^{k+1} = x^k - \big(\nabla^2 f(x)\big)^{-1} \nabla f\big(x\big).$$

Градиентный спуск работает с линейной аппроксимацией функции. Теперь будем минимизировать её квадратичную аппроксимацию:

$$f\big(x\big) \approx f\big(x^k\big) + \big\langle \nabla f\big(x^k\big), x - x^k \big\rangle + \frac{1}{2}\big\langle \nabla^2 f\big(x^k\big)\big(x - x^k\big), x - x^k \big\rangle.$$

Построим x^{k+1} как точку, где зануляется градиент квадратичной аппроксимации:

$$\nabla f\big(x^k\big) + \nabla^2 f\big(x^k\big)\big(x - x^k\big) = 0 \implies x^{k+1} = x^k - \big(\nabla^2 f\big(x^k\big)\big)^{-1} \nabla f\big(x^k\big).$$

Действительно, получили то, что ожидали:

Алгоритм Л7.1 Метод Ньютона

Вход: стартовая точка $x^0 \in \mathbb{R}^d$, количество итераций K
for $k = 0, 1, \ldots, K-1$ **do**
    $x^{k+1} = x^k - \big(\nabla^2 f\big(x^k\big)\big)^{-1} \nabla f\big(x^k\big)$
end for
Выход: x^K

Пример Л7.1. Рассмотрим задачу:

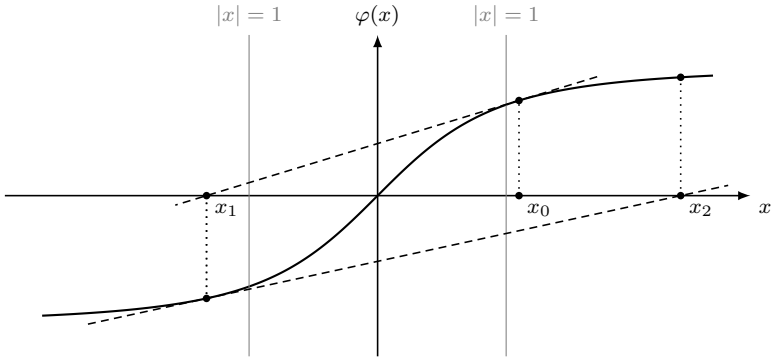
$$\min_{x \in \mathbb{R}^d} \left[f(x) = \frac{1}{2}\langle x, Ax \rangle - \langle b, x \rangle + c \right],$$

где $A \in \mathbb{S}^d_+$.
Пусть x^* — решение задачи. Тогда нетрудно заметить, что

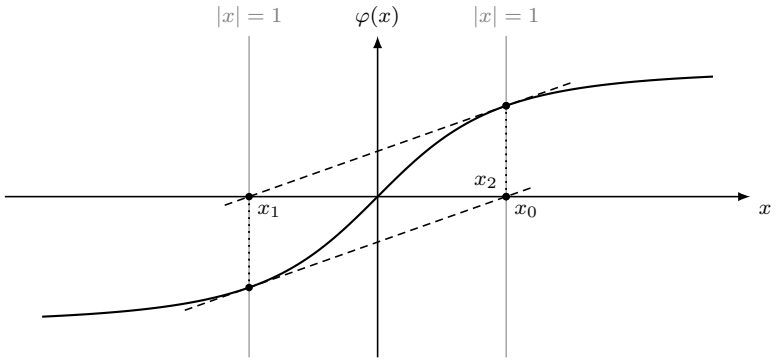
$$\nabla f\big(x^*\big) = Ax^* - b = 0,$$

то есть $x^* = A^{-1}b$. При этом также нетрудно заметить, что

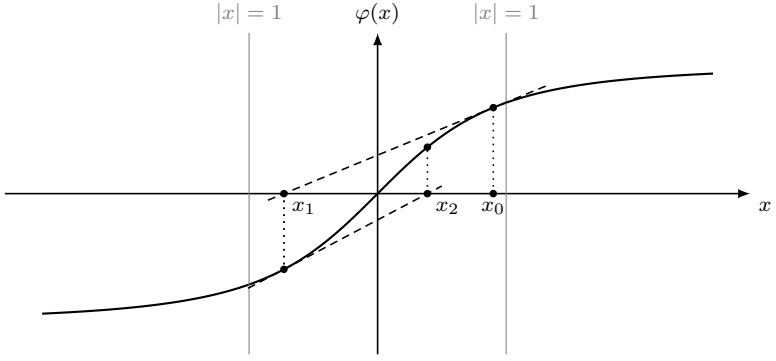
$$\big(\nabla^2 f\big(x\big)\big)^{-1} \nabla f\big(x\big) = A^{-1}\big(Ax - b\big) = x - A^{-1}b.$$



$|x^0| > 1$: расходимость



$|x^0| = 1$: колебания $1 \rightleftarrows -1$



$|x^0| < 1$: сходимость к 0

Рис. Л7.1: Геометрические шаги Ньютона для $\varphi(x) = \frac{x}{\sqrt{1+x^2}}$ в трёх режимах. На каждом подграфике: точка $(x_k, \varphi(x_k))$, касательная через неё и $(x_{k+1}, 0)$ (где $x_{k+1} = -x_k^3$), затем вертикаль к графику для следующей точки.

Таким образом, метод Ньютона для квадратичной задачи сходится за одну итерацию. Важно отметить, что вычислительная стоимость этой итерации очень велика, не только из-за использования гессиана, но еще и из-за его обращения за $\mathcal{O}(d^3)$.

Далее будем работать в предположении, что целевая функция в задаче безусловной минимизации является дважды непрерывно дифференцируемой, μ-сильно выпуклой $\left(\nabla^2 f(x) \succeq \mu I_d\right)$, а также имеет M-Липшицев гессиан.

Определение Л7.1. Пусть дана дважды непрерывно дифференцируемая на $\mathbb{R}^d$ функция $f : \mathbb{R}^d \to \mathbb{R}$. Будем говорить, что данная функция имеет *M-Липшицев гессиан*, если для любых $x, y \in \mathbb{R}^d$ выполнено

$$\left\|\nabla^2 f(x) - \nabla^2 f(y)\right\|_2 \leq M\|x - y\|_2.$$

Утверждение Л7.1. Пусть функция f имеет M-Липшицев гессиан. Введем обозначение

$$G_k = \nabla^2 f\left(x^k\right) - \int\limits_0^1 \nabla^2 f\left(x^* + \tau\left(x^k - x^*\right)\right) \mathrm{d}\tau\,.$$

Тогда справедлива оценка:

$$\left\|G_k\right\|_2 \leq \frac{M}{2}\left\|x^k - x^*\right\|_2.$$

Доказательство.

$$\begin{aligned}
\left\|G_k\right\|_2 &= \left\|\nabla^2 f\left(x^k\right) - \int\limits_0^1 \nabla^2 f\left(x^* + \tau\left(x^k - x^*\right)\right) \mathrm{d}\tau\right\|_2 \\
&= \left\|\int\limits_0^1 \left(\nabla^2 f\left(x^k\right) - \nabla^2 f\left(x^* + \tau\left(x^k - x^*\right)\right)\right) \mathrm{d}\tau\right\|_2 \\
&\leq \int\limits_0^1 \left\|\nabla^2 f\left(x^k\right) - \nabla^2 f\left(x^* + \tau\left(x^k - x^*\right)\right)\right\|_2 \mathrm{d}\tau \\
&\leq \int\limits_0^1 M(1-\tau)\left\|x^k - x^*\right\|_2 \mathrm{d}\tau \\
&= M\left\|x^k - x^*\right\|_2 \int\limits_0^1 (1-\tau)\,\mathrm{d}\tau \\
&= \frac{M}{2}\left\|x^k - x^*\right\|_2.
\end{aligned}$$

■

Теперь можем сформулировать теорему о сходимости метода Ньютона.

Теорема Л7.1. Пусть задача безусловной оптимизации (Л7.1) с M-Липшицевым гессианом, μ-сильно выпуклой целевой функцией f, решается с помощью метода Ньютона (Алгоритм Л7.1) Тогда справедлива следующая оценка сходимости:

$$\left\|x^{k+1} - x^*\right\|_2 \leq \frac{M}{2\mu}\left\|x^k - x^*\right\|_2^2.$$

Доказательство. Будем изучать, как меняется расстояние до решения:

$$x^{k+1} - x^* = x^k - \left(\nabla^2 f\left(x^k\right)\right)^{-1}\nabla f\left(x^k\right) - x^*.$$

Воспользуемся формулой Ньютона-Лейбница для интеграла вдоль кривой:

$$\nabla f\left(x^k\right) - \nabla f\left(x^*\right) = \int\limits_0^1 \nabla^2 f\left(x^* + \tau\left(x^k - x^*\right)\right)\left(x^k - x^*\right)\mathrm{d}\tau\,.$$

Зная, что $\nabla f\left(x^*\right) = 0$, получим:

$$x^{k+1} - x^* = x^k - x^* - \left(\nabla^2 f\left(x^k\right)\right)^{-1}\int\limits_0^1 \nabla^2 f\left(x^* + \tau\left(x^k - x^*\right)\right)\left(x^k - x^*\right)\mathrm{d}\tau\,.$$

Используем «умную единицу»:

$$\begin{aligned} x^{k+1} - x^* &= x^k - x^* - \left(\nabla^2 f\left(x^k\right)\right)^{-1}\int\limits_0^1 \nabla^2 f\left(x^* + \tau\left(x^k - x^*\right)\right)\left(x^k - x^*\right)\mathrm{d}\tau \\ &= \left(\nabla^2 f\left(x^k\right)\right)^{-1}\nabla^2 f\left(x^k\right)\left(x^k - x^*\right) \\ &\quad - \left(\nabla^2 f\left(x^k\right)\right)^{-1}\int\limits_0^1 \nabla^2 f\left(x^* + \tau\left(x^k - x^*\right)\right)\left(x^k - x^*\right)\mathrm{d}\tau\,. \end{aligned}$$

Заметим, что $x^k - x^*$ можно вынести за пределы интеграла:

$$\begin{aligned} x^{k+1} - x^* &= \left(\nabla^2 f\left(x^k\right)\right)^{-1}\left(\nabla^2 f\left(x^k\right) - \int\limits_0^1 \nabla^2 f\left(x^* + \tau\left(x^k - x^*\right)\right)\mathrm{d}\tau\right)\left(x^k - x^*\right) \\ &= \left(\nabla^2 f\left(x^k\right)\right)^{-1} G_k\left(x^k - x^*\right). \end{aligned}$$

Спектральная норма матрицы согласована с евклидовой нормой вектора. Таким образом:

$$\left\|x^{k+1} - x^*\right\|_2 \leq \left\|\left(\nabla^2 f\left(x^k\right)\right)^{-1}\right\|_2 \left\|G_k\right\|_2 \left\|x^k - x^*\right\|_2.$$

Норму обратного гессиана оценим как $\left\|\left(\nabla^2 f\left(x^k\right)\right)^{-1}\right\|_2 = \frac{1}{\lambda_{\min}\left(\nabla^2 f\left(x^k\right)\right)} \leq \frac{1}{\mu}$. Остается лишь применить результат Утверждения Л7.1, и получим требуемое:

$$\left\|x^{k+1} - x^*\right\|_2 \leq \frac{M}{2\mu}\left\|x^k - x^*\right\|_2^2.$$

■

Замечание Л7.2. Сходимость является локальной. А именно, чтобы гарантировать

$$\left\|x^1 - x^*\right\|_2 < \left\|x^0 - x^*\right\|_2,$$

нужно предположить, что:

$$\left\|x^0 - x^*\right\|_2 < \frac{2\mu}{M}.$$

Приведем пример, когда метод Ньютона расходится при старте из плохого приближения.

Пример Л7.2. Рассмотрим функцию $f : \mathbb{R} \to \mathbb{R}$:

$$f\big(x\big) = \frac{x^2}{2} + 10 \log \cosh x.$$

Её минимум находится в точке $x^* = 0$.
Найдём константу сильной выпуклости μ и константу липшицевости гессиана M.
Первые три производные имеют вид:

$$f'(x) = x + 10 \tanh x,$$
$$f''(x) = 1 + \frac{10}{\cosh^2 x},$$
$$f'''(x) = -20 \frac{\tanh x}{\cosh^2 x}.$$

Так как $\cosh^2 x \geq 1$, имеем

$$\mu = \min_{x \in \mathbb{R}} f''(x) = 1.$$

Для константы липшицевости гессиана в одномерном случае:

$$M = \sup_{x \in \mathbb{R}} \left|f'''(x)\right|.$$

Положим $t = \tanh x \in (-1, 1)$, тогда $\frac{1}{\cosh^2 x} = 1 - t^2$ и

$$\left|f'''(t)\right| = 20|t|\big(1 - t^2\big).$$

Это выражение достигает максимума при $t = \pm\frac{1}{\sqrt{3}}$, и, следовательно,

$$M = \frac{40}{3\sqrt{3}} \approx 7.692.$$

Согласно теоретическим оценкам, метод сходится при начальном приближении x^0:

$$\left|x^0 - x^*\right| \leq \frac{2\mu}{M} = \frac{3\sqrt{3}}{20} \approx 0.260.$$

Однако, оценка достаточно грубая, поскольку использует всего 2 константы, связанных с функцией. Найдем точно, где метод начинает расходиться.
Итерация метода Ньютона имеет вид

$$x^{k+1} = x^k - \frac{x^k + 10 \tanh x^k}{1 + \frac{10}{\cosh^2 x^k}}.$$

Поскольку минимум находится в нуле, функция f четная, то на границе области сходимости модуль следующей точки совпадает с текущим, а знак меняется:

$$x^{k+1} = -x^k.$$

Это условие даёт уравнение

$$2x = \frac{x + 10\tanh x}{1 + \frac{10}{\cosh^2 x}},$$

положительный корень которого:

$$\bar{x} \approx 1.22258$$

и задаёт границу области сходимости.
Итого, поведение метода Ньютона:

- $\left|x^0\right| < \bar{x}$ — сходимость к 0 квадратичная.
- $\left|x^0\right| = \bar{x}$ — колебания.
- $\left|x^0\right| > \bar{x}$ — расходимость.

Замечание Л7.3. Существует несколько способов сделать сходимость метода Ньютона глобальной. Один из них — так называемый демпфированный метод с шагом

$$x^{k+1} = x^k - \gamma_k \big(\nabla^2 f\big(x^k\big)\big)^{-1} \nabla f\big(x^k\big).$$

Стоит отметить, что на практике демпфированные методы сходятся медленнее.

Л7.3 Метод Ньютона с кубической регуляризацией

Вспомним идейно один из способов получить интуицию метода градиентного спуска. Пользуясь гладкостью градиента целевой функции, можно аппроксимировать её значение некоторым эллиптическим параболоидом. Далее можно явно выписать его минимум — это и будет шаг градиентного спуска. Попробуем распространить эту идею на методы высших порядков. Предположим, что целевая функция имеет M-липшицев гессиан. Тогда можно записать аппроксимацию второго порядка:

$$\xi_{2,x}(y) = f\big(x\big) + \big\langle \nabla f\big(x\big), y - x\big\rangle + \frac{1}{2}\big\langle \nabla^2 f\big(x\big)\big(y - x\big), y - x\big\rangle + \frac{M}{6}\big\|x - y\big\|_2^3.$$

Она устроена также, как и квадратичная парабола, которая мы строилась для L-гладких функций. Эта аппроксимация подпирает сверху целевую функцию, поэтому $f(y) \le \xi_{2,x}(y)$. Построим итерационную схему, минимизирующую верхнюю оценку:

$$x^{k+1} = \operatorname*{argmin}_{y \in \mathbb{R}^d} \xi_{2,x^k}(y).$$

Для простоты анализа определим

$$T_M(x) \in \operatorname*{argmin}_{y \in \mathbb{R}^d} \left[\big\langle \nabla f\big(x\big), y - x\big\rangle + \frac{1}{2}\big\langle \nabla^2 f\big(x\big)\big(y - x\big), y - x\big\rangle + \frac{M}{6}\big\|x - y\big\|_2^3\right].$$

Здесь мы константа Липшица гессиана будет параметром, чтобы иметь возможность варьировать шаг. Такой подход называется кубической регуляризацией метода Ньютона.

Алгоритм Л7.2 Метод Ньютона с кубической регуляризацией

Вход: стартовая точка $x^0 \in \mathbb{R}^d$, параметры $\{M_k\}_{k=0}^{K-1}$, количество итераций K

1: **for** $k = 0, \dots, K-1$ **do**
2: $\quad x^{k+1} = T_{M_k}(x^k)$
3: **end for**

Выход: x^K

Далее мы предполагаем, что собственные значения Гессиана пронумерованы в порядке убывания.

Теорема Л7.2 (Теорема 3 из [16]). Пусть задача безусловной оптимизации (Л7.1) с M-липшицевым гессианом целевой функцией f решается с помощью метода Ньютона с кубической регуляризацией (Алгоритм Л7.2). Тогда при

$$\nabla^2 f(x^0) \succ 0, \quad \frac{M\|\nabla f(x^0)\|_2}{\lambda_d^2(\nabla^2 f(x^0))} \leq \frac{1}{4}$$

справедлива следующая оценка сходимости:

$$\|\nabla f(x^k)\|_2 \leq \lambda_d^2(\nabla^2 f(x^0)) \frac{9e^{3/2}}{16M} \left(\frac{1}{2}\right)^{2^k}.$$

Полученный метод представляет интерес, поскольку не требует выпуклости, только M-липшицевость гессиана и условия из теоремы в стартовой точке. Сходимость при этом квадратичная по норме градиента, то есть, очень быстрая.

Стоит отметить, что рассматриваемая задача поиска аргминимума, вообще говоря, не выпуклая из-за кубического члена и может иметь локальные минимумы, в то время как нам нужно отыскать глобальный. Тем не менее, эта задача эквивалентна задаче выпуклой одномерной оптимизации.

Теорема Л7.3 (Теорема 10 из [16]). Задача

$$\min_{y \in \mathbb{R}^d} \left[\langle \nabla f(x), y - x \rangle + \frac{1}{2} \langle \nabla^2 f(x)(y-x), y - x \rangle + \frac{M}{6} \|x - y\|_2^3 \right]$$

эквивалентна задаче

$$\min_{r \in \mathcal{D}} \left[\frac{1}{2} \left\langle \left(\nabla^2 f(x) + \frac{Mr}{2} I_d \right)^{-1} \nabla f(x), \nabla f(x) \right\rangle + \frac{M}{12} r^3 \right],$$

где

$$\mathcal{D} = \left\{ r \in \mathbb{R}_+ \;\middle|\; \nabla^2 f(x) + \frac{Mr}{2} I_d \succ 0 \right\}.$$

Л7.4 Квазиньютоновские методы

Метод Ньютона обеспечивает быструю локальную сходимость, однако вычислительная стоимость одной итерации часто оказывается слишком высокой: требуется находить обратный гессиан $\left(\nabla^2 f(x^k)\right)^{-1}$. Идея квазиньютоновских методов заключается в том, чтобы заменить его на приближение H_k, сохраняющее ключевые свойства обратного гессиана, но гораздо более дешёвое в обновлении.

Начнём с линейной аппроксимации приращения градиента:

$$\nabla f\big(x^{k+1}\big) - \nabla f\big(x^k\big) \approx \nabla^2 f\big(x^{k+1}\big)\big(x^{k+1} - x^k\big).$$

Введём обозначения

$$s^k = x^{k+1} - x^k, \quad y^k = \nabla f\big(x^{k+1}\big) - \nabla f\big(x^k\big)$$

и положим

$$B_{k+1} \approx \nabla^2 f\big(x^{k+1}\big), \quad H_{k+1} \approx \big(\nabla^2 f\big(x^{k+1}\big)\big)^{-1}.$$

Тогда аппроксимация выше принимает вид

$$y^k = B_{k+1}s^k \iff s^k = H_{k+1}y^k. \tag{Л7.2}$$

Это уравнение называют *квазиньютоновским*.

Замечание Л7.4. Условие симметричности и квазиньютоновское уравнение дают

$$\frac{d(d-1)}{2} + d = \frac{d(d+1)}{2}$$

линейно независимых ограничений на $d \times d$ матрицу. Поскольку общее число элементов матрицы равно d^2, остаётся $\frac{d(d-1)}{2}$ свободных параметров. Это означает, что множество допустимых приближений очень велико, и существует широкий спектр стратегий построения H_{k+1} или B_{k+1}. Ниже мы рассмотрим наиболее популярные из них.

Общая схема квазиньютоновских методов выглядит следующим образом:

Алгоритм Л7.3 Общая схема квазиньютоновских методов

Вход: стартовая точка $x^0 \in \mathbb{R}^d$, стартовая матрица H_0, размеры шагов $\{\gamma_k\}_{k=0} > 0$, количество итераций K

1: **for** $k = 0, 1, \ldots, K-1$ **do**
2: $\quad x^{k+1} = x^k - \gamma_k H_k \nabla f\big(x^k\big)$
3: $\quad$ Обновить $H_k \to H_{k+1}$ так, чтобы выполнялось (Л7.2)
4: **end for**

Выход: x^K

Замечание Л7.5. Размеры шагов $\{\gamma_k\}_{k=0}^{K-1}$ могут выбираться с помощью процедуры линейного поиска. При этом можно как находить шаг наискорейшего спуска вдоль направления $H_k \nabla f\big(x^k\big)$, так и требовать выполнения условий Вольфа (Алгоритм Л3.6).

Л7.4.1 SR1

Будем использовать для обновления матрицы дешевую с точки зрения вычислений симметричную одноранговую добавку (отсюда и название Symmetric Rank 1):

$$H_{k+1} = H_k + \mu_k q^k \left(q^k\right)^\top,$$

где $\mu_k \in \mathbb{R}$ и $q^k \in \mathbb{R}^d$. Их мы подбираем исходя из квазиньютоновского уравнения:

$$s^k = H_{k+1}y^k = \left(H_k + \mu_k q^k \left(q^k\right)^\top\right)y^k = H_k y^k + \mu_k \left(\left(q^k\right)^\top y^k\right)q^k.$$

Откуда

$$\mu_k \left(\left(q^k\right)^\top y^k\right)q^k = s^k - H_k y^k.$$

Возьмём $\mu_k = \frac{1}{(q^k)^\top y^k}$ и получим, что $q^k = s^k - H_k y^k$.

Итак, `SR1` способ обновления матрицы H_k:

$$H_{k+1} = H_k + \frac{\left(s^k - H_k y^k\right)\left(s^k - H_k y^k\right)^\top}{\left(s^k - H_k y^k\right)^\top y^k}.$$

Л7.4.2 Broyden

Подойдем к задаче иначе. Запишем квазиньютоновское уравнение для матрицы B:

$$y^k = B_{k+1} s^k.$$

Потребуем минимальность фробениусовой нормы поправки при ограничении в виде квазиньютоновского уравнения:

$$\begin{aligned} B_{k+1} = &\operatorname*{argmin}_{B \in \mathbb{R}^{d\times d}} \left\|B - B_k\right\|_F^2 \\ \text{s.t.} \quad & Bs^k = y^k. \end{aligned}$$

Иными словами, хотим найти наиболее близкую матрицу к B_k, которая будет удовлетворять квазиньютоновскому уравнению. Решением такой задачи будет формула обновления:

$$B_{k+1} = B_k + \frac{\left(y^k - B_k s^k\right)\left(s^k\right)^\top}{\left(s^k\right)^\top s^k}.$$

Обратим внимание, что полученная поправка является однoранговой, хотя явно мы этого не требовали. Однако, она не является симметричной, что плохо согласуется со свойствами гессиана, который мы хотим аппроксимировать.

Чтобы не обращать матрицу B_{k+1} и не решать системы линейных уравнений, можем получить явную формулу для $H_{k+1} = B_{k+1}^{-1}$, используя формулу Шермана-Моррисона-Вудбери, впервые предложенную в [21].

Утверждение Л7.2. Пусть $A \in \mathbb{R}^{d\times d}$ — обратимая матрица, а $U, V \in \mathbb{R}^{d\times n}$. Тогда, если также обратима матрица $I_d + V^\top A^{-1} U$, то верна формула:

$$\left(A + UV^\top\right)^{-1} = A^{-1} - A^{-1}U\left(I_d + V^\top A^{-1}U\right)^{-1} V^\top A^{-1}.$$

После ее применения получаем обновление H_k:

$$H_{k+1} = H_k + \frac{\left(s^k - H_k y^k\right)\left(s^k\right)^\top H_k}{\left(s^k\right)^\top H_k \left(y^k\right)^\top}.$$

Л7.4.3 DFP

Модифицируем задачу поиска B_{k+1}. Добавим требование на симметричность матрицы и заменим обычную норму Фробениуса на взвешенную:

$$\left\|A\right\|_W = \left\|W^{1/2} A W^{1/2}\right\|_F,$$

где W — произвольная матрица, для которой выполняется $Wy^k = s^k$. Для определенности, такой матрицей может быть обратный усредненный гессиан $W = \bar{G}_k^{-1}$, определяющийся следующим образом:

$$\bar{G}_k = \int_0^1 \nabla^2 f\left(x^k - \tau\gamma_k H_k \nabla f\left(x^k\right)\right) \mathrm{d}\tau\,.$$

Получим задачу оптимизации для поиска B_{k+1}:

$$\begin{aligned} B_{k+1} = & \operatorname*{argmin}_{B \in \mathbb{R}^{d \times d}} \|B - B_k\|_W^2 \\ \text{s.t.} \quad & Bs^k = y^k, \\ & B^\top = B. \end{aligned}$$

Решением этой задачи является формула обновления `DFP` (Davidson-Fletcher-Powell):

$$B_{k+1} = \left(I - \frac{y^k \left(s^k\right)^\top}{\left(y^k\right)^\top s^k}\right) B_k \left(I - \frac{s^k \left(y^k\right)^\top}{\left(y^k\right)^\top s^k}\right) + \frac{y^k \left(y^k\right)^\top}{\left(y^k\right)^\top s^k}.$$

Обращая ее по формуле Шермана-Моррисона-Вудбери, получаем выражение для H_{k+1}:

$$H_{k+1} = H_k - \frac{H_k y^k \left(y^k\right)^\top H_k}{\left(y^k\right)^\top H_k y^k} + \frac{s^k \left(s^k\right)^\top}{\left(s^k\right)^\top y^k}. \tag{Л7.3}$$

Теорема Л7.4 (Теорема 4.2 из [20]). Пусть задача безусловной оптимизации (Л7.1) с M-липшицевым гессианом, L-гладкой, μ-сильно выпуклой целевой функцией f решается с помощью метода `DFP` (Алгоритм Л7.3 с формулой обновления (Л7.3)). Введем обозначение

$$\lambda_f(x) = \sqrt{\left\langle \nabla f(x), \left(\nabla^2 f(x)\right)^{-1} \nabla f(x) \right\rangle}.$$

Тогда при выборе шагов $\gamma_k = 1$ и начальной точке x^0 достаточно близкой к решению, что выполняется

$$\lambda_f(x^0) \leq \frac{\log \frac{3}{2}}{4} \frac{\mu^{5/2}}{LM},$$

справедлива следующая оценка сходимости:

$$\lambda_f(x^K) \leq \left(\frac{11 d L^2}{\mu^2 K}\right)^{K/2} \lambda_f(x^0).$$

Л7.4.4 BFGS

Поступим так же, как и в `DFP`, только будем решать задачу сразу для H. Она получена из таких же рассуждений, меняется только ограничение:

$$\begin{aligned} H_{k+1} = & \operatorname*{argmin}_{H \in \mathbb{R}^{d \times d}} \|H - H_k\|_W^2 \\ \text{s.t.} \quad & s^k = Hy^k, \\ & H^\top = H. \end{aligned}$$

Решением будет формула `BFGS` (Broyden–Fletcher–Goldfarb–Shanno):

$$H_{k+1} = \left(I_d - \frac{s^k \left(y^k\right)^\top}{\left(y^k\right)^\top s^k}\right) H_k \left(I_d - \frac{y^k \left(s^k\right)^\top}{\left(y^k\right)^\top s^k}\right) + \frac{s^k \left(s^k\right)^\top}{\left(y^k\right)^\top s^k}. \tag{Л7.4}$$

Теорема Л7.5 (Теорема 4.2 из [20])**.** Пусть задача безусловной оптимизации (Л7.1) с M-липшицевым гессианом, L-гладкой, μ-сильно выпуклой целевой функцией f решается с помощью метода `BFGS` (Алгоритм Л7.3 с формулой обновления (Л7.4)). Введем обозначение

$$\lambda_f(x) = \sqrt{\left\langle \nabla f(x), \left(\nabla^2 f(x)\right)^{-1} \nabla f(x) \right\rangle}.$$

Тогда при выборе шагов $\gamma_k = 1$ и начальной точке x^0 достаточно близкой к решению, что выполняется

$$\lambda_f(x^0) \leq \frac{\log \frac{3}{2}}{4} \frac{\mu^{5/2}}{LM},$$

справедлива следующая оценка сходимости:

$$\lambda_f(x^K) \leq \left(\frac{11dL}{\mu K}\right)^{K/2} \lambda_f(x^0).$$

Замечание Л7.6. Существует много способов инициализировать матрицу H_0, но на практике достаточно брать $H_0 = I_d$.

Замечание Л7.7. Сложность всех арифметических операций на одной итерации квазиньютоновских методов $\mathcal{O}(d^2)$. Это дешевле обращения гессиана, но все еще довольно плохой результат для задач экстремально больших размерностей.

Л7.4.5 Сравнительная таблица

Соберем все формулы для рассмотренных методов в одном месте.

Метод	B^{new}	$H^{\text{new}} = \left(B^{\text{new}}\right)^{-1}$
`SR1`	$B + \frac{(y-Bs)(y-Bs)^\top}{(y-Bs)^\top s}$	$H + \frac{(s-Hy)(s-Hy)^\top}{(s-Hy)^\top y}$
`Broyden`	$B + \frac{y-Bs}{s^\top s} s^\top$	$H + \frac{(s-Hy)s^\top H}{s^\top Hy}$
`DFP`	$\left(I_d - \frac{ys^\top}{y^\top s}\right) B \left(I_d - \frac{sy^\top}{y^\top s}\right) + \frac{yy^\top}{y^\top s}$	$H + \frac{ss^\top}{s^\top y} - \frac{Hyy^\top H}{y^\top Hy}$
`BFGS`	$B + \frac{yy^\top}{y^\top s} - \frac{Bss^\top B}{s^\top Bs}$	$\left(I_d - \frac{sy^\top}{y^\top s}\right) H \left(I_d - \frac{ys^\top}{y^\top s}\right) + \frac{ss^\top}{y^\top s}$

Таблица 1: Обновления матриц B и H в различных квазиньютоновских методах. Здесь B, H, s и y соответствуют последней итерации, а B^{new} и H^{new} — их новые значения.

Л7.4.6 L-BFGS

При переходе от метода Ньютона к квазиньютоновским методам мы уже получили ускорение с $\mathcal{O}(d^3)$ до $\mathcal{O}(d^2)$ арифметических операций на итерацию. В памяти же в обоих подходах требуется хранить $\mathcal{O}(d^2)$ значений. Квадратичная стоимость итерации и по арифметическим операциям, и по памяти — это очень дорого для задач большой размерности.

Это подводит нас к следующему методу — `L-BFGS` из [12], где `L` означает Limited-memory. Базовая идея — вместо матрицы $H_k \in \mathbb{R}^{d\times d}$, аппроксимирующей Гессиан, хранить несколько подходящих векторов длины d. За улучшение по памяти приходится платить ухудшением скорости сходимости. Тем не менее, оказывается возможным показать линейную сходимость такого метода.
В случае `BFGS` имели следующее обновление матрицы после применения формулы Шермана–Моррисона–Вудберри:

$$H_{k+1} = V_k^\top H_k V_k + \frac{s^k \left(s^k\right)^\top}{\left(y^k\right)^\top s^k},$$

где $V_k = I_d - \frac{y^k\left(s^k\right)^\top}{\left(y^k\right)^\top s^k}$. Предлагается на итерации k выбирать некоторое начальное приближение Гессиана H_k^0 и затем последовательно применить (Л7.4) по последним m парам $\{s^k, y^k\}$:

$$\begin{aligned} H_{k+1} = {} & \left(V_{k-1}^\top \dots V_{k-m}^\top\right) H_k^0 \left(V_{k-m} \dots V_{k-1}\right) \\ & + \rho^{k-m}\left(V_{k-1}^\top \dots V_{k-m+1}^\top\right) s^{k-m}\left(s^{k-m}\right)^\top \left(V_{k-m+1} \dots V_{k-1}\right) \\ & + \dots \\ & + \rho^{k-1} s^{k-1}\left(s^{k-1}\right)^\top. \end{aligned}$$

Здесь мы обозначили $\rho^k = \frac{1}{\left(y^k\right)^\top s^k}$. Экспериментальные результаты показывают, что m может быть выбрано относительно небольшим (меньше двадцати). Это дает существенное улучшение метода по памяти. Начальное приближение H_k^0 рекомендуется выбирать следующим образом:

$$H_k^0 = \frac{\left(s^{k-1}\right)^\top y^{k-1}}{\left(y^{k-1}\right)^\top y^{k-1}} I_d.$$

Множитель перед единичной матрицей реализует шкалирование матрицы вдоль последнего направления поиска. Для поиска шага для робастности важно использовать условия Вольфа (либо сильные условия Вольфа).
Основные недостатки метода `L-BFGS` заключаются в том, что он часто сходится медленно, особенно на плохо обусловленных задачах.

Л7.5 Trust-Region шаг

Методы линейного поиска основаны на поиске подходящего размера шага γ_k вдоль фиксированного направления p^k, которым может быть градиент $\nabla f\left(x^k\right)$ в случае градиентного спуска, или $\left(\nabla^2 f\left(x^k\right)\right)^{-1}\nabla f\left(x^k\right)$ в случае демпфированного метода Ньютона.
Метод области доверия (Trust–Region) работает иначе. На каждом шаге алгоритма строится квадратичная модель m_k целевой функции f:

$$m_k(p) = f\left(x^k\right) + \left\langle \nabla f\left(x^k\right), p\right\rangle + \frac{1}{2}\left\langle p, B_k p\right\rangle,$$

где B_k — некоторая симметричная матрица, приближающая гессиан $\nabla^2 f\left(x^k\right)$. Например, в квазиньютоновских методах мы знаем аппроксимацию гессиана B_k.

Эта квадратичная модель совпадает с разложением Тейлора целевой функции f с точностью до первых двух членов. Если B_k в точности равна $\nabla^2 f(x^k)$, то совпадение в первых трех членах.
Чтобы получить шаг согласно стратегии Trust–Region решается подзадача

$$\begin{aligned} \min_{p\in\mathbb{R}^d} \quad & m_k(p) = f(x^k) + \langle \nabla f(x^k), p\rangle + \frac{1}{2}\langle p, B_k p\rangle \\ \text{s.t.} \quad & \|p\|_2 \le \Delta_k, \end{aligned} \tag{Л7.5}$$

где Δ_k — радиус области доверия. В этой подзадаче и целевая функция m_k, и ограничение, которое можно переписать как

$$\langle p, p\rangle \le \Delta_k^2,$$

являются квадратичными. В случае, когда

$$\|B_k^{-1}\nabla f(x^k)\|_2 \le \Delta_k,$$

то решение (Л7.5) просто

$$p_k^* = -B_k^{-1}\nabla f(x^k).$$

В ином случае, решение не так очевидно, однако, может быть найдено не так дорого. Но зачастую достаточно найти лишь приближенное решение для хорошей сходимости на практике.
Поговорим о стратегии подбора радиуса Δ_k. Если выбирать его слишком маленьким, шаги в направлении оптимума будут небольшими, и метод сходиться будет медленно. Если же брать Δ_k чересчур большим, то построенная модель m_k будет плохо приближать целевую функцию f вдали от нуля. Шаги будут большими, но неточными, и можем получить расходимость. Чтобы численно описывать, насколько наша модель точно приближает целевую функцию, определим соотношение

$$\rho_k = \frac{f(x^k) - f(x^k + p^k)}{m_k(0) - m_k(p^k)}. \tag{Л7.6}$$

Числитель этой дроби назовем *действительным убыванием*, а знаменатель — *предсказанным убыванием*. Поскольку p^k — точка оптимума m_k, то предсказанное убывание всегда неотрицательно. Отсюда следует, что если $\rho_k < 0$, то $f(x^k + p^k) > f(x^k)$, и такой шаг должен быть отклонен. С другой стороны, если $\rho_k \approx 1$, то модель и функция хорошо согласуются, и можно увеличивать доверительную область на последующих итерациях. Если ρ_k положительно, но не вблизи 1, то не изменяем Δ_k. Эту стратегию можно описать с помощью алгоритма.

Алгоритм Л7.4 Метод доверительной области

Вход: максимальное значение доверительного радиуса $\overline{\Delta} > 0$, начальный доверительный радиус $\Delta_0 \in (0, \overline{\Delta})$, ограничение снизу на отношение убываний $\eta \in \left[0, \frac{1}{4}\right]$

for $k = 0, \ldots, K-1$ **do**
  Найти p_k (приближённо) решением (Л7.5)
  Вычислить ρ_k по формуле (Л7.6)
  if $\rho_k < \frac{1}{4}$ **then**
    $\Delta_{k+1} = \frac{1}{4}\|p_k\|_2$
  else
    if $\rho_k > \frac{3}{4}$ **и** $\|p_k\|_2 = \Delta_k$ **then**
      $\Delta_{k+1} = \min\left(2\Delta_k, \overline{\Delta}\right)$
    else
      $\Delta_{k+1} = \Delta_k$
    end if
  end if
  if $\rho_k > \eta$ **then**
    $x_{k+1} = x_k + p_k$
  else
    $x_{k+1} = x_k$
  end if
end for

В алгоритме мы дополнительно добавили ограничение на максимальный размер шага $\overline{\Delta}$, а также ограничение на минимальное значение отношения ρ_k, при котором совершается шаг, в виде константы η.

Более подробно про приближенное решение подзадачи (Л7.5) рассказано в работе [18].

Л8 Негладкая оптимизация. Адаптивные методы

До сих пор мы рассматривали только безусловные гладкие выпуклые задачи оптимизации, то есть, задачи поиска минимума выпуклой функции с липшицевым градиентом на $\mathbb{R}^d$. Оказывается, что это достаточно сильные требования на целевую функцию, следующий пример это показывает.

Пример Л8.1. Рассмотрим функцию $f : \mathbb{R} \to \mathbb{R}$ следующего вида:

$$f(x) = |x|.$$

Эта функция достаточно простая и часто встречается, например, в виде ℓ_1-регуляризатора. При этом из-за отсутствия дифференцируемости в нуле она не является гладкой. Более того, наше определение выпуклости (Определение Л2.3) работает только для дифференцируемых функций, то есть, не применимо для модуля.

Это подталкивает нас к развитию теории для негладких задач.

Л8.1 Негладкие задачи

Хотелось бы ослабить требования на функции, в частности перестать требовать дифференцируемость целевой функции. В Параграфе Л2 было введено определение выпуклости для дифференцируемой функции (Определение Л2.3). Теперь обобщим его, не требуя дифференцируемости.

Определение Л8.1. Рассмотрим функцию $f : \mathbb{R}^d \to \mathbb{R}$. Функция f называется *выпуклой* на выпуклом множестве $\mathcal{X}$, если для любых $x_1, \ldots, x_k \in \mathbb{R}^d$ и любых $\alpha_1, \ldots, \alpha_k$, $\alpha_i \geq 0$, $\sum\limits_{i=1}^{k} \alpha_i = 1$ выполняется

$$f\left(\sum_{i=1}^{k} \alpha_i x_i\right) \leq \sum_{i=1}^{k} \alpha_i f(x_i).$$

Замечание Л8.1. В Параграфе С5 более подробно изучаются различные определения выпуклых функций. В частности, в случае дифференцируемой функции показывается, что Определения Л2.3 и Л8.1 эквивалентны (Теорема С5.1).

Требование на дифференцируемость функции осталось только в гладкости. Заменим условие L-липшицевости градиента на M-липшицевость самой функции.

Определение Л8.2. Пусть дана функция $f : \mathbb{R}^d \to \mathbb{R}$. Будем говорить, что она является *M-липшицевой*, если для любых $x, y \in \mathbb{R}^d$ выполнено:

$$|f(x) - f(y)| \leq M\|x - y\|_2.$$

Мы отказались от дифференцируемости целевой функции. Однако, можно ввести новую сущность, которая будет обладать некоторыми такими же полезными свойствами, что и градиент для дифференцируемых выпуклых функций.

Определение Л8.3. Пусть дана выпуклая функция $f : \mathbb{R}^d \to \mathbb{R}$. Вектор $g \in \mathbb{R}^d$ будем называть *субградиентом* функции f в точке $x \in \mathbb{R}^d$, если для любого

$y \in \mathbb{R}^d$ выполнено

$$f(y) \geq f(x) + \langle g, y - x \rangle.$$

Множество $\partial f(x)$ всех субградиентов f в точке x называется *субдифференциалом*.

Сформулируем критерий минимума выпуклой функции, по аналогии с тем, как было для дифференцируемых функций (Теорема Л2.1 и Теорема Л2.2).

Теорема Л8.1. Точка x^* — минимум выпуклой функции f тогда и только тогда, когда

$$0 \in \partial f(x^*).$$

Доказательство. $\Rightarrow$ Если $f(x) \geq f(x^*)$ для любых $x \in \mathcal{X}$, то для вектора 0 выполнено

$$f(x) \geq f(x^*) + \langle 0, x - x^* \rangle$$

для любого $x \in \mathbb{R}^d$. Следовательно,

$$0 \in \partial f(x^*).$$

Доказано по определению субградиента.

$\Leftarrow$ Пусть $0 \in \partial f(x^*)$. Тогда по определению субградиента и выпуклости:

$$f(x) \geq f(x^*) + \langle 0, x - x^* \rangle \geq f(x^*)$$

для любого $x \in \mathbb{R}^d$. Доказано по определению минимума на множестве.

■

Теорема Л8.2. Пусть дана выпуклая функция $f : \mathbb{R}^d \to \mathbb{R}$. Тогда функция f является M-липшицевой тогда и только тогда, когда для любого $x \in \mathbb{R}^d$ и $g \in \partial f(x)$ имеем $\|g\|_2 \leq M$.

Доказательство. $\Rightarrow$ Пусть дополнительно к выпуклости функция f еще и M-липшицева. Рассмотрим $g \in \partial f(x)$, тогда по определению субградиента (Определение Л8.3) для любого $y \in \mathbb{R}^d$:

$$f(y) - f(x) \geq \langle g, y - x \rangle.$$

Из M-липшицевости f (Определение Л8.2):

$$M\|y - x\|_2 \geq f(y) - f(x) \geq \langle g, y - x \rangle.$$

Возьмем $y = g + x \in \mathbb{R}^d$, тогда

$$M\|g\|_2 = M\|y - x\|_2 \geq \langle g, y - x \rangle = \|g\|_2^2.$$

Откуда и получаем:

$$\|g\|_2 \leq M.$$

$\Leftarrow$ Пусть дополнительно к выпуклости у функции f все субградиенты равномерно ограничены: $\|g\|_2 \leq M$ для любого $x \in \mathbb{R}^d$ и $g \in \partial f(x)$. Рассмотрим $g \in \partial f(x)$, тогда по определению субградиента (Определение Л8.3) для любого $y \in \mathbb{R}^d$:

$$f(x) - f(y) \leq \langle g, x - y \rangle.$$

Используем неравенством Коши–Буняковского–Шварца (0.3):

$$f(x) - f(y) \leq \|g\|_2 \cdot \|x - y\|_2.$$

Воспользуемся предположением об ограниченности нормы субградиента и получим:

$$f(x) - f(y) \leq M\|x - y\|_2.$$

Если переобозначить $x \to y, y \to x$, то можем записать:

$$f(y) - f(x) \leq M\|y - x\|_2.$$

Если объединить оба неравенства, то можем прийти к более краткой записи:

$$|f(y) - f(x)| \leq M\|x - y\|_2.$$

Это и есть свойство M-липшицевости f.

■

Л8.2 Субградиентный метод

Рассматриваем безусловную задачу:

$$\min_{x \in \mathbb{R}^d} f(x), \tag{Л8.1}$$

где функция f выпуклая на $\mathbb{R}^d$ и M-липшицева.
Простая идея — снова модифицировать градиентный спуск, теперь вместо градиента используя какой-то субградиент в текущей точке:

Алгоритм Л8.1 Субградиентный метод

Вход: размеры шага $\gamma > 0$, стартовая точка $x^0 \in \mathbb{R}^d$, количество итераций K
for $k = 0, 1, \ldots, K-1$ **do**
  Вычислить $g^k \in \partial f(x^k)$
  $x^{k+1} = x^k - \gamma g^k$
end for
Выход: $\frac{1}{K}\sum_{k=0}^{K-1} x^k$

Замечание Л8.2. Сходимость у субградиентного метода (Алгоритм Л8.1) по средней точке, а не по последней.

Теорема Л8.3. Пусть задача безусловной оптимизации с M-липшицевой, выпуклой целевой функцией f решается с помощью субградиентного метода (Алгоритм Л8.1). Тогда при $\gamma_k = \gamma$ справедлива следующая оценка сходимости:

$$f\left(\frac{1}{K}\sum_{k=0}^{K-1} x^k\right) - f^* \leq \frac{\|x^0 - x^*\|_2^2}{2\gamma K} + \frac{\gamma M^2}{2}.$$

Доказательство. Начнем с рассмотрения квадрата нормы $\|x^{k+1} - x^*\|_2^2$. Подставим формулу обновления

$$x^{k+1} = x^k - \gamma g^k,$$

после чего раскроем квадрат нормы:

$$\begin{aligned}\left\|x^{k+1}-x^*\right\|_2^2 =& \left\|x^k-\gamma g^k-x^*\right\|_2^2 \\ =& \left\|x^k-x^*\right\|_2^2 - 2\gamma\langle g^k, x^k-x^*\rangle + \gamma^2\left\|g^k\right\|_2^2.\end{aligned}$$

Из M-липшицевости f следует, что норма субградиента g^k ограничена M (Теорема Л8.2):

$$\left\|x^{k+1}-x^*\right\|_2^2 \le \left\|x^k-x^*\right\|_2^2 - 2\gamma\langle g^k, x^k-x^*\rangle + \gamma^2 M^2.$$

Из определения субградиента (Определение Л8.3) оценим скалярное произведение:

$$\left\|x^{k+1}-x^*\right\|_2^2 \le \left\|x^k-x^*\right\|_2^2 - 2\gamma\big(f(x^k)-f^*\big) + \gamma^2 M^2.$$

Перенесем разность значений функции f влево, а норму вправо:

$$2\gamma(f(x^k)-f^*) \le \left\|x^k-x^*\right\|_2^2 - \left\|x^{k+1}-x^*\right\|_2^2 + \gamma^2 M^2.$$

Просуммируем по всем k от 0 до $K-1$:

$$2\gamma\sum_{k=0}^{K-1}(f(x^k)-f^*) \le \sum_{k=0}^{K-1}\Big(\left\|x^k-x^*\right\|_2^2 - \left\|x^{k+1}-x^*\right\|_2^2 + \gamma^2 M^2\Big).$$

Раскроем телескопическую сумму справа, из которой остается только два слагаемых с нормой, а также разделим на $2\gamma K$:

$$\frac{1}{K}\sum_{k=0}^{K-1}(f(x^k)-f^*) \le \frac{\left\|x^0-x^*\right\|_2^2 - \left\|x^K-x^*\right\|_2^2}{2\gamma K} + \frac{\gamma M^2}{2}.$$

Можно откинуть неположительное $-\frac{\left\|x^K-x^*\right\|_2^2}{2\gamma K}$ в правой части, а также воспользоваться выпуклостью функции f (Определение Л8.1) в левой:

$$f\left(\frac{1}{K}\sum_{k=0}^{K-1}x^k\right) - f^* \le \frac{1}{K}\sum_{k=0}^{K-1}(f(x^k)-f^*) \le \frac{\left\|x^0-x^*\right\|_2^2}{2\gamma K} + \frac{\gamma M^2}{2}.$$

■

Утверждение Л8.1. Пусть задача удовлетворяет условиям Теоремы Л8.3 и выбрано значение шага $\gamma_k = \frac{\left\|x^0-x^*\right\|_2}{M\sqrt{K}}$. Тогда справедлива следующая оценка сходимости:

$$f\left(\frac{1}{K}\sum_{k=0}^{K-1}x^k\right) - f^* \le \frac{M\left\|x^0-x^*\right\|_2}{\sqrt{K}}.$$

Более того, чтобы добиться точности ε по функции $\big(f\big(\frac{1}{K}\sum_{k=0}^{K-1}x^k\big) - f^* \le \varepsilon\big)$, необходимо

$$K = \mathcal{O}\left(\frac{M^2\left\|x^0-x^*\right\|_2^2}{\varepsilon^2}\right) \text{ итераций.}$$

Доказательство. Запишем оценку сходимости из Теоремы Л8.3:

$$f\left(\frac{1}{K}\sum_{k=0}^{K-1} x^k\right) - f^* \leq \frac{\left\|x^0 - x^*\right\|_2^2}{2\gamma K} + \frac{\gamma M^2}{2}. \tag{Л8.2}$$

Минимизируя правую часть оценки по $\gamma > 0$, находим оптимальное значение шага $\gamma^* = \frac{\left\|x^0 - x^*\right\|_2}{M\sqrt{K}}$.
Подставим его в (Л8.2) и получим:

$$f\left(\frac{1}{K}\sum_{k=0}^{K-1} x^k\right) - f^* \leq \frac{M\left\|x^0 - x^*\right\|_2}{\sqrt{K}}.$$

Теперь потребуем точность ε по функции, ограничив правую часть неравенства:

$$\frac{M\left\|x^0 - x^*\right\|_2}{\sqrt{K}} \leq \varepsilon.$$

Решая это неравенство относительно K, приходим к

$$K \geq \left(\frac{M\left\|x^0 - x^*\right\|_2}{\varepsilon}\right)^2,$$

что и даёт заявленную оценку на число итераций. ■

Субградиентный метод является обобщением градиентного спуска на негладкие задачи. Сходимость получилась медленнее, чем у градиентного спуска, но можно ли улучшить результат? Например, улучшить анализ или использовать моментум. В общем случае результат для субградиентного метода является неулучшаемым для выпуклых и сильно-выпуклых задач, то есть, он оптимален.

Л8.3 Адаптивные методы

В полученной оценке для субградиентного метода шаг выбирается равным

$$\gamma = \frac{\left\|x^0 - x^*\right\|_2}{M\sqrt{K}}.$$

Однако такая формула требует знания заранее числа итераций K, константы Липшица M и расстояния $\left\|x^0 - x^*\right\|_2$ до решения. На практике эти параметры обычно неизвестны, и их использование существенно ограничивает применимость метода. Возникает естественный вопрос: как построить выбор шага, который не требует этих величин, но при этом сохраняет их смысл и обеспечивает аналогичное поведение алгоритма?

Л8.3.1 AdaGradNorm

Первым шагом можно избавиться от необходимости знать константу Липшица M. Согласно Теореме Л8.2 имеем неравенство:

$$\left\|g\right\|_2 \leq M$$

для всех $g \in \partial f(x)$ во всех точках $x \in \mathbb{R}^d$. Отсюда можно оценить

$$M \sim \left\|g\right\|_2.$$

Однако этого недостаточно: в оптимальном значении шага в знаменателе стоит $M\sqrt{K}$. Тогда мы можем брать норму не одного субградиента в какой-то точке, а суммировать по всем предыдущим:

$$M\sqrt{K} \sim \sqrt{\sum_{t=0}^{K-1} \left\|g^t\right\|_2^2}.$$

Неизвестным остается только расстояние до решения $\left\|x^0 - x^*\right\|_2$. Его можно оценить лишь какой-то константой D, на практике часто подбираемой эмпирически:

$$\left\|x^0 - x^*\right\|_2 \leq D.$$

Замечание Л8.3. Условие $\left\|x^0 - x^*\right\|_2 \leq D$ выглядит менее искусственным, если мы знаем, что решение x^* лежит в некотором ограниченном множестве $\mathcal{X}$. Тогда константу D можно оценить как диаметр множества:

$$D \leq \sup_{x,y \in \mathcal{X}} \left\|x - y\right\|_2.$$

Собирая все вместе, имеем шаг

$$\gamma_k = \frac{D}{\sqrt{\sum_{t=0}^{k} \left\|g^t\right\|_2^2 + \varepsilon}},$$

где для вычислительной устойчивости в знаменатель была добавлена константа ε, обычно выбирающаяся достаточно малой.
Таким образом шаг автоматически уменьшается по мере роста накопленной нормы градиента, и метод работает без знания M и K. Метод с таким адаптивным подбором шага называется AdaGradNorm [24].

Алгоритм Л8.2 AdaGradNorm

Вход: стартовая точка $x^0 \in \mathbb{R}^d$, сумма квадратов норм субградиентов $G^{-1} = 0$, количество итераций K, параметры $\varepsilon \sim 10^{-8}, D > 0$
1: **for** $k = 0, 1, \ldots, K-1$ **do**
2: Вычислить $g^k \in \partial f(x^k)$
3: $G^k = G^{k-1} + \left\|g^k\right\|_2^2$
4: $x^{k+1} = x^k - \frac{D}{\sqrt{G^k + \varepsilon}} g^k$
5: **end for**
Выход: $\frac{1}{K} \sum_{k=0}^{K-1} x^k$

Л8.3.2 DoG: Distance over Gradients

В методе AdaGradNorm шаг имеет вид

$$\gamma_k = \frac{D}{\sqrt{\sum_{t=0}^{k} \left\|g^t\right\|_2^2 + \varepsilon}},$$

где $D = \left\|x^0 - x^*\right\|_2$ остаётся неизвестным параметром. Для его замены предлагается использовать наблюдаемую величину

$$d_k = \max_{0 \leq t \leq k} \left\|x^t - x^0\right\|_2,$$

которая служит естественной оценкой D по траектории метода. В самом деле, мы заменили оценку на диаметр множества максимальным расстоянием, на которое удалялись от стартовой точки.
Подставляя эту оценку вместо D, получаем метод без гиперпараметров.

Алгоритм Л8.3 DoG (Distance over Gradients)

Вход: стартовая точка $x^0 \in \mathbb{R}^d$, сумма квадратов норм субградиентов $G^{-1} = 0$, оценка диаметра $d_{-1} = d_\varepsilon > 0$, количество итераций K, параметр $\varepsilon \sim 10^{-8}$

1: **for** $k = 0, 1, \ldots, K-1$ **do**
2: Вычислить $g^k \in \partial f(x^k)$
3: $G^k = G^{k-1} + \left\|g^k\right\|_2^2$
4: $d_k = \max\left(d_{k-1}, \left\|x^k - x^0\right\|_2\right)$
5: $x^{k+1} = x^k - \frac{d^k}{\sqrt{G^k + \varepsilon}} g^k$
6: **end for**

Выход: $\frac{1}{K}\sum_{k=0}^{K-1} x^k$

Л8.3.3 AdaGrad

Следующий шаг — учесть неоднородность по координатам. Тогда оптимальный шаг для каждой i-ой компоненты зависит от своей константы Липшица M_i и своего расстояния $\left|x_i^0 - x_i^*\right|$. Вместо глобального масштаба можно адаптивно накапливать информацию о каждом направлении:

$$M_i\sqrt{K} \sim \sqrt{\sum_{t=0}^{K-1}(g_i^t)^2},$$

а также оценивать расстояние до решения по каждой координате:

$$\left|x_i^0 - x_i^*\right| \le D_i.$$

Тогда шаг для каждой координаты берётся как

$$\gamma_{k,i} = \frac{D_i}{\sqrt{\sum_{t=0}^{k}(g_i^t)^2 + \varepsilon}}.$$

Такой метод учитывает частоту и величину обновлений в каждой компоненте и особенно эффективен при разреженных признаках. Этот алгоритм получил название AdaGrad [7].

Алгоритм Л8.4 AdaGrad

Вход: стартовая точка $x^0 \in \mathbb{R}^d$, суммы квадратов норм субградиентов $G_i^{-1} = 0$, количество итераций K, параметры $\varepsilon \sim 10^{-8}, D_i > 0$

1: **for** $k = 0, 1, \ldots, K-1$ **do**
2: Вычислить $g^k \in \partial f\left(x^k\right)$
3: $G_i^k = G_i^{k-1} + \left(g_i^k\right)^2$
4: $x_i^{k+1} = x_i^k - \frac{D_i}{\sqrt{G_i^k + \varepsilon}} g_i^k$
5: **end for**

Выход: $\frac{1}{K}\sum_{k=0}^{K-1} x^k$

Для алгоритма сформулируем и докажем теорему о сходимости.

Теорема Л8.4. Пусть задача оптимизации (Л8.1) с M-липшицевой, выпуклой целевой функцией f решается с помощью AdaGrad (Алгоритм Л8.4). Тогда справедлива следующая оценка сходимости:

$$f\left(\frac{1}{K}\sum_{k=0}^{K-1} x^k\right) - f^* \leq \frac{3M\tilde{D}}{2\sqrt{K}},$$

где $\tilde{D} = \sum_{i=1}^{d} D_i, \left|x_i^k - x_i^*\right| \leq D_i \ \forall i = \overline{1,d}, k = \overline{0,K-1}$.
Более того, чтобы добиться точности ε по функции ($f(x^K) - f^* \leq \varepsilon$), необходимо

$$K = \mathcal{O}\left(\frac{M^2\tilde{D}^2}{\varepsilon^2}\right) \text{ итераций.}$$

Доказательство. Начнем с квадрата ошибки по i-ой координате $\left(x_i^{k+1} - x_i^*\right)^2$. Подставим итерацию метода:

$$\begin{aligned}\left(x_i^{k+1} - x_i^*\right)^2 &= \left(x_i^k - \gamma_{k,i} g_i^k - x_i^*\right)^2 \\ &= \left(x_i^k - x_i^*\right)^2 - 2\gamma_{k,i} g_i^k (x_i^k - x_i^*) + \gamma_{k,i}^2 \left(g_i^k\right)^2.\end{aligned}$$

Откуда выражаем перекрестное слагаемое:

$$g_i^k (x_i^k - x_i^*) = \frac{1}{2\gamma_{k,i}}\left(x_i^k - x_i^*\right)^2 - \frac{1}{2\gamma_{k,i}}\left(x_i^{k+1} - x_i^*\right)^2 + \frac{\gamma_{k,i}}{2}\left(g_i^k\right)^2.$$

Собираем скалярное произведение, суммируя по всем координатам i от 1 до d:

$$\left\langle g^k, x^k - x^*\right\rangle = \frac{1}{2}\sum_{i=1}^{d}\left[\frac{1}{\gamma_{k,i}}\left(x_i^k - x_i^*\right)^2 - \frac{1}{\gamma_{k,i}}\left(x_i^{k+1} - x_i^*\right)^2 + \gamma_{k,i}\left(g_i^k\right)^2\right].$$

Оцениваем левую часть, пользуясь определением субградиента (Определение Л8.3):

$$f(x^k) - f^* \leq \frac{1}{2}\sum_{i=1}^{d}\left[\frac{1}{\gamma_{k,i}}\left(x_i^k - x_i^*\right)^2 - \frac{1}{\gamma_{k,i}}\left(x_i^{k+1} - x_i^*\right)^2 + \gamma_{k,i}\left(g_i^k\right)^2\right].$$

Суммируем по всем k от 0 до $K-1$ и усредняем, а также меняем порядок суммирования:

$$\begin{aligned}\frac{1}{K}\sum_{k=0}^{K-1}\left(f(x^k) - f^*\right) &\leq \frac{1}{2K}\sum_{k=0}^{K-1}\sum_{i=1}^{d}\left[\frac{1}{\gamma_{k,i}}\left(x_i^k - x_i^*\right)^2 - \frac{1}{\gamma_{k,i}}\left(x_i^{k+1} - x_i^*\right)^2 + \gamma_{k,i}\left(g_i^k\right)^2\right] \\ &= \frac{1}{2K}\sum_{i=1}^{d}\sum_{k=0}^{K-1}\left[\frac{1}{\gamma_{k,i}}\left(x_i^k - x_i^*\right)^2 - \frac{1}{\gamma_{k,i}}\left(x_i^{k+1} - x_i^*\right)^2 + \gamma_{k,i}\left(g_i^k\right)^2\right].\end{aligned}$$

Следующим шагом будет приведение подобных слагаемых. Для этого сдвинем индекс суммирования k и аккуратно поработаем с крайними слагаемыми. Для удобства по-

ложим $\gamma_{-1,i} = +\infty$ и преобразуем:

$$\begin{aligned}
\frac{1}{K}\sum_{k=0}^{K-1}\big(f(x^k)-f^*\big) &\le \frac{1}{2K}\sum_{i=1}^{d}\sum_{k=0}^{K-1}\left[\frac{1}{\gamma_{k,i}}\big(x_i^k-x_i^*\big)^2+\gamma_{k,i}\big(g_i^k\big)^2\right]\\
&\quad-\frac{1}{2K}\sum_{i=1}^{d}\sum_{k=0}^{K-1}\left[\frac{1}{\gamma_{k,i}}\big(x_i^{k+1}-x_i^*\big)^2\right]\\
&=\frac{1}{2K}\sum_{i=1}^{d}\sum_{k=0}^{K-1}\left[\frac{1}{\gamma_{k,i}}\big(x_i^k-x_i^*\big)^2+\gamma_{k,i}\big(g_i^k\big)^2\right]\\
&\quad-\frac{1}{2K}\sum_{i=1}^{d}\sum_{k=0}^{K-1}\left[\frac{1}{\gamma_{k-1,i}}\big(x_i^k-x_i^*\big)^2\right]-\frac{1}{2K}\sum_{i=1}^{d}\frac{1}{\gamma_{K-1,i}}(x_i^K-x_i^*)^2\\
&\le\frac{1}{2K}\sum_{i=1}^{d}\sum_{k=0}^{K-1}\left[\left(\frac{1}{\gamma_{k,i}}-\frac{1}{\gamma_{k-1,i}}\right)\big(x_i^k-x_i^*\big)^2+\gamma_{k,i}\big(g_i^k\big)^2\right].
\end{aligned}$$

В последнем переходе мы откинули неположительное слагаемое. Теперь воспользуемся константами $D_i : \big(x_i^k - x_i^*\big)^2 \le D_i^2$:

$$\frac{1}{K}\sum_{k=0}^{K-1}\big(f(x^k)-f^*\big)\le\frac{1}{2K}\sum_{i=1}^{d}\sum_{k=0}^{K-1}\left[\left(\frac{1}{\gamma_{k,i}}-\frac{1}{\gamma_{k-1,i}}\right)D_i^2+\gamma_{k,i}\big(g_i^k\big)^2\right].$$

Подставляем $\gamma_{k,i}=\frac{D_i}{\sqrt{\sum_{t=0}^{k}(g_i^t)^2}}$ и сворачиваем телескопическую сумму по k:

$$\begin{aligned}
\frac{1}{K}\sum_{k=0}^{K-1}\big(f(x^k)-f^*\big)&\le\frac{1}{2K}\sum_{i=1}^{d}\sum_{k=0}^{K-1}\left[\left(\sqrt{\sum_{t=0}^{k}\big(g_i^t\big)^2}-\sqrt{\sum_{t=0}^{k-1}\big(g_i^t\big)^2}\right)D_i+\frac{D_i\big(g_i^k\big)^2}{\sqrt{\sum_{t=0}^{k}\big(g_i^t\big)^2}}\right]\\
&=\frac{1}{2K}\sum_{i=1}^{d}D_i\left[\sqrt{\sum_{t=0}^{K-1}\big(g_i^t\big)^2}+\sum_{k=0}^{K-1}\frac{\big(g_i^k\big)^2}{\sqrt{\sum_{t=0}^{k}\big(g_i^t\big)^2}}\right].
\end{aligned}$$

Воспользуемся техническим фактом, который говорит, что для любых неотрицательных чисел $\{a_k\} \ge 0$ выполнено:

$$\sum_{k=0}^{K-1}\frac{\big(a_k\big)^2}{\sqrt{\sum_{t=0}^{k}\big(a_i^k\big)^2}}\le 2\sqrt{\sum_{k=0}^{K-1}\big(a_i^k\big)^2}.$$

Применяем его для второго слагаемого в сумме, получаем:

$$\frac{1}{K}\sum_{k=0}^{K-1}\big(f(x^k)-f^*\big)\le\frac{3}{2K}\sum_{i=1}^{d}D_i\sqrt{\sum_{t=0}^{K-1}\big(g_i^t\big)^2}.$$

Из неизвестных значений в правой части остается только $\sum_{t=0}^{K-1}\big(g_i^t\big)^2$. Оценим её с помощью свойства M-липшицевых функций (Теорема Л8.2):

$$\sum_{t=0}^{K-1}\big(g_i^t\big)^2\le\sum_{t=0}^{K-1}\big\|g^t\big\|_2^2\le KM^2.$$

Подставляем в основное неравенство:

$$\frac{1}{K}\sum_{k=0}^{K-1}\left(f(x^k)-f^*\right) \le \frac{3}{2K}\sum_{i=1}^{d} D_i\sqrt{KM^2} = \frac{3M}{2\sqrt{K}}\sum_{i=1}^{d} D_i = \frac{3M\tilde{D}}{2\sqrt{K}}.$$

Чтобы получить оценку из условия, остается применить определение выпуклости функции (Определение Л8.1):

$$f\left(\frac{1}{K}\sum_{k=0}^{K-1} x^k\right) - f^* \le \frac{3M\tilde{D}}{2\sqrt{K}}.$$

Чтобы получить оценку на количество итераций для достижения точности ε по функции ($f\left(\frac{1}{K}\sum_{k=0}^{K-1} x^k\right) - f^* \le \varepsilon$), ограничим правую часть:

$$\frac{3M\tilde{D}}{2\sqrt{K}} \le \varepsilon.$$

Откуда и получаем:

$$K \ge \frac{9M^2\tilde{D}^2}{4\varepsilon^2}.$$

■

Замечание Л8.4. Константы D_i могут выглядеть немного искусственными, однако, их можно обосновать. Например, когда нам известно, что решение x^* лежит в ограниченном множестве $\mathcal{X}$, а стартовая и все последующие точки также лежат внутри этого множества.

Замечание Л8.5. Константы D_i можно также оценивать динамически. В частности, можно использовать идею DoG, адаптированную под координатную запись:

$$d_{k,i} = \max_{0\le t\le k} |x_i^t - x_i^0|.$$

Л8.3.4 RMSProp

У AdaGrad знаменатель монотонно растёт, и шаг со временем может становиться слишком малым. Чтобы избежать затухания шага, вместо суммы используют экспоненциальное скользящее среднее квадратов градиентов:

$$G_i^k = \beta G_i^{k-1} + (1-\beta)(g_i^k)^2,$$

где моментум $\beta \in [0,1]$ выбирается порядка ~ 0.99. Такой приём позволяет учитывать преимущественно недавнюю историю и адаптироваться к меняющейся геометрии задачи. Метод получил название RMSProp [23].

Алгоритм Л8.5 RMSProp

Вход: стартовая точка $x^0 \in \mathbb{R}^d$, сглаженные суммы квадратов норм субградиентов $G_i^{-1} = 0$, количество итераций K, параметры $\varepsilon \sim 10^{-8}, D_i > 0, \beta \in [0,1]$

for $k = 0, 1, \ldots, K-1$ **do**
  Вычислить $g^k \in \partial f(x^k)$
  $G_i^k = \beta G_i^{k-1} + (1-\beta)(g_i^k)^2$
  $x_i^{k+1} = x_i^k - \frac{D_i}{\sqrt{G_i^k+\varepsilon}} g_i^k$
end for

Выход: $\frac{1}{K}\sum_{k=0}^{K-1} x^k$

Л8.3.5 Adam

Алгоритм Adam объединяет идеи RMSProp и момента. Помимо экспоненциального сглаживания квадратов градиентов, он усредняет сами градиенты:

$$v^k = \beta_1 v^{k-1} + (1-\beta_1) g^k,$$
$$G_i^k = \beta_2 G_i^{k-1} + (1-\beta_2)(g_i^k)^2,$$

где два моментума $\beta_1, \beta_2 \in [0,1]$, обычно выбираются 0.9 и 0.999 соответственно. Чтобы устранить смещение на ранних итерациях, вводится коррекция:

$$\hat{v}^k = \frac{v^k}{1-\beta_1^{k+1}},$$
$$\hat{G}^k = \frac{G^k}{1-\beta_2^{k+1}}.$$

В итоге шаг получается

$$x_i^{k+1} = x_i^k - \frac{D_i}{\sqrt{\hat{G}_i^k + \varepsilon}} \hat{v}_i^k.$$

Такой метод даёт устойчивые шаги при шумных и разреженных градиентах и требует минимальной настройки гиперпараметров [8].

Алгоритм Л8.6 Adam

Вход: стартовая точка $x^0 \in \mathbb{R}^d$, суммы $v^{-1} = 0, G_i^{-1} = 0$, количество итераций K, параметры $\beta_1, \beta_2 \in [0,1], \varepsilon \sim 10^{-8}, D_i > 0$

1: **for** $k = 0, 1, \ldots, K-1$ **do**
2: Вычислить $g^k \in \partial f(x^k)$
3: $v^k = \beta_1 v^{k-1} + (1-\beta_1) g^k$
4: $\hat{v}^k = v^k / (1-\beta_1^{k+1})$
5: $G_i^k = \beta_2 G_i^{k-1} + (1-\beta_2)(g_i^k)^2$
6: $\hat{G}^k = G^k / (1-\beta_2^{k+1})$
7: $x_i^{k+1} = x_i^k - \frac{D_i}{\sqrt{\hat{G}_i^k + \varepsilon}} \hat{v}_i^k$
8: **end for**

Выход: $\frac{1}{K} \sum_{k=0}^{K-1} x^k$

Л8.3.6 NAdam

Вариант NAdam сочетает Adam с идеей ускорения Нестерова. То есть, сначала делаем шаг по инерции, потом вычисляем субградиент и делаем шаг вдоль него [6].

Алгоритм Л8.7 NAdam

Вход: стартовая точка $x^0 \in \mathbb{R}^d$, суммы $v^{-1} = 0, G_i^{-1} = 0$, количество итераций K, параметры $\beta_1, \beta_2 \in [0,1], \varepsilon \sim 10^{-8}, D_i > 0$

1: **for** $k = 0, 1, \ldots, K-1$ **do**
2: Вычислить $g^k \in \partial f(x^k)$
3: $v^k = \beta_1 v^{k-1} + (1-\beta_1) g^k$
4: $\hat{v}^k = v^k / (1 - \beta_1^{k+1})$
5: $G_i^k = \beta_2 G_i^{k-1} + (1-\beta_2)(g_i^k)^2$
6: $\hat{G}^k = G^k / (1 - \beta_2^{k+1})$
7: $m^k = \beta_1 \hat{v}^k + \frac{1-\beta_1}{1-\beta_1^{k+1}} g^k$
8: $x_i^{k+1} = x_i^k - \frac{D_i}{\sqrt{\hat{G}_i^k + \varepsilon}} m_i^k$
9: **end for**

Выход: $\frac{1}{K} \sum_{k=0}^{K-1} x^k$

Л8.3.7 AdamW

В классическом Adam если применяется ℓ_2-регуляризация, то градиент от слагаемого $\frac{\lambda}{2}\|x\|_2^2$ добавляется в общий градиент и масштабируется адаптивным шагом, что делает её неэквивалентной классическому weight decay. AdamW отделяет регуляризацию: стягивание весов $-\lambda x^k$ добавляется отдельно от градиентного обновления. Это позволяет контролировать регуляризацию независимо от величины шага и улучшает обобщающую способность моделей [14].

Алгоритм Л8.8 AdamW

Вход: стартовая точка $x^0 \in \mathbb{R}^d$, параметры $\beta_1 = 0.9, \beta_2 = 0.99, \lambda > 0$, суммы $v^{-1} = 0, G_i^{-1} = 0$, количество итераций K, параметры $\varepsilon \sim 10^{-8}, D_i > 0$

1: **for** $k = 0, 1, \ldots, K-1$ **do**
2: Вычислить $g^k \in \partial f(x^k)$
3: $v^k = \beta_1 v^{k-1} + (1-\beta_1) g^k$
4: $\hat{v}^k = v^k / (1 - \beta_1^{k+1})$
5: $G_i^k = \beta_2 G_i^{k-1} + (1-\beta_2)(g_i^k)^2$
6: $\hat{G}^k = G^k / (1 - \beta_2^{k+1})$
7: $x_i^{k+1} = x_i^k - \frac{D_i}{\sqrt{\hat{G}_i^k + \varepsilon}} \hat{v}_i^k - \lambda x_i^k$
8: **end for**

Выход: $\frac{1}{K} \sum_{k=0}^{K-1} x^k$

Л8.4 Проксимальный оператор

Негладкие задачи «более сложные» по сравнению с гладкими задачами. Может быть получится «спрятать под ковер» отсутствие гладкости? Такую возможность дает проксимальный оператор.

Определение Л8.4. Для функции $r : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ проксимальный оператор определяется следующим образом:

$$\operatorname{prox}_r(x) = \operatorname*{argmin}_{\tilde{x} \in \mathbb{R}^d} \left(r(\tilde{x}) + \frac{1}{2}\|x - \tilde{x}\|_2^2 \right).$$

Пока что мы не потребовали ничего от функции r и мало что можем сказать про проксимальный оператор. Оказывается, что если наложить на r условие выпуклости, то проксимальный оператор для нее становится однозначно определён.

Утверждение Л8.2. Пусть $r : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ — выпуклая функция, для которой определен prox_r. Если существует $\hat{x} \in \mathbb{R}^d$ с $r(\hat{x}) < +\infty$, то проксимальный оператор однозначно определен.

Доказательство. Проксимальный оператор возвращает минимум некоторой задачи оптимизации. Задача сильно выпуклая, так как целевая функция — это сумма выпуклой и сильно выпуклой функции. А значит существует строго один минимум (Теорема Л2.3). А существование $\hat{x}$ необходимо для того, чтобы $r(\tilde{x}) + \frac{1}{2}\|x - \tilde{x}\|_2^2$ хотя бы где-то принимала конечное значение. ■

Л8.4.1 Примеры проксимального оператора

Рассмотрим, какие операторы мы будем получать в зависимости от того, какую проксимальную функцию мы используем.

- ℓ_1**-норма.**
  Пусть $r(x) = \lambda\|x\|_1$, где $\lambda > 0$. Тогда

  $$[\text{prox}_r(x)]_i = \big[|x_i| - \lambda\big]_+ \cdot \text{sign}(x_i),$$

  что известно как *оператор трешхолдинга.*

  Доказательство. Рассмотрим задачу

  $$\begin{aligned}\text{prox}_r(x) &= \operatorname*{argmin}_{\tilde{x}\in\mathbb{R}^d}\left(\lambda\|\tilde{x}\|_1 + \frac{1}{2}\|x - \tilde{x}\|_2^2\right)\\ &= \operatorname*{argmin}_{\tilde{x}\in\mathbb{R}^d}\sum_{i=1}^{d}\left(\frac{1}{2}(\tilde{x}_i - x_i)^2 + \lambda|\tilde{x}_i|\right).\end{aligned}$$

  Задача распадается по координатам. Для каждой координаты условие оптимальности имеет вид

  $$0 \in \tilde{x}_i^* - x_i + \lambda \cdot \partial|\tilde{x}_i^*|,$$

  что приводит к правилу софт-трешхолдинга:

  $$\tilde{x}_i^* = \big[|x_i| - \lambda\big]_+ \cdot \text{sign}(x_i).$$

  ■

- ℓ_2**-норма.**
  Пусть $r(x) = \frac{\lambda}{2}\|x\|_2^2$, где $\lambda > 0$. Тогда

  $$\text{prox}_r(x) = \frac{x}{1+\lambda}.$$

Доказательство. Рассмотрим

$$\operatorname{prox}_r(x) = \operatorname*{argmin}_{\tilde{x}\in\mathbb{R}^d} \left(\frac{\lambda}{2}\|\tilde{x}\|_2^2 + \frac{1}{2}\|x-\tilde{x}\|_2^2 \right).$$

Необходимое условие минимума (Теорема Л2.2):

$$\lambda\tilde{x}^* + \tilde{x}^* - x = 0.$$

Откуда и получаем решение:

$$\tilde{x}^* = \frac{x}{1+\lambda}.$$

■

- **Индикатор множества.**
  Пусть $r(x) = \mathbb{I}_{\mathcal{X}}(x)$ для замкнутого множества $\mathcal{X} \subseteq \mathbb{R}^d$, где

$$\mathbb{I}_{\mathcal{X}}(x) = \begin{cases} 0, & x \in \mathcal{X}, \\ +\infty, & x \notin \mathcal{X}. \end{cases}$$

  Тогда

$$\operatorname{prox}_r(x) = \operatorname*{argmin}_{\tilde{x}\in\mathcal{X}} \frac{1}{2}\|x-\tilde{x}\|_2^2.$$

  Смысл такого оператора в том, что он возвращает наиближайшую к x точку из множества $\mathcal{X}$.

  Доказательство. Имеем

$$\begin{aligned} \operatorname{prox}_r(x) &= \operatorname*{argmin}_{\tilde{x}\in\mathbb{R}^d} \left(\mathbb{I}_{\mathcal{X}}(\tilde{x}) + \frac{1}{2}\|x-\tilde{x}\|_2^2 \right) \\ &= \operatorname*{argmin}_{\tilde{x}\in\mathcal{X}} \frac{1}{2}\|x-\tilde{x}\|_2^2. \end{aligned}$$

  ■

Л8.4.2 Свойства проксимального оператора

Для последующего теоретического анализа алгоритмов на основе проксимального оператора, докажем полезные свойства.

Утверждение Л8.3. Пусть $r : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ — выпуклая функция, для которой определен prox_r. Тогда для любых $x, y \in \mathbb{R}^d$ следующие три условия являются эквивалентными:

1. $\operatorname{prox}_r(x) = y$,
2. $x - y \in \partial r(y)$,
3. $\langle x-y, z-y\rangle \leq r(z) - r(y)$ для любого $z \in \mathbb{R}^d$.

Доказательство. Запишем первое условие по определению проксимального оператора (Определение Л8.4):

$$y = \operatorname*{argmin}_{\tilde{x}\in\mathbb{R}^d} \left(r(\tilde{x}) + \frac{1}{2}\|x - \tilde{x}\|_2^2 \right).$$

Из условия оптимальности для выпуклой функции r (Теорема Л8.1) это эквивалентно

$$0 \in \partial \left(r(\tilde{x}) + \frac{1}{2}\|x - \tilde{x}\|_2^2 \right)\Big|_{\tilde{x}=y} = \partial r(y) + y - x.$$

Получили эквивалентность первого и второго условий.
Теперь свяжем второе и третье. Из определения субдифференциала (Определение Л8.3), для любого субградиента $g \in \partial f(y)$ и для любого $z \in \mathbb{R}^d$:

$$\langle g, z - y\rangle \le r(z) - r(y).$$

В частности справедливо и для $g = x - y \in \partial r(y)$ по предположению:

$$\langle x - y, z - y\rangle \le r(z) - r(y).$$

Получили следствие из 2 в 3. Теперь в обратную сторону. Пусть выполняется соотношение 3:

$$\langle x - y, z - y\rangle \le r(z) - r(y).$$

Но это же означает, что $x-y$ является субградиентом r в точке y, то есть, $g \in \partial r(y)$. ■

Этим свойством мы связали действие проксимального оператора и субградиент. Следующее утверждение позволит работать с действиями проксимального оператора под знаками скалярного произведения и нормы.

Утверждение Л8.4. Пусть $r : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ — выпуклая функция, для которой определен prox_r. Тогда для любых $x, y \in \mathbb{R}^d$ выполнено следующее:

- $\langle x - y, \operatorname{prox}_r(x) - \operatorname{prox}_r(y)\rangle \ge \|\operatorname{prox}_r(x) - \operatorname{prox}_r(y)\|_2^2$,
- $\|\operatorname{prox}_r(x) - \operatorname{prox}_r(y)\|_2 \le \|x - y\|_2$.

Доказательство. Для удобства обозначим $u = \operatorname{prox}_r(x), v = \operatorname{prox}_r(y)$. Начнем с неравенства из Утверждения Л8.3. Запишем для троек точек (x, u, v) и (y, v, u):

$$\langle x - u, v - u\rangle \le r(v) - r(u),$$
$$\langle y - v, u - v\rangle \le r(u) - r(v).$$

Сложим неравенства, справа получим 0:

$$\langle x - u - (y - v), v - u\rangle \le 0.$$

Разнесем слагаемые в разные стороны:

$$\langle x - y, u - v\rangle \ge \|u - v\|_2^2.$$

Это и есть первое свойство. Из него логично вытекает и второе свойство. Достаточно лишь применить к неравенству выше неравенство Коши–Буняковского–Шварца (0.3):

$$\|u - v\|_2^2 \le \langle x - y, u - v\rangle \le \|x - y\|_2 \|u - v\|_2.$$

Случай $u = v$ тривиален и удовлетворяет неравенству, так что имеем:

$$\|u - v\|_2 \le \|x - y\|_2.$$

■

Л8.4.3 Проксимальный градиентный метод

Настало время применить проксимальный оператор для построения метода оптимизации. Рассмотрим следующую задачу:

$$\min_{x\in\mathbb{R}^d}\big[f(x)+r(x)\big], \tag{Л8.3}$$

где f является L-гладкой выпуклой функцией, а r — просто выпуклой. Она не обязана быть гладкой, но должна быть проксимально-дружественной, то есть, для нее должен быть определен проксимальный оператор. Задача (Л8.3) называется *композитной*. Вновь модифицируем градиентный спуск: добавим применение проксимального оператора $\mathrm{prox}_{\gamma r}$ после градиентного шага $-\gamma\nabla f(x^k)$.

Алгоритм Л8.9 Проксимальный градиентный метод

Вход: стартовая точка $x^0\in\mathbb{R}^d$, размеры шагов $\{\gamma_k\}_{k=0}>0$, количество итераций K

1: **for** $k=0,1,\dots,K-1$ **do**
2: $\quad x^{k+1}=\mathrm{prox}_{\gamma r}(x^k-\gamma\nabla f(x^k))$
3: **end for**

Выход: x^K

Покажем, какую физику несёт метод. Для этого воспользуемся результатами Утверждения Л8.3, тогда $x^{k+1}=\mathrm{prox}_{\gamma r}(x^k-\gamma\nabla f(x^k))$ эквивалентно записи:

$$(x^k-\gamma\nabla f(x^k))-x^{k+1}\in\gamma\partial r(x^{k+1}).$$

Или, если переписать в более привычном виде:

$$x^{k+1}\in x^k-\gamma\big(\nabla f(x^k)+\partial r(x^{k+1})\big).$$

Получили неявную запись метода, из которой видно, что мы делаем шаг вдоль $\nabla f(x^k)+g$, где $g\in\partial r(x^{k+1})$.

Утверждение Л8.5. Пусть $f:\mathbb{R}^d\to\mathbb{R}\cup\{+\infty\}$, $r:\mathbb{R}^d\to\mathbb{R}\cup\{+\infty\}$ — выпуклые функции. Дополнительно предположим, что f является непрерывно дифференцируемой и L-гладкой, а для r определен prox_r. Тогда x^* — решение композитной задачи оптимизации тогда и только тогда, когда для любого $\gamma>0$ выполнено:

$$x^*=\mathrm{prox}_{\gamma r}(x^*-\gamma\nabla f(x^*)).$$

Доказательство. Обратимся к Утверждению Л8.3. Согласно нему, выражение

$$x^*=\mathrm{prox}_{\gamma r}(x^*-\gamma\nabla f(x^*))$$

эквивалентно записи:

$$\big(x^*-\gamma\nabla f(x^*)\big)-x^*\in\gamma\partial r(x^*).$$

Сокращаем на $\gamma>0$ и получаем условие оптимальности:

$$0\in\nabla f(x^*)+\partial r(x^*).$$

■

Теперь мы готовы доказывать теорему о скорости сходимости.

Теорема Л8.5. Пусть композитная задача оптимизации (Л8.3) с L-гладкой, μ-сильно выпуклой целевой функцией f и выпуклой (необязательно гладкой, но) проксимально дружественной функцией r решается с помощью проксимального градиентного метода (Алгоритм Л8.9). Тогда при $\gamma_k \leq \frac{1}{L}$ справедлива следующая оценка сходимости:

$$\|x^K - x^*\|_2^2 \leq \left(\prod_{k=0}^{K-1} (1 - \mu\gamma_k)\right) \|x^0 - x^*\|_2^2.$$

Доказательство. Рассматриваем $\|x^{k+1} - x^*\|_2^2$. Подставим итерацию метода:

$$\|x^{k+1} - x^*\|_2^2 = \|\mathrm{prox}_{\gamma r}(x^k - \gamma_k \nabla f(x^k)) - x^*\|_2^2.$$

Воспользуемся Утверждением Л8.5:

$$\|x^{k+1} - x^*\|_2^2 = \|\mathrm{prox}_{\gamma r}(x^k - \gamma_k \nabla f(x^k)) - \mathrm{prox}_{\gamma r}(x^* - \gamma_k \nabla f(x^*))\|_2^2.$$

А теперь применяем неравенство из Утверждения Л8.4:

$$\|x^{k+1} - x^*\|_2^2 \leq \|(x^k - \gamma_k \nabla f(x^k)) - (x^* - \gamma_k \nabla f(x^*))\|_2^2.$$

Раскрываем квадрат нормы:

$$\|x^{k+1} - x^*\|_2^2 \leq \|x^k - x^*\|_2^2 - 2\gamma_k \langle x^k - x^*, \nabla f(x^k) - \nabla f(x^*)\rangle + \gamma_k^2 \|\nabla f(x^k) - \nabla f(x^*)\|.$$

Оценим последнее слагаемое, используя свойство L-гладкой выпуклой f (неравенство (Л2.5) из Теоремы Л2.5):

$$\begin{aligned}\|x^{k+1} - x^*\|_2^2 &\leq \|x^k - x^*\|_2^2 - 2\gamma_k \langle x^k - x^*, \nabla f(x^k) - \nabla f(x^*)\rangle \\ &\quad + \gamma_k^2 L \langle \nabla f(x^k) - \nabla f(x^*), x^k - x^*\rangle \\ &= \|x^k - x^*\|_2^2 - \gamma_k(2 - \gamma_k L)\langle \nabla f(x^k) - \nabla f(x^*), x^k - x^*\rangle.\end{aligned}$$

Два раза запишем определение μ-сильной выпуклости f (Определение Л2.4):

$$\begin{aligned} f(x^k) &\geq f(x^*) + \langle \nabla f(x^*), x^k - x^*\rangle + \frac{\mu}{2}\|x^k - x^*\|_2^2, \\ f(x^*) &\geq f(x^k) + \langle \nabla f(x^k), x^* - x^k\rangle + \frac{\mu}{2}\|x^* - x^k\|_2^2. \end{aligned} \tag{Л8.4}$$

Сложим их, получим:

$$\langle \nabla f(x^k) - \nabla f(x^*), x^k - x^*\rangle \geq \mu\|x^k - x^*\|_2^2.$$

Подставляем в основное неравенство (Л8.4):

$$\begin{aligned}\|x^{k+1} - x^*\|_2^2 &\leq \|x^k - x^*\|_2^2 - \mu\gamma_k(2 - \gamma_k L)\|x^k - x^*\|_2^2 \\ &= \big(1 - \mu\gamma_k(2 - \gamma_k L)\big)\|x^k - x^*\|_2^2.\end{aligned}$$

Поскольку $\gamma_k \leq \frac{1}{L}$, то $2 - \gamma_k L > 1$. Продолжая цепочку неравенств, имеем:

$$\|x^{k+1} - x^*\|_2^2 \leq \big(1 - \mu\gamma_k\big)\|x^k - x^*\|_2^2.$$

Остается лишь развернуть рекурсию с K-ой итерации по нулевую:

$$\begin{aligned}\left\|x^K - x^*\right\|_2^2 &\leq \left(1 - \mu\gamma_{K-1}\right)\left\|x^{K-1} - x^*\right\|_2^2 \\ &\leq \left(1 - \mu\gamma_{K-1}\right)\left(1 - \mu\gamma_{K-2}\right)\left\|x^{K-2} - x^*\right\|_2^2 \\ &\leq \dots \\ &\leq \left(\prod_{k=0}^{K-1}\left(1 - \mu\gamma_k\right)\right)\left\|x^0 - x^*\right\|_2^2.\end{aligned}$$

■

Утверждение Л8.6. Пусть задача удовлетворяет условиям Теоремы Л8.5 и выбрано значение шага $\gamma_k = \frac{1}{L}$. Тогда справедлива следующая оценка сходимости:

$$\left\|x^K - x^*\right\|_2^2 \leq \left(1 - \frac{\mu}{L}\right)^K \left\|x^0 - x^*\right\|_2^2.$$

Более того, чтобы добиться точности ε по аргументу $\left(\left\|x^K - x^*\right\|_2 \leq \varepsilon\right)$, необходимо

$$K = \mathcal{O}\left(\frac{L}{\mu}\log\frac{\left\|x^0 - x^*\right\|_2}{\varepsilon}\right) = \tilde{\mathcal{O}}\left(\frac{L}{\mu}\right) \text{ итераций.}$$

Доказательство. Повторяет доказательство Утверждения Л3.1. ■

Замечание Л8.6. Проксимальный градиентный спуск для композитной задачи с L-гладкой μ-сильно выпуклой функцией f и выпуклой проксимально-дружественной функцией r имеет такую же скорость сходимости, что и метод градиентного спуска для функции f (Теорема Л3.1). Свойства гладкости/негладкости r при этом не влияют.

Кажется, что положив $f \equiv 0$, с помощью такого метода можно решать любую негладкую задачу. Действительно, если разрешить приближенное вычисление проксимального оператора, то формально можно решать любую выпуклую негладкую задачу. Однако с теоретической точки зрения это не приводит к преимуществам по сравнению с субградиентным методом: для нахождения значения проксимального оператора приходится решать подзадачу, которая сама по себе требует использования дополнительного метода (например, того же субградиентного спуска).

Л9 Метод зеркального спуска

Обратимся к итерации метода градиентного спуска (Алгоритм Л3.1):

$$x^{k+1} = x^k - \gamma_k \nabla f\big(x^k\big).$$

Здесь точка x рассматривается как элемент банахова пространства $\big(E, \|\cdot\|\big)$. При этом градиент $\nabla f\big(x^k\big)$ принадлежит сопряжённому пространству $\big(E^*, \|\cdot\|_*\big)$, а не исходному $\big(E, \|\cdot\|\big)$. Таким образом, в общем случае итерация складывает объекты из разных пространств. В ситуации, когда $\|\cdot\| = \|\cdot\|_2$, это несоответствие не возникает: пространства совпадают, то есть $\big(E, \|\cdot\|\big) = \big(E^*, \|\cdot\|_*\big)$, и формула градиентного спуска корректна.
Однако измерять расстояние вовсе не обязательно в евклидовой норме. Даже несмотря на то, что в конечномерных пространствах все нормы эквивалентны, для различных задач могут быть предпочтительны более «естественные» способы измерения расстояния. Например, при работе с распределениями вероятностей использование евклидовой нормы выглядит искусственным.
Эти соображения приводят к идее, предложенной А. Немировским и Д. Юдиным: выполнять шаг градиентного спуска не в исходном пространстве, а в сопряжённом. Для этого вводится отображение $\varphi : E \to E^*$, и итерация метода принимает вид

$$\varphi(x^{k+1}) = \varphi(x^k) - \gamma \nabla f(x^k),$$

где обратное отображение $\varphi^{-1} : E^* \to E$ возвращает результат обратно в исходное пространство E.

Замечание Л9.1. Пространство E^* в этом контексте называют «зеркальным». Именно в нём выполняется шаг градиентного спуска, а затем с помощью φ^{-1} результат отображается обратно в исходное пространство. Для практического применения метода необходимо иметь явную форму для вычисления x^{k+1}.

Л9.1 Дивергенция Брэгмана

Обобщим определение μ-сильной выпуклости на случай произвольной нормы.

Определение Л9.1. Пусть дана непрерывно дифференцируемая на выпуклом множестве $\mathcal{X}$ функция $d : \mathcal{X} \to \mathbb{R}$. Будем говорить, что она является *μ-сильно выпуклой* ($\mu > 0$) относительно нормы $\|\cdot\|$ на множестве $\mathcal{X}$, если для любых $x, y \in \mathcal{X}$ выполнено

$$d(x) \geq d(y) + \big\langle \nabla d(y), x - y \big\rangle + \frac{\mu}{2}\|x - y\|^2.$$

Теперь можем ввести один из основных инструментов этой лекции.

Определение Л9.2. Пусть дана дифференцируемая 1-сильно выпуклая относительно нормы $\|\cdot\|$ на множестве $\mathcal{X}$ функция d. Дивергенцией Брэгмана, порожденной функцией d на множестве $\mathcal{X}$, называется функция двух аргументов $V(x, y) : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ такая, что для любых $x, y \in \mathcal{X}$ выполняется

$$V(x, y) = d(x) - d(y) - \big\langle \nabla d(y), x - y \big\rangle.$$

Глядя на определение μ-сильной выпуклости становится понятно, откуда взялось выражение для $V(x, y)$.

Замечание Л9.2. Дивергенцию Брэгмана можно воспринимать как расстояние между векторами, но стоит обратить внимание, что она не является метрикой и даже не симметрична относительно перестановки переменных.

Рассмотрим несколько примеров $V(x, y)$, в том числе какими функциями $d(x)$ они порождаются.

Пример Л9.1. Квадрат евклидовой нормы $d : \mathbb{R}^d \to \mathbb{R}$:

$$d(x) = \frac{1}{2}\|x\|_2^2.$$

1-сильная выпуклость относительно $\|\cdot\|_2$ тривиальна. Подставляем в определение дивергенции Брэгмана (Определение Л9.2):

$$\begin{aligned} V(x,y) &= \frac{1}{2}\|x\|_2^2 - \frac{1}{2}\|y\|_2^2 - \langle y, x - y\rangle = \frac{1}{2}\Big(\|x\|_2^2 - \|y\|_2^2 - 2\langle y, x\rangle + 2\|y\|_2^2\Big) \\ &= \frac{1}{2}\Big(\|x\|_2^2 - 2\langle y, x\rangle + \|y\|_2^2\Big) = \frac{1}{2}\|x - y\|_2^2. \end{aligned}$$

Получили просто квадрат евклидова расстояния.

Пример Л9.2. Энтропия Шеннона $d : \triangle_{d-1} \to \mathbb{R}$:

$$d(x) = \sum_{i=1}^{d} x_i \log x_i,$$

где

$$\triangle_{d-1} = \left\{ x \in \mathbb{R}^d \;\middle|\; x_i > 0, \sum_{i=1}^{d} x_i = 1 \right\}.$$

Градиент d в точке y:

$$\big(\nabla d(y)\big)_i = 1 + \log y_i,$$

Подставляем в определение дивергенции Брэгмана (Определение Л9.2):

$$\begin{aligned} V(x,y) &= \sum_{i=1}^{d} x_i \log x_i - \sum_{i=1}^{d} y_i \log y_i - \sum_{i=1}^{d} \big(1 + \log y_i\big)(x_i - y_i) \\ &= \sum_{i=1}^{d} x_i \big(\log x_i - \log y_i\big) - \sum_{i=1}^{d} (x_i - y_i) \\ &= \sum_{i=1}^{d} x_i \log \frac{x_i}{y_i}. \end{aligned}$$

Слагаемое $\sum_{i=1}^{d}(x_i - y_i) = 0$, так как $x, y \in \triangle_{d-1}$ и $\sum_{i=1}^{d} x_i = \sum_{i=1}^{d} y_i = 1$. Получили дивергенцию Кульбака–Лейблера (KL-дивергенцию).
Здесь же можем показать 1-сильную выпуклость энтропии относительно $\|\cdot\|_1$. Согласно неравенству Пинскера получим:

$$V(x,y) = d(x) - d(y) - \big\langle \nabla d(y), x - y\big\rangle = \sum_{i=1}^{d} x_i \log \frac{x_i}{y_i} \geq \frac{1}{2}\|x - y\|_1^2.$$

Пример Л9.3. Энтропия фон Неймана $d : \mathbb{S}_1^d \to \mathbb{R}$:

$$d(X) = \operatorname{Tr}(X \log X),$$

где

$$\mathbb{S}_1^d = \left\{ X \in \mathbb{S}_{++}^d \;\middle|\; \operatorname{Tr}(X) = 1 \right\},$$

а $\log X$ — матричный логарифм, определяется как отображение, обратное матричной экспоненте e^X. Также для него верно разложение в ряд для матриц:

$$\log X = \sum_{k=1}^{\infty} (-1)^{k+1} \frac{(X-I)^k}{k}, \;\; \|I - X\|_2 < 1.$$

Покажем, почему $\|I - X\|_2 \leq 1$ выполняется для матриц $X \in \mathbb{S}_1^d$. Известно, что матрица X положительно определена. Значит, все ее собственные числа $\lambda_i(X), \;\; i \in \overline{1,d}$ больше нуля. При этом, известен след матрицы:

$$\operatorname{Tr} X = \sum_{i=1}^{d} \lambda_i(X) = 1.$$

Откуда следует, что каждое из собственных значений X лежит на интервале: $\lambda_i(X) \in (0,1), \;\; i \in \overline{1,d}$. Но тогда имеем, что $1 - \lambda_i(X) \in (0,1)$, а это спектр матрицы $I - X$. Вторая норма для симметричной матрицы $I - X$ выражается как:

$$\|I - X\|_2 = |\lambda|_{\max}(I - X) < 1.$$

Следовательно, можем применять формулу через ряд при поиске дифференциала:

$$\mathrm{d} \operatorname{Tr}(X \log X) = \operatorname{Tr}((\mathrm{d}X) \log X) + \operatorname{Tr}(X \, \mathrm{d}(\log X)).$$

С последним слагаемым поработаем отдельно:

$$\begin{aligned}
\operatorname{Tr}(X \, \mathrm{d}(\log X)) &= \operatorname{Tr}\left(X \, \mathrm{d} \left(\sum_{k=1}^{\infty} (-1)^{k+1} \frac{(X-I)^k}{k} \right) \right) \\
&= \operatorname{Tr}\left(X \sum_{k=1}^{\infty} (-1)^{k+1} \frac{\mathrm{d}\big((X-I)^k\big)}{k} \right) \\
&= \operatorname{Tr}\left(X \sum_{k=1}^{\infty} (-1)^{k+1} \frac{k(X-I)^{k-1} \mathrm{d}X}{k} \right) \\
&= \operatorname{Tr}\left(X \sum_{k=0}^{\infty} (I-X)^k \mathrm{d}X \right).
\end{aligned}$$

При переходе от $\mathrm{d}\big((X-I)^k\big)$ к $k(X-I)^{k-1}\mathrm{d}X$ существенно, что X и $X - I$ коммутируют, а под знаком Tr можно циклически переставлять матрицы. Далее можно воспользоваться рядом Неймана:

$$(I - T)^{-1} = \sum_{k=0}^{\infty} T^n, \;\; |\lambda|_{\max}(T) < 1.$$

Применяем его для матрицы $I - X$:

$$\begin{aligned}\operatorname{Tr}(X\,\mathrm{d}(\log X)) &= \operatorname{Tr}\Big(X\big(I-(I-X)\big)^{-1}\mathrm{d}X\Big)\\ &= \operatorname{Tr}(I\mathrm{d}X).\end{aligned}$$

Собираем изначальный дифференциал:

$$\begin{aligned}\mathrm{d}\operatorname{Tr}(X\log X) &= \operatorname{Tr}((\mathrm{d}X)\log X) + \operatorname{Tr}(I\mathrm{d}X)\\ &= \operatorname{Tr}((I+\log X)\mathrm{d}X).\end{aligned}$$

Следовательно, градиент $\nabla d(Y)$ равен:

$$\nabla d(Y) = I + (\log Y)^{\top}.$$

Теперь находим дивергенцию Брэгмана V, порождаемую d:

$$\begin{aligned}V(X,Y) &= \operatorname{Tr}(X\log X) - \operatorname{Tr}(Y\log Y) - \big\langle I + (\log Y)^{\top}, X - Y\big\rangle\\ &= \operatorname{Tr}(X\log X - Y\log Y - (X-Y) - X\log Y + Y\log Y)\\ &= \operatorname{Tr}(X\log X - X\log Y))\end{aligned}$$

Слагаемое $\operatorname{Tr}(X-Y) = 0$, поскольку $\operatorname{Tr}X = \operatorname{Tr}Y = 1$. Получили выражение для дивергенции фон Неймана.

Пример Л9.4. Логарифмический барьер $d : \mathbb{S}^d_{++} \to \mathbb{R}$:

$$d(X) = -\log(\det X).$$

Вычислим дифференциал:

$$\begin{aligned}\mathrm{d}(-\log(\det X)) &= -\frac{1}{\det X}\,\mathrm{d}(\det X)\\ &= -\frac{1}{\det X}\det X\big\langle X^{-\top}, \mathrm{d}X\big\rangle\\ &= -\big\langle X^{-\top}, \mathrm{d}X\big\rangle.\end{aligned}$$

Подставляем в определение дивергенции Брэгмана (Определение Л9.2):

$$\begin{aligned}V(X,Y) &= -\log(\det X) + \log(\det Y) - \big\langle -Y^{-\top}, X - Y\big\rangle\\ &= -\log(\det X/\det Y) + \operatorname{Tr}\big(XY^{-1} - I\big)\\ &= \operatorname{Tr}\big(XY^{-1} - I\big) - \log\det\big(XY^{-1}\big).\end{aligned}$$

Изучим подробнее свойства нового объекта.

Утверждение Л9.1. Дивергенция Брэгмана в общем случае обладает следующими свойствами:

1. Асимметричность: не всегда верно $V(x,y) = V(y,x)$.
2. Ограниченность снизу:

$$V(x,y) \geq \frac{1}{2}\|x-y\|^2 \;\forall x, y \in \mathcal{X}.$$

3. Неотрицательность: $V(x, y) \geq 0 \ \forall x, y \in \mathcal{X}$.

4. 1-сильная выпуклость по первому аргументу.

5. По второму аргументу может не быть выпуклости.

Доказательство. Кратко поясним каждый пункт.

1. В качестве примера асимметричной дивергенции Брегмана можем рассмотреть KL-дивергенцию (Пример Л9.2). Из записи видно отсутствие симметрии:

$$V(x, y) = \sum_{i=1}^{d} x_i \log \frac{x_i}{y_i}.$$

2. Ограниченность снизу получается из Определения Л9.2 и 1-сильной выпуклости d:

$$V(x, y) = d(x) - d(y) - \langle \nabla d(y), x - y \rangle \geq \frac{1}{2} \|x - y\|^2 \ \forall x, y \in \mathcal{X}.$$

3. Неотрицательность следует из предыдущего пункта. Также можем утверждать, что равенство нулю $V(x, y) = 0$ достигается тогда и только тогда, когда аргументы равны: $x = y$.

4. Если всмотреться в выражение $V(x, y)$ через d, то относительно x это сумма 1-сильно выпуклой d и линейной функции, то есть, 1-сильно выпуклая функция.

5. Если же смотреть относительно y, то перед 1-сильно выпуклой d стоит минус, а скалярное произведение обращается к градиенту $\nabla d(y)$. Из-за такого сложного вида в общем случае дивергенция Брэгмана не является выпуклой по y.

■

Утверждение Л9.2 (Равенство параллелограмма/теорема Пифагора). Для любых точек $x, y, z \in \mathcal{X}$ выполняется

$$V(z, x) + V(x, y) - V(z, y) = \langle \nabla d(y) - \nabla d(x), z - x \rangle.$$

Доказательство. Распишем :

$$\begin{aligned} V(z, x) + V(x, y) &= d(z) - d(x) - \langle \nabla d(x), z - x \rangle + d(x) - d(y) - \langle \nabla d(y), x - y \rangle \\ &= d(z) - d(y) - \langle \nabla d(y), z - y \rangle - \langle \nabla d(x) - \nabla d(y), z - x \rangle \\ &= V(z, y) - \langle \nabla d(x) - \nabla d(y), z - x \rangle. \end{aligned}$$

■

Теперь можем переходить к основной части лекции.

Л9.2 Зеркальный спуск

Решаем задачу оптимизации

$$\min_{x \in \mathcal{X}} f(x), \tag{Л9.1}$$

где множество $\mathcal{X}$ и функция f выпуклы.
Обратимся к градиентному спуску (Алгоритм Л3.1). Мы записывали его итерацию в готовом виде:

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

Но можно зайти и с точки зрения оптимизации. Рассмотрим задачу:

$$\operatorname*{argmin}_{x \in \mathbb{R}^d} \left\{ f(x^k) + \left\langle \nabla f(x^k), x - x^k \right\rangle + \frac{1}{2\gamma_k} \left\| x - x^k \right\|_2^2 \right\}.$$

В самом деле, Теоремы Л2.1 и Л2.2 дают нам необходимое и достаточное условия глобального минимума выпуклой функции:

$$\nabla_x \left(f(x^k) + \left\langle \nabla f(x^k), x - x^k \right\rangle + \frac{1}{2\gamma_k} \left\| x - x^k \right\|_2^2 \right) = \nabla f(x^k) + \frac{1}{\gamma_k}(x - x^k) = 0,$$

откуда и получается формула для итерации градиентного спуска.
В этом параграфе мы ввели понятие дивергенции Брэгмана, которая имеет смысл расстояния. Используя его, можем построить метод на решении модифицированной задачи оптимизации: заменим $\frac{1}{2}\left\| x - x^k \right\|_2^2$ на $V(x, x^k)$, множество $\mathbb{R}^d$ заменим на выпуклое $\mathcal{X} \subseteq \mathbb{R}^d$:

$$\operatorname*{argmin}_{x \in \mathcal{X}} \left\{ f(x^k) + \left\langle \nabla f(x^k), x - x^k \right\rangle + \frac{1}{\gamma_k} V(x, x^k) \right\}.$$

Можно упростить запись, избавившись от константы $f(x^k)$ и домножив целевую функцию на положительное γ_k. Получим итеративный метод:

$$x^{k+1} = \operatorname*{argmin}_{x \in \mathcal{X}} \left\{ \left\langle \gamma_k \nabla f(x^k), x - x^k \right\rangle + V(x, x^k) \right\}.$$

Это и есть формула итерации метода зеркального спуска.

Алгоритм Л9.1 Зеркальный спуск

Вход: стартовая точка $x^0 \in \mathcal{X}$, размеры шагов $\{\gamma_k\}_{k=0} > 0$, количество итераций K

1: **for** $k = 0, 1, \ldots, K-1$ **do**
2: $\quad x^{k+1} = \operatorname{argmin}_{x \in \mathcal{X}} \left\{ \left\langle \gamma_k \nabla f(x^k), x - x^k \right\rangle + V(x, x^k) \right\}$
3: **end for**

Выход: $\frac{1}{K} \sum_{k=1}^{K} x^k$

Рассмотрим, какие методы будут получаться при выборе различных дивергенций Брэгмана.

Пример Л9.5. Квадрат евклидовой нормы $V : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, $\mathcal{X} \subseteq \mathbb{R}^d$:

$$V(x, y) = \frac{1}{2}\left\| x - y \right\|_2^2, \ x, y \in \mathcal{X}.$$

Запишем итерацию зеркального спуска:

$$x^{k+1} = \operatorname*{argmin}_{x \in \mathcal{X}} \left\{ \langle \gamma_k \nabla f(x^k), x - x^k \rangle + \frac{1}{2} \|x - x^k\|_2^2 \right\}.$$

Чтобы посчитать argmin, добавим слагаемые не зависящие от x и выделим квадрат нормы:

$$\begin{aligned} x^{k+1} &= \operatorname*{argmin}_{x \in \mathcal{X}} \left\{ \frac{1}{2}\gamma^2 \|\nabla f(x^k)\|_2^2 + \langle \gamma \nabla f(x^k), x - x^k \rangle + \frac{1}{2} \|x - x^k\|_2^2 \right\} \\ &= \operatorname*{argmin}_{x \in \mathcal{X}} \|x - (x^k - \gamma \nabla f(x^k))\|_2^2 = \Pi_{\mathcal{X}}(x^k - \gamma \nabla f(x^k)). \end{aligned}$$

Получился обычный градиентный спуск с евклидовой проекцией.

Пример Л9.6. Дивергенция Кульбака–Лейблера $V : \triangle_{d-1} \times \triangle_{d-1} \to \mathbb{R}$:

$$V(x, y) = \sum_{i=1}^{d} x_i \log \frac{x_i}{y_i}, \ x, y \in \triangle_{d-1}.$$

Запишем итерацию зеркального спуска как задачу с ограничениями:

$$\begin{aligned} \min_{x \in \mathbb{R}^d} \ & \langle \gamma \nabla f(x^k), x \rangle + V(x, x^k) \\ \text{s.t.} \ & -x_i \le 0 \\ & \sum_{i=1}^{d} x_i - 1 = 0. \end{aligned}$$

Лагранжиан для такой задачи:

$$\begin{aligned} L(x, \lambda, \nu) &= \langle \gamma \nabla f(x^k), x \rangle + V(x, x^k) + \sum_{i=1}^{d} \lambda_i (-x_i) + \nu \left(\sum_{i=1}^{d} x_i - 1 \right) \\ &= \sum_{i=1}^{d} \left(\gamma [\nabla f(x^k)]_i + \log \frac{x_i}{x_i^k} - \lambda_i + \nu \right) x_i - \nu. \end{aligned}$$

Условие оптимальности по x_i:

$$\frac{\partial}{\partial x_i} L(x, \lambda, \nu) = 1 + \gamma [\nabla f(x^k)]_i + \log \frac{x_i}{x_i^k} - \lambda_i + \nu = 0.$$

Отсюда находим x^*, доставляющий минимум лагранжиана:

$$x_i^* = x_i^k \exp\left(-1 - \gamma [\nabla f(x^k)]_i + \lambda_i - \nu\right), \ i = \overline{1, d}.$$

Записываем минимум лагранжиана:

$$\inf_{x \in \mathbb{R}^d} L(x, \lambda, \nu) = L(x^*, \lambda, \nu) = \sum_{i=1}^{d} -x_i^k \exp\Big(-1 + \lambda_i - \gamma [\nabla f(x^k)]_i - \nu\Big) - \nu.$$

Переходим к двойственной задаче:

$$\max_{\lambda_i \geq 0, \nu \in \mathbb{R}} \left[\sum_{i=1}^{d} -x_i^k \exp\left(-1 + \lambda_i - \gamma[\nabla f(x^k)]_i - \nu\right) - \nu \right].$$

Видно, что увеличение λ_i увеличивает экспоненту, поэтому $\lambda_i^* = 0$. Это все что нужно было от двойственной задачи.
Условие Каруша–Куна–Такера с подстановкой $\lambda_i^* = 0$):

$$\nabla_x \left(\left\langle \gamma \nabla f(x^k), x \right\rangle + V(x, x^k) + \nu^* \left(\sum_{i=1}^{d} x_i - 1 \right) \right) = 0.$$

Откуда:

$$\log\left(\frac{x_i^*}{x_i^k}\right) + 1 + \gamma[\nabla f(x^k)]_i + \nu^* = 0.$$

Преобразуем и получаем:

$$x_i^* = x_i^k \exp\Big(-\gamma[\nabla f(x^k)]_i\Big) \cdot \exp(1 + \nu^*).$$

Чтобы записать выражение для ν^*, вспомним ограничения задачи $\sum_{i=1}^{d} x_i^* = 1$. Отсюда получаем нормировку и можем записать:

$$x_i^{k+1} = x_i^* = \frac{x_i^k \exp\big(-\gamma[\nabla f(x^k)]_i\big)}{\sum_{i=1}^{d} x_i^k \exp\big(-\gamma[\nabla f(x^k)]_i\big)}.$$

Это и есть итерации зеркального спуска для симплекса.

Замечание Л9.3. В случае симплекса и KL-дивергенции можно получить выигрыш в $\frac{d}{\log d}$ раз по сравнению с градиентным спуском с евклидовой проекцией.

Чтобы добавить интуиции и понимания, покажем, как соотносятся зеркальный спуск и идея шага в сопряженном пространстве.

Пример Л9.7. Рассмотрим метод при $\mathcal{X} = \mathbb{R}^d$ и некоторой d:

$$\begin{aligned} x^{k+1} &= \operatorname*{argmin}_{x \in \mathcal{X}} \left\{ \left\langle \gamma \nabla f(x^k), x - x^k \right\rangle + V(x, x^k) \right\} \\ &= \operatorname*{argmin}_{x \in \mathbb{R}^d} \left\{ \left\langle \gamma \nabla f(x^k), x - x^k \right\rangle + d(x) - d(x^k) - \left\langle \nabla d(x^k), x - x^k \right\rangle \right\}. \end{aligned}$$

Запишем условие оптимальности (Теорема Л2.2):

$$\gamma \nabla f(x^k) + \nabla d(x) - \nabla d(x^k) = 0.$$

Это и есть идея Немировского и Юдина, идея «зеркальности».

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

То есть, $\nabla d : E \to E^*$ и есть то преобразование φ, переносящее в зеркальное пространство.

Замечание Л9.4. С помощью такого инструмента, как дивергенция Брэгмана, мы научились переходить от задачи в «зеркальном» пространстве к простому поиску argmin,

который во многих случаях либо имеет аналитическое решение, либо может быть найден численно с хорошей точностью.

Прежде чем переходить к доказательству сходимости, обобщим некоторые свойства:

Определение Л9.3. Пусть дана непрерывно дифференцируемая на $\mathcal{X}$ функция $f : \mathcal{X} \to \mathbb{R}$. Будем говорить, что данная функция имеет *L-Липшицев градиент* (является *L-гладкой*) относительно нормы $\|\cdot\|$ на $\mathcal{X}$, если для любых $x, y \in \mathcal{X}$ выполнено

$$\big\|\nabla f(x) - \nabla f(y)\big\|_* \leq L\big\|x - y\big\|.$$

Теорема Л9.1. Пусть дана L-гладкая относительно нормы $\|\cdot\|$ функция $f : \mathcal{X} \to \mathbb{R}$. Тогда для любых $x, y \in \mathcal{X}$ выполнено

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

Доказательство. Используем (см. страницу 84 из [28]) формулу Ньютона-Лейбница для криволинейного интеграла второго рода по кривой, заданной вектор функцией $r(\tau)$:

$$\int\limits_a^b \big\langle \nabla f\big(r(\tau)\big), \mathrm{d}r(\tau)\big\rangle = f\big(r(b)\big) - f\big(r(a)\big).$$

В нашем случае выберем кривую следующим образом $r(\tau) = x + \tau(y - x)$, где $\tau \in [0, 1]$. Тогда

$$\begin{aligned} f\big(y\big) - f\big(x\big) &= \int\limits_0^1 \big\langle \nabla f\big(x + \tau(y - x)\big), y - x\big\rangle \,\mathrm{d}\tau \\ &= \big\langle \nabla f\big(x\big), y - x\big\rangle + \int\limits_0^1 \big\langle \nabla f\big(x + \tau(y - x)\big) - \nabla f\big(x\big), y - x\big\rangle \,\mathrm{d}\tau\,. \end{aligned}$$

Переместив скалярное произведение влево и взяв модуль от обоих частей, получим:

$$\begin{aligned} \big|f\big(y\big) - f\big(x\big) - \big\langle \nabla f\big(x\big), y - x\big\rangle\big| &= \left|\int\limits_0^1 \big\langle \nabla f\big(x + \tau(y - x)\big) - \nabla f(x), y - x\big\rangle \,\mathrm{d}\tau\right| \\ &\leq \int\limits_0^1 \big|\big\langle \nabla f\big(x + \tau(y - x)\big) - \nabla f\big(x\big), y - x\big\rangle\big| \,\mathrm{d}\tau\,. \end{aligned}$$

В последнем переходе мы использовали факт, что модуль суммы не превосходит сумму моделей слагаемых. Далее воспользуемся неравенством Гёльдера (0.2), а затем L-гладкостью относительно нормы $\|\cdot\|$ (Определение Л9.3):

$$\begin{aligned} \big|f\big(y\big) - f\big(x\big) - \big\langle \nabla f\big(x\big), y - x\big\rangle\big| &\leq \int\limits_0^1 \big\|\nabla f\big(x + \tau(y - x)\big) - \nabla f\big(x\big)\big\|_*\big\|y - x\big\| \,\mathrm{d}\tau \\ &\leq L\big\|y - x\big\|^2 \int\limits_0^1 \tau \,\mathrm{d}\tau = \frac{L}{2}\big\|x - y\big\|^2. \end{aligned}$$

■

Теперь мы готовы перейти к доказательству сходимости зеркального спуска.

Теорема Л9.2. Пусть задача оптимизации на выпуклом множестве $\mathcal{X}$ (Л9.1) с L-гладкой относительно нормы $\|\cdot\|$, выпуклой целевой функцией f решается с помощью зеркального спуска (Алгоритм Л9.1). Тогда при $\gamma_k = \gamma \leq \frac{1}{L}$ справедлива следующая оценка сходимости:

$$f\left(\frac{1}{K}\sum_{k=1}^{K} x^k\right) - f^* \leq \frac{V(x^*, x^0)}{\gamma K}.$$

Доказательство. Начнем с итерации зеркального спуска:

$$\begin{aligned} x^{k+1} &= \operatorname*{argmin}_{x\in\mathcal{X}} \left\{\left\langle \gamma\nabla f(x^k), x - x^k\right\rangle + V(x, x^k)\right\} \\ &= \operatorname*{argmin}_{x\in\mathcal{X}} \left\{\left\langle \gamma\nabla f(x^k), x - x^k\right\rangle + d(x) - d(x^k) - \left\langle \nabla d(x^k), x - x^k\right\rangle\right\}. \end{aligned}$$

Запишем условие оптимальности на выпуклом множестве $\mathcal{X}$ (Теорема **??**) для функции под argmin:

$$\left\langle \gamma\nabla f(x^k) + \nabla d(x^{k+1}) - \nabla d(x^k), x - x^{k+1}\right\rangle \geq 0$$

для всех $x \in \mathcal{X}$. Далее применяем равенство параллелограмма для дивергенции Брэгмана (Утверждение Л9.2):

$$V(x, x^{k+1}) + V(x^{k+1}, x^k) - V(x, x^k) = \left\langle \nabla d(x^k) - \nabla d(x^{k+1}), x - x^{k+1}\right\rangle.$$

Подставляем в неравенство выше, учитывая знак:

$$\left\langle \gamma\nabla f(x^k), x^{k+1} - x\right\rangle + V(x, x^{k+1}) + V(x^{k+1}, x^k) - V(x, x^k) \leq 0. \tag{Л9.2}$$

Теперь запишем свойство L-гладкости (Теорема Л9.1):

$$f(x^{k+1}) - f(x^k) - \left\langle \nabla f(x^k), x^{k+1} - x^k\right\rangle \leq \frac{L}{2}\left\|x^k - x^{k+1}\right\|^2. \tag{Л9.3}$$

Сложим неравенство (Л9.2) и домноженное на γ неравенство (Л9.3):

$$\begin{aligned} \left\langle \gamma\nabla f(x^k), x^{k+1} - x\right\rangle + V(x, x^{k+1}) + V(x^{k+1}, x^k) - V(x, x^k) \\ +\gamma\big(f(x^{k+1}) - f(x^k) - \left\langle \nabla f(x^k), x^{k+1} - x^k\right\rangle\big) \leq \frac{\gamma L}{2}\left\|x^k - x^{k+1}\right\|^2. \end{aligned}$$

Перепишем, оставив слева только слагаемые с f:

$$\begin{aligned} \left\langle \gamma\nabla f(x^k), x^k - x\right\rangle + \gamma\big(f(x^{k+1}) - f(x^k)\big) \leq V(x, x^k) - V(x, x^{k+1}) \\ - V(x^{k+1}, x^k) + \frac{\gamma L}{2}\left\|x^k - x^{k+1}\right\|^2. \end{aligned} \tag{Л9.4}$$

Из определения выпуклости (Определение Л2.3) можно оценить:

$$f(x^k) - f(x) + \left\langle \nabla f(x^k), x - x^k\right\rangle \leq 0. \tag{Л9.5}$$

Домножим неравенство (Л9.5) на γ и сложим с неравенством (Л9.4) (скалярное произведение $\left\langle \gamma\nabla f(x^k), x^k - x\right\rangle$ и $f(x^k)$ уничтожаются):

$$\gamma\big(f(x^{k+1}) - f(x)\big) \leq V(x, x^k) - V(x, x^{k+1}) - V(x^{k+1}, x^k) + \frac{\gamma L}{2}\left\|x^k - x^{k+1}\right\|^2. \tag{Л9.6}$$

Второе свойство дивергенции Брэгмана из Утверждения Л9.1:

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

Подставляем в неравенство (Л9.6):

$$\gamma\big(f(x^{k+1}) - f(x)\big) \leq V(x, x^k) - V(x, x^{k+1}) + \big(\gamma L - 1\big)V(x^{k+1}, x^k).$$

Применяем $\gamma \leq \frac{1}{L}$ и отбрасываем неположительное слагаемое:

$$\gamma\big(f(x^{k+1}) - f(x)\big) \leq V(x, x^k) - V(x, x^{k+1}).$$

Усредняем по всем k от 0 до $K-1$:

$$\frac{\gamma}{K}\sum_{k=0}^{K-1}\big(f(x^{k+1}) - f(x)\big) \leq \frac{1}{K}\sum_{k=0}^{K-1}\big(V(x, x^k) - V(x, x^{k+1})\big).$$

Далее справа сворачиваем телескопическую сумму:

$$\frac{1}{K}\sum_{k=0}^{K-1}\big(f(x^{k+1}) - f(x)\big) \leq \frac{1}{\gamma K}\big(V(x, x^0) - V(x, x^K)\big) \leq \frac{V(x, x^0)}{\gamma K}.$$

Для левой части применяем неравенство Йенсена (Неравенство (0.1))

$$f\left(\frac{1}{K}\sum_{k=1}^{K} x^k\right) - f(x) \leq \frac{V(x, x^0)}{\gamma K}.$$

В силу произвольности x, можем взять $x = x^*$:

$$f\left(\frac{1}{K}\sum_{k=1}^{K} x^k\right) - f^* \leq \frac{V(x^*, x^0)}{\gamma K}.$$

■

Сформулируем следствие о сходимости и оценке на количество итераций при оптимальном значении шага.

Утверждение Л9.3. Пусть выполнены условия Теоремы Л9.2 и выбрано значение шага $\gamma_k = \frac{1}{L}$. Тогда справедлива следующая оценка сходимости:

$$f\left(\frac{1}{K}\sum_{k=1}^{K} x^k\right) - f^* \leq \frac{LV(x^*, x^0)}{K}.$$

Более того, чтобы добиться точности ε по функции $f\big(\frac{1}{K}\sum_{k=1}^{K} x^k\big) - f^* \leq \varepsilon)$, необходимо

$$K = \mathcal{O}\left(\frac{LV(x^*, x^0)}{\varepsilon}\right) \text{ итераций.}$$

Доказательство. Выпишем оценку из Теоремы Л9.2 и подставим $\gamma = \frac{1}{L}$:

$$f\left(\frac{1}{K}\sum_{k=1}^{K} x^k\right) - f^* \leq \frac{V(x^*, x^0)}{\gamma K} = \frac{LV(x^*, x^0)}{K}.$$

Чтобы получить оценку на число итераций, отметим, что мы хотим гарантировать:

$$f\left(\frac{1}{K}\sum_{k=1}^{K} x^k\right) - f^* \leq \frac{LV(x^*, x^0)}{K} \leq \varepsilon.$$

Тогда

$$K \geq \frac{LV(x^*, x^0)}{\varepsilon}.$$

■

Полученный результат сходимости зеркального спуска по форме аналогичен оценке для градиентного спуска с проекцией. Это объясняется тем, что зеркальный спуск является обобщением градиентного спуска. Возникает вопрос: может ли зеркальный спуск давать лучшие оценки скорости сходимости? Да, это возможно при более благоприятных значениях константы гладкости L и дивергенции Брэгмана V.
Константа L определяется через условие гладкости функции:

$$\left\|\nabla f(x) - \nabla f(y)\right\|_q \leq L\left\|x - y\right\|_p,$$

где $p, q \in [1, +\infty]$ образуют гёльдеровскую пару, то есть, $\frac{1}{p} + \frac{1}{q} = 1$.
Выбор нормы в этом определении влияет на значение L. Если $1 \leq p \leq 2$, то выполняются соотношения

$$\left\|z\right\|_q \leq \left\|z\right\|_2 \leq \left\|z\right\|_p,$$

что влечёт $L \leq L_2$. В удачных случаях константа L может быть существенно меньше евклидовой константы гладкости L_2.
При $1 \leq p \leq 2$ для дивергенции Брэгмана справедлива нижняя оценка (Утверждение Л9.1):

$$V(x, y) \geq \frac{1}{2}\left\|x - y\right\|_p^2 \geq \frac{1}{2}\left\|x - y\right\|_2^2.$$

Здесь ситуация противоположная: дивергенция Брэгмана не меньше квадрата нормы, что может ухудшить оценку.
Следовательно, преимущество зеркального спуска проявляется тогда, когда уменьшение константы гладкости компенсирует рост дивергенции Брэгмана. Более точно, выигрыш достигается в случае

$$\frac{L_2}{L} \gg \sup_{x,y \in \mathcal{X}} \frac{2V(x, y)}{\left\|x - y\right\|_2^2}.$$

Список литературы

[1] Necdet Serhat Aybat, Alireza Fallah, Mert Gurbuzbalaban, and Asuman Ozdaglar. A universally optimal multistage accelerated stochastic gradient method, 2019.

[2] Francis Bach and Eric Moulines. Non-asymptotic analysis of stochastic approximation algorithms for machine learning. In *Advances in Neural Information Processing Systems*, volume 24, pages 451–459, December 2011.

[3] Aleksandr Beznosikov, Valentin Samokhin, and Alexander Gasnikov. Distributed saddle point problems: lower bounds, near-optimal and robust algorithms*. *Optimization Methods and Software*, page 1–18, March 2025.

[4] A. Cauchy. Méthode générale pour la résolution des systèmes d'équations simultanées. *Comptes rendus hebdomadaires des séances de l'Académie des sciences*, 55:536–538, 1847.

[5] Yudong Chen. Lecture 9–10: Accelerated gradient descent, 2023. UW-Madison CS/ISyE/Math/Stat 726 Spring 2023 Lecture Notes.

[6] Timothy Dozat. Incorporating nesterov momentum into adam. ICLR Workshop, 2016. OpenReview; NAdam method.

[7] John Duchi, Elad Hazan, and Yoram Singer. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research*, 12:2121–2159, 2011.

[8] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2015.

[9] Peter D Lax. *Linear algebra and its applications*. John Wiley & Sons, 2007.

[10] Laurent Lessard, Benjamin Recht, and Andrew Packard. Analysis and design of optimization algorithms via integral quadratic constraints. *SIAM Journal on Optimization*, 26(1):57–95, January 2016.

[11] Hongzhou Lin, Julien Mairal, and Zaid Harchaoui. A universal catalyst for first-order optimization, 2015.

[12] Dong C. Liu and Jorge Nocedal. On the limited memory bfgs method for large scale optimization. *Mathematical Programming, Series B*, 45(3):503–528, 1989.

[13] Jun Liu. A concise lyapunov analysis of nesterov's accelerated gradient method, 2025.

[14] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019.

[15] Y. Nesterov. A method for solving the convex programming problem with convergence rate o(1/k2), 1983.

[16] Yurii Nesterov and B. T. Polyak. Cubic regularization of newton method and its global performance. *Mathematical Programming*, 108(1):177–205, 2006.

[17] Isaac Newton. *De analysi per aequationes numero terminorum infinitas*. 1669. Unpublished manuscript; first printed edition in 1711.

[18] J. Nocedal and S. Wright. *Numerical Optimization*. Springer Series in Operations Research and Financial Engineering. Springer New York, 2006.

[19] B.T. Polyak. Some methods of speeding up the convergence of iteration methods. *USSR Computational Mathematics and Mathematical Physics*, 4(5):1–17, 1964.

[20] Anton Rodomanov and Yurii Nesterov. New results on superlinear convergence of classical quasi-newton methods. *Journal of Optimization Theory and Applications*, 188(3):744–769, 2021.

[21] Jack Sherman and Winifred J. Morrison. Adjustment of an inverse matrix corresponding to changes in the elements of a given column or a given row of the original matrix. *Annals of Mathematical Statistics*, 1949. Abstract.

[22] Sebastian U. Stich. Unified optimal analysis of the (stochastic) gradient method, 2019.

[23] Tijmen Tieleman and Geoffrey Hinton. Lecture 6.5 – rmsprop: Divide the gradient by a running average of its recent magnitude. COURSERA: Neural Networks for Machine Learning, 2012.

[24] Rachel Ward, Xiaoxia Wu, and Léon Bottou. Adagrad stepsizes: Sharp convergence over nonconvex landscapes. *Journal of Machine Learning Research*, 21(1):1–30, 2020.

[25] Дмитрий Беклемишев. *Курс аналитической геометрии и линейной алгебры.* Litres, 2016.

[26] Эрнест Винберг. *Курс алгебры.* Litres, 2015.

[27] Г. Е. Иванов. *Лекции по математическому анализу. Часть 1.* МФТИ, Москва, 2011.

[28] Г. Е. Иванов. *Лекции по математическому анализу. Часть 2.* МФТИ, Москва, 2016.

[29] Ю. Е. Нестеров. *Методы выпуклой оптимизации.* Изд-во МЦНМО, Москва, 2010.

[30] Альберт Николаевич Ширяев. *Вероятность – 1: Элементарная теория вероятностей. Математические основания. Предельные теоремы.* Московский центр непрерывного математического образования, Москва, 3-е, перераб. и доп. edition, 2004.