

Определения и формулировки

1. Положительно определённая матрица.

💡 Матрица $A \in \mathbb{S}$ называется положительно (отрицательно) определённой, если для $\forall x \neq 0 : x^T A x > (<) 0$. Обозначение: $A \prec 0$ ($A \succ 0$).
Аналогично определяется полуопределённость, только там неравенства нестрогие.

2. Евклидова норма вектора.



$$\|x\|_2 = \sqrt{\sum_{i=1}^n |x_i|^2}$$

Данная норма соответствует расстоянию в реальном мире. Иначе называется 2-норма (см. p -норма вектора)

3. Неравенство треугольника для нормы.



Норма должна удовлетворять следующим свойствам:

1. $\|\alpha x\| = |\alpha| \|x\|, \alpha \in \mathbb{R}$
2. $\|x\| = 0 \Rightarrow x = 0$
3. $\|x + y\| \leq \|x\| + \|y\|$ – неравенство треугольника

4. p -норма вектора.



$$\|x\|_p = \left(\sum_{i=0}^n |x_i|^p \right)^{\frac{1}{p}}$$

Важные частные случаи:

- Норма Чебышева: $\|x\|_\infty = \max_i |x_i|$
- Манхэттенское расстояние или $L1$ норма: $\|x\|_1 = \sum_{i=0}^n |x_i|$

5. Как выглядит единичный шар в p -норме на плоскости для $p = 1, 2, \infty$?

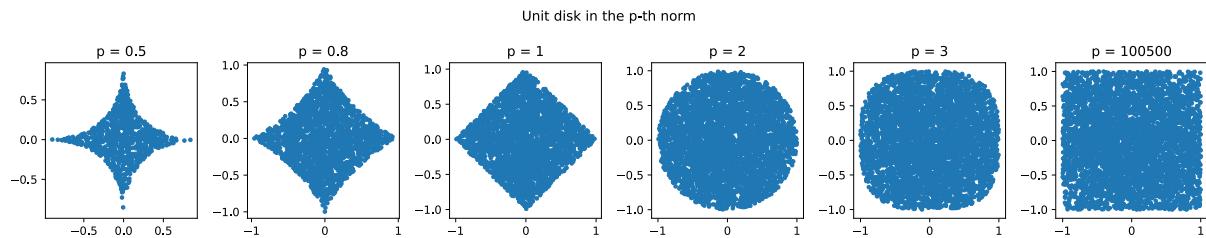


Рисунок 1: Шары в разных нормах

6. Норма Фробениуса для матрицы.



$$\|A\|_F = \left(\sum_{i=1}^m \sum_{j=1}^n |a_{ij}|^2 \right)^{\frac{1}{2}}$$

7. Спектральная норма матрицы.



$$\|A\|_2 = \sup_{x \neq 0} \frac{\|Ax\|_2}{\|x\|_2} = \sigma_1(A) = \sqrt{\lambda_{max}(A^\top A)}$$

Где $\sigma_1(A)$ – старшее сингулярное значение A , $\lambda_{max}(A^\top A)$ – наибольшее собственное значение $A^\top A$.

8. Скалярное произведение двух векторов.



Пусть $x, y \in \mathbb{R}^n$, тогда их скалярное произведение это

$$\langle x, y \rangle = x^T y = \sum_{i=1}^n x_i y_i = y^T x = \langle y, x \rangle$$

9. Скалярное произведение двух матриц, согласованное с нормой Фробениуса.



Пусть $X, Y \in \mathbb{R}^{m \times n}$, тогда их скалярное произведение это

$$\langle X, Y \rangle = \text{tr}(X^T Y) = \sum_{i=1}^m \sum_{j=1}^n X_{ij} Y_{ij} = \text{tr}(Y^T X) = \langle Y, X \rangle$$

Связь с нормой Фробениуса: $\langle X, X \rangle = \|X\|_F^2$

10. Собственные значения матрицы. Спектр матрицы.

💡 Скаляр λ является собственным значением для матрицы A , если существует вектор q , такой что $Aq = \lambda q$. В таком случае q называют собственным вектором.
Спектр матрицы – совокупность её собственных значений.

11. Связь спектра матрицы и её определенности.

💡 Матрица положительно (неотрицательно) определена \Leftrightarrow её спектр (все её собственные значения) положителен (неотрицателен).

12. Спектральное разложение матрицы.

💡 Спектральное разложение матрицы, или разложение матрицы на основе собственных векторов, — это представление квадратной матрицы A в виде произведения трёх матриц $A = V\Lambda V^{-1}$, где V — матрица, столбцы которой являются собственными векторами матрицы A , Λ — диагональная матрица с соответствующими собственными значениями на главной диагонали. В таком виде могут быть представлены только матрицы, обладающие полным набором собственных векторов.

Тогда $A^n = V\Lambda^n V^{-1}$.

13. Сингулярное разложение матрицы.

💡 $A \in \mathbb{R}^{m \times n}$, $\text{rank } A = r$.

$$A = U\Sigma V^T$$

$U \in \mathbb{R}^{m \times r}$, $U^T U = I$, $V \in \mathbb{R}^{n \times r}$, $V^T V = I$, Σ is a diagonal matrix with

$$\Sigma = \text{diag}(\sigma_1, \dots, \sigma_r)$$

such that

$$\sigma_1 \geq \sigma_2 \geq \dots \geq \sigma_r > 0$$

Столбцы U , V - левые и правые собственные векторы A , σ_i - сингулярные значения.

$$A = \sum_{i=1}^r \sigma_i u_i v_i^T$$

14. Связь определителя и собственных чисел для квадратной матрицы.

💡 Если у матрицы A собственные значения $\lambda_1, \lambda_2, \dots, \lambda_n$, то её определитель равен:

$$\det(A) = \lambda_1 \cdot \lambda_2 \cdot \dots \cdot \lambda_n$$

15. Связь следа и собственных чисел для квадратной матрицы.

💡 Если у матрицы A собственные значения $\lambda_1, \lambda_2, \dots, \lambda_n$, то её след равен:

$$\text{tr}(A) = \lambda_1 + \lambda_2 + \dots + \lambda_n$$

16. Линейная сходимость последовательности.

- 💡 Пусть есть последовательность $\{ |x_k - x^*|_2 \}$ в \mathbb{R} , сходящаяся к 0. Линейная сходимость при $q \in (0, 1)$ (скорость сходимости) и $C \in (0, \infty)$ (константа сходимости) определяется одним из двух способов:

$$\|x_{k+1} - x^*\| \leq Cq^k \text{ или } \|x_{k+1} - x^*\| \leq q\|x_k - x^*\|$$

Чем меньше q , тем быстрее сходится последовательность.

По-другому, говорят, что последовательность x_k сходится к числу L . Мы говорим, что эта последовательность линейно сходится к L , если \exists число $\mu \in (0, 1)$, такое, что

$$\lim_{k \rightarrow \infty} \frac{|x_{k+1} - L|}{|x_k - L|} = \mu$$

и μ называется скоростью сходимости.

17. Сублинейная сходимость последовательности.

- 💡 Если последовательность r_k сходится к нулю, но не обладает линейной сходимостью, то говорят, что она сходится сублинейно. Иногда мы можем рассматривать следующий класс сублинейной сходимости:

$$|x_{k+1} - x^*|_2 \leq Ck^q,$$

где $q < 0$ и $0 < C < \infty$.

18. Сверхлинейная сходимость последовательности.

- 💡 Мы определяем сверхлинейную сходимость как сходимость последовательности, которая быстрее любой линейной сходимости. Иногда рассматривают более специальный класс. Тогда говорят, что сверхлинейная сходимость при $q > 1, C > 0$ определяется следующим образом:

$$|x_{k+1} - x^*| \leq C|x_k - x^*|^q$$

19. Квадратичная сходимость последовательности.

- 💡 Квадратичная сходимость является частным случаем сверхлинейной сходимости, когда $q = 2$. Она определяется следующим образом:

$$|x_{k+1} - x^*| \leq C|x_k - x^*|^2$$

Или по-другому:

$$\lim_{k \rightarrow \infty} \frac{|x_{k+1} - L|}{|x_k - L|^2} = \mu$$

где $\mu > 0$.

20. Тест корней для определения скорости сходимости последовательности.

💡 Пусть $(r_k)_{k=m}^{\infty}$ - последовательность неотрицательных чисел, сходящаяся к нулю, и пусть $\alpha := \limsup_{k \rightarrow \infty} r_k^{1/k}$. (Заметим, что $\alpha \geq 0$.)

1. Если $0 \leq \alpha < 1$, то $(r_k)_{k=m}^{\infty}$ сходится линейно с константой α .
2. В частности, если $\alpha = 0$, то $(r_k)_{k=m}^{\infty}$ сходится сверхлинейно.
3. Если $\alpha = 1$, то $(r_k)_{k=m}^{\infty}$ сходится сублинейно.
4. Случай $\alpha > 1$ невозможен.

21. Тест отношений для определения скорости сходимости последовательности.

💡 Пусть $r_{k=k_m}^{\infty}$ - последовательность строго положительных чисел, сходящаяся к нулю. Пусть

$$q = \lim_{k \rightarrow \infty} \frac{r_{k+1}}{r_k}$$

1. Если существует q и $0 \leq q < 1$, то $r_{k=k_m}^{\infty}$ имеет линейную сходимость с константой q .
2. В частности, если $q = 0$, то $r_{k=k_m}^{\infty}$ имеет сверхлинейную сходимость.
3. Если q не существует, но $q = \lim_{k \rightarrow \infty} \sup_k \frac{r_{k+1}}{r_k} < 1$, то $r_{k=k_m}^{\infty}$ имеет линейную сходимость с константой, не превышающей q .
4. Если $\lim_{k \rightarrow \infty} \inf_k \frac{r_{k+1}}{r_k} = 1$, то $r_{k=k_m}^{\infty}$ имеет сублинейную сходимость.
5. Случай $\lim_{k \rightarrow \infty} \inf_k \frac{r_{k+1}}{r_k} > 1$ невозможен.

22. Унимодальная функция.

💡 Функция $f(x)$ называется унимодальной на $[a, b]$, если существует $x^* \in [a, b]$, такое, что

1. $f(x_1) > f(x_2)$ для всех $a \leq x_1 < x_2 < x^*$
2. $f(x_1) < f(x_2)$ для всех $x^* < x_1 < x_2 \leq b$

23. Метод дихотомии.

💡 Наша цель - решить следующую задачу: $\min_{x \in [a,b]} f(x)$ Мы делим отрезок на две равные части и выбираем ту, которая содержит решение задачи, используя значения функции, опираясь на ключевое свойство, описанное выше. Наша цель после одной итерации метода - уменьшить область поиска решения в два раза (в среднем). Метод описан на рисунках ниже.

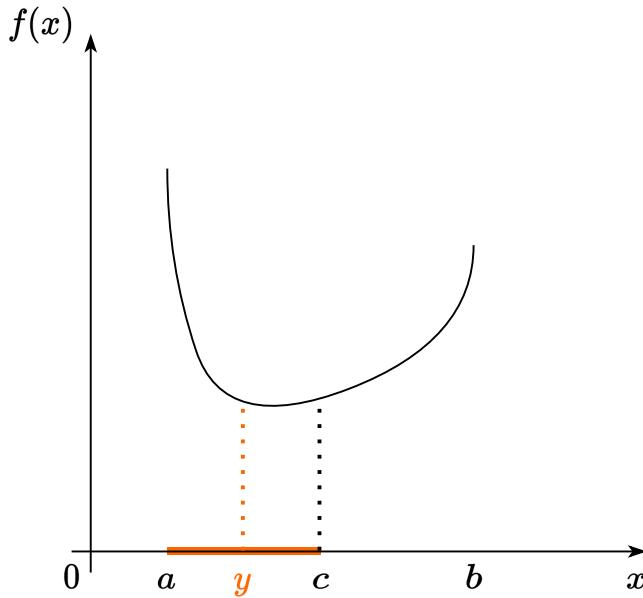


Рисунок 2: Диаграмма метода дихотомии

Длина отрезка на $(k + 1)$ -ой итерации:

$$\Delta_{k+1} = b_{k+1} - a_{k+1} = \frac{1}{2^k}(b - a)$$

Для унимодальных функций:

$$|x_{k+1} - x^*| \leq \frac{\Delta_{k+1}}{2} \leq \frac{1}{2^{k+1}}(b - a) \leq (0.5)^{k+1} \cdot (b - a)$$

Заметим, что на каждой итерации мы обращаемся к оракулу не более чем два раза, поэтому число вычислений функции равно $N = 2 \cdot k$, что подразумевает:

$$|x_{k+1} - x^*| \leq (0.5)^{\frac{N}{2}+1} \cdot (b - a) \leq (0.707)^N \frac{b - a}{2}$$

24. Метод золотого сечения.

💡 Общая идея: хотим поделить отрезок на 3 части так, чтобы потом когда одна из частей отпадет на следующей итерации одно из нужных значений функций будет уже известно.

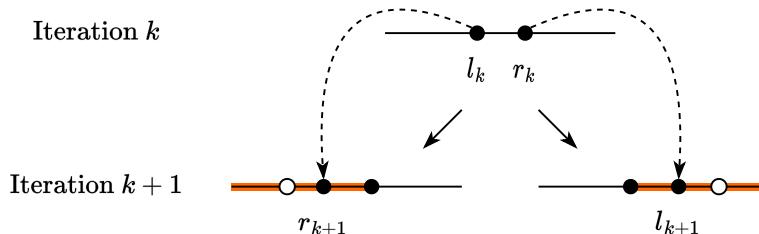


Рисунок 3: Иллюстрация метода золотого сечения

```
def golden_search(f, a, b, epsilon):
    tau = (sqrt(5) + 1) / 2
    y = a + (b - a) / tau**2
    z = a + (b - a) / tau
    while b - a > epsilon:
        if f(y) <= f(z):
            b = z
            z = y
            y = a + (b - a) / tau**2
        else:
            a = y
            y = z
            z = a + (b - a) / tau
    return (a + b) / 2
```

25. Метод параболической интерполяции.

💡 Идея метода: берем 3 точки, по этим 3 точкам однозначно строим параболу, находим ее минимум, и из этих 4 точек оставляем 3 так, чтобы между первой и третьей находился минимум.

```
def parabola_search(f, x1, x2, x3, epsilon):
    f1, f2, f3 = f(x1), f(x2), f(x3)
    while x3 - x1 > epsilon:
        u = x2 - ((x2 - x1)**2*(f2 - f3) - (x2 - x3)**2*(f2 - f1))/(2*((x2 - x1)*(f2 - f3))
        fu = f(u)

        if x2 <= u:
            if f2 <= fu:
                x1, x2, x3 = x1, x2, u
                f1, f2, f3 = f1, f2, fu
            else:
                x1, x2, x3 = x2, u, x3
                f1, f2, f3 = f2, fu, f3
        else:
            if fu <= f2:
                x1, x2, x3 = x1, u, x2
                f1, f2, f3 = f1, fu, f2
            else:
                x1, x2, x3 = u, x2, x3
                f1, f2, f3 = fu, f2, f3
    return (x1 + x3) / 2
```

Сходится сверхлинейно, но метод довольно неустойчивый. Если $f(x)$ не похожа на параболу, нам конец. Если она обратна параболе, то мы и вовсе уйдём искать максимум.

26. Условие достаточного убывания для неточного линейного поиска.

💡 Неточный линейный поиск:

$$x_{k+1} = x_k - \alpha \nabla f(x_k) \alpha = \operatorname{argmin}_{\alpha} f(x_{k+1})$$

Хотим приближенно найти α . Сведем задачу к поиску минимума следующей функции:

$$\phi(\alpha) = f(x_k) - \alpha \nabla f(x_k)^T \nabla f(x_k), \alpha \geq 0$$

Приблизим ее через первые 2 члена ряда Тейлора:

$$\phi(\alpha) \approx f(x_k) - \alpha \nabla f(x_k)^T \nabla f(x_k)$$

Тогда условием достаточного убывания (Armijo condition) является:

$$f(x_k - \alpha \nabla f(x_k)) \leq f(x_k) - c_1 \cdot \alpha \nabla f(x_k)^T \nabla f(x_k), c_1 \in (0, 1)$$

Иллюстрация для понимания:

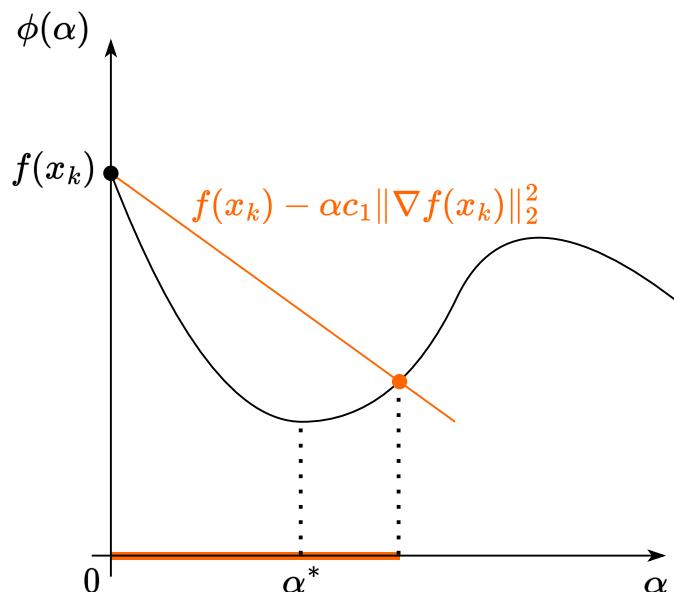


Рисунок 4: Иллюстрация условия достаточного убывания

27. Условия Гольдштейна для неточного линейного поиска.

💡 Определим ϕ_1 и ϕ_2 следующим образом ($c_1 > c_2$)

$$\phi_1(\alpha) = f(x_k) - c_1 \alpha \|\nabla f(x_k)\|^2$$

$$\phi_2(\alpha) = f(x_k) - c_2 \alpha \|\nabla f(x_k)\|^2$$

Тогда условие Гольдштейна заключается в том, что $\phi_1(\alpha) \leq \phi(\alpha) \leq \phi_2(\alpha)$.
Иллюстрация для понимания:

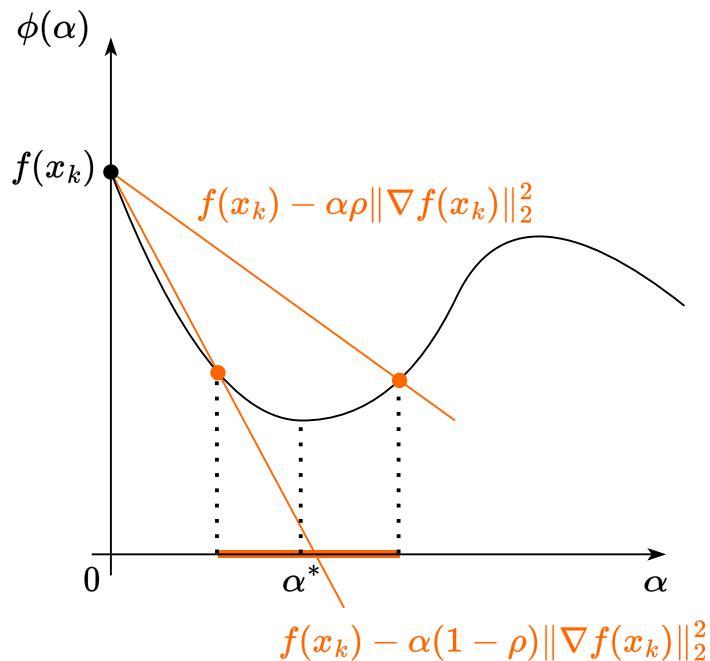


Рисунок 5: Иллюстрация условий Гольдштейна

28. Условие ограничения на кривизну для неточного линейного поиска.



$$-\nabla f(x_k - \alpha \nabla f(x_k))^\top \nabla f(x_k) \geq c_2 \nabla f(x_k)^\top (-\nabla f(x_k)),$$

где $c_2 \in (c_1, 1)$, и c_1 взято из условия достаточного убывания.

Иллюстрация для понимания:

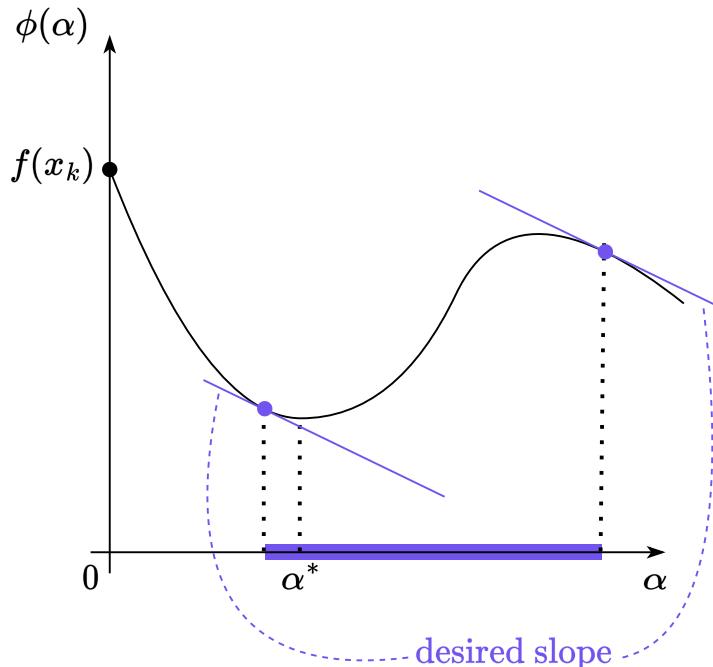


Рисунок 6: Иллюстрация условия ограничения на кривизну

29. Градиент функции $f(x) : \mathbb{R}^n \rightarrow \mathbb{R}$.

💡 $\nabla f(x)$, вектор частных производных функции f .

30. Гессиан функции $f(x) : \mathbb{R}^n \rightarrow \mathbb{R}$.



$$f''(x) = \nabla^2 f(x) = \frac{\partial^2 f}{\partial x_i \partial x_j} = \begin{pmatrix} \frac{\partial^2 f}{\partial x_1 \partial x_1} & \frac{\partial^2 f}{\partial x_1 \partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_1 \partial x_n} \\ \frac{\partial^2 f}{\partial x_2 \partial x_1} & \frac{\partial^2 f}{\partial x_2 \partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_2 \partial x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 f}{\partial x_n \partial x_1} & \frac{\partial^2 f}{\partial x_n \partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_n \partial x_n} \end{pmatrix}$$

31. Якобиан функции $f(x) : \mathbb{R}^n \rightarrow \mathbb{R}^m$.



$$J_f = f'(x) = \frac{df}{dx^T} = \begin{pmatrix} \frac{\partial f_1}{\partial x_1} & \frac{\partial f_2}{\partial x_1} & \dots & \frac{\partial f_m}{\partial x_1} \\ \frac{\partial f_1}{\partial x_2} & \frac{\partial f_2}{\partial x_2} & \dots & \frac{\partial f_m}{\partial x_2} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial f_1}{\partial x_n} & \frac{\partial f_2}{\partial x_n} & \dots & \frac{\partial f_m}{\partial x_n} \end{pmatrix}$$

32. Формула для аппроксимации Тейлора первого порядка $f_{x_0}^I(x)$ функции $f(x) : \mathbb{R}^n \rightarrow \mathbb{R}$ в точке x_0 .



Для дифференцируемой f :

$$f_{x_0}^I(x) = f(x_0) + \nabla f(x_0)^T(x - x_0)$$

33. Формула для аппроксимации Тейлора второго порядка $f_{x_0}^{II}(x)$ функции $f(x) : \mathbb{R}^n \rightarrow \mathbb{R}$ в точке x_0 .



Для дважды дифференцируемой f :

$$f_{x_0}^{II}(x) = f(x_0) + \nabla f(x_0)^T(x - x_0) + \frac{1}{2}(x - x_0)^T \nabla^2 f(x_0)(x - x_0)$$

34. Связь дифференциала функции df и градиента ∇f для функции $f(x) : \mathbb{R}^n \rightarrow \mathbb{R}$.



$$df(x) = \langle \nabla f(x), dx \rangle$$

35. Связь второго дифференциала функции $d^2 f$ и гессиана $\nabla^2 f$ для функции $f(x) : \mathbb{R}^n \rightarrow \mathbb{R}$.



$$d(df) = d^2 f(x) = \langle \nabla^2 f(x) dx_1, dx \rangle = \langle H_f(x) dx_1, dx \rangle$$

36. Формула для приближенного вычисления производной функции $f(x) : \mathbb{R}^n \rightarrow \mathbb{R}$ по k -ой координате с помощью метода конечных разностей.



$$\frac{\partial f}{\partial x_k}(x) \approx \frac{f(x + \varepsilon e_k) - f(x)}{\varepsilon}, \quad e_k = (0, \dots, \underset{k}{1}, \dots, 0)$$

Время работы: $2dT$, где вызов $f(x)$ занимает $T, x \in \mathbb{R}^d$

37. Пусть $f = f(x_1(t), \dots, x_n(t))$. Формула для вычисления $\frac{\partial f}{\partial t}$ через $\frac{\partial x_i}{\partial t}$ (Forward chain rule).



$$\frac{\partial f}{\partial t} = \sum_{i=1}^n \frac{\partial f}{\partial x_i} \frac{\partial x_i}{\partial t}$$

38. Пусть L - функция, возвращающая скаляр, а v_k - функция, возвращающая вектор $x \in \mathbb{R}^t$. Формула для вычисления $\frac{\partial L}{\partial v_k}$ через $\frac{\partial L}{\partial x_i}$ (Backward chain rule).



$$\frac{\partial L}{\partial v_k} = \sum_{i=1}^t \frac{\partial L}{\partial x_i} \frac{\partial x_i}{\partial v_k}$$

39. Идея Хатчинсона для оценки следа матрицы с помощью матвек операций.

💡 $X \in \mathbb{R}^{d \times d}$, $v \in \mathbb{R}^d$ - случайный вектор: $\mathbb{E}[vv^T] = I$, на каждой координате которого с одинаковой вероятностью стоит 1 или -1 .

$$\text{Tr}(X) = \mathbb{E}[v^T X v] = \frac{1}{V} \sum_{i=1}^V v_i^T X v_i.$$

40. Афинное множество. Афинная комбинация. Афинная оболочка.

💡 Множество A называется аффинным если для любых x_1, x_2 из A прямая, проходящая через x_1, x_2 , тоже лежит в A . То есть:

$$\forall \theta \in \mathbb{R}, \forall x_1, x_2 \in A : \theta x_1 + (1 - \theta) x_2 \in A$$

Пример аффинного множества: \mathbb{R}^n

Пусть $x_1, x_2, \dots, x_k \in S$. Тогда точка $\theta_1 x_1 + \theta_2 x_2 + \dots + \theta_k x_k$ называется аффинной комбинацией, если

$$\forall i \in \{1, \dots, k\} : \theta_i \in \mathbb{R}, \quad \sum_{i=1}^k \theta_i = 1$$

Аффинная оболочка – множество всех возможных аффинных комбинаций элементов множества.

$$\text{aff}(S) = \left\{ \sum_{i=1}^k \theta_i x_i \middle| k > 0, x_i \in S, \theta_i \in \mathbb{R}, \sum_{i=1}^k \theta_i = 1 \right\}$$

41. Выпуклое множество. Выпуклая комбинация. Выпуклая оболочка.

💡 Множество S называется выпуклым если для любых x_1, x_2 из S отрезок между x_1, x_2 тоже лежит в S . То есть:

$$\forall \theta \in [0, 1], \forall x_1, x_2 \in S : \theta x_1 + (1 - \theta) x_2 \in S$$

Пусть $x_1, x_2, \dots, x_k \in S$. Тогда точка $\theta_1 x_1 + \theta_2 x_2 + \dots + \theta_k x_k$ называется выпуклой комбинацией, если

$$\forall i \in \{1, \dots, k\} : \theta_i \geq 0, \quad \sum_{i=1}^k \theta_i = 1$$

Выпуклая оболочка – множество всех возможных выпуклых комбинаций элементов множества.

$$\text{conv}(S) = \left\{ \sum_{i=1}^k \theta_i x_i \middle| k > 0, x_i \in S, \theta_i \geq 0, \sum_{i=1}^k \theta_i = 1 \right\}$$

42. Конус. Выпуклый конус. Коническая комбинация. Коническая оболочка.

💡 Множество S называется конусом если для любого x из S луч, проходящий из 0 через x , тоже лежит в S . То есть:

$$\forall \theta \geq 0, \forall x \in S : \theta x \in S$$

Множество S называется выпуклым конусом если для любых $x_1, x_2 \in S$ их коническая комбинация тоже лежит в S . То есть:

$$\forall x_1, x_2 \in S, \theta_1, \theta_2 \geq 0 : \theta_1 x_1 + \theta_2 x_2 \in S$$

Пусть $x_1, x_2, \dots, x_k \in S$. Тогда точка $\theta_1 x_1 + \theta_2 x_2 + \dots + \theta_k x_k$ называется конической комбинацией, если

$$\forall i \in \{1, \dots, k\} : \theta_i \geq 0$$

Коническая оболочка - множество всех возможных конических комбинаций элементов множества.

$$\text{coni}(S) = \left\{ \sum_{i=1}^k \theta_i x_i \mid k > 0, x_i \in S, \theta_i \geq 0 \right\}$$

43. Внутренность множества.

💡 Внутренность множества - совокупность всех точек множества, содержащих вместе с собой в множестве некоторую окрестность вокруг себя.

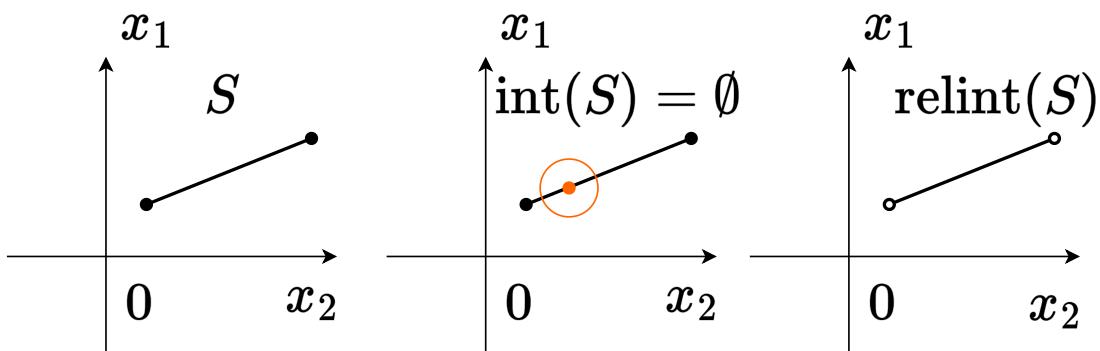
44. Относительная внутренность множества.

💡 Относительная внутренность множества - внутренность множества в его аффинной оболочке. Может быть полезной при работе с множествами меньшей размерности чем пространство, в котором они находятся.

$$\text{relint}(S) = \{x \in S \mid \exists \varepsilon > 0, N_\varepsilon(x) \cap \text{aff}(S) \subseteq S\}$$

$N_\varepsilon(x)$ – шар радиуса ε с центром в x , $\text{aff}(S)$ – аффинная оболочка S

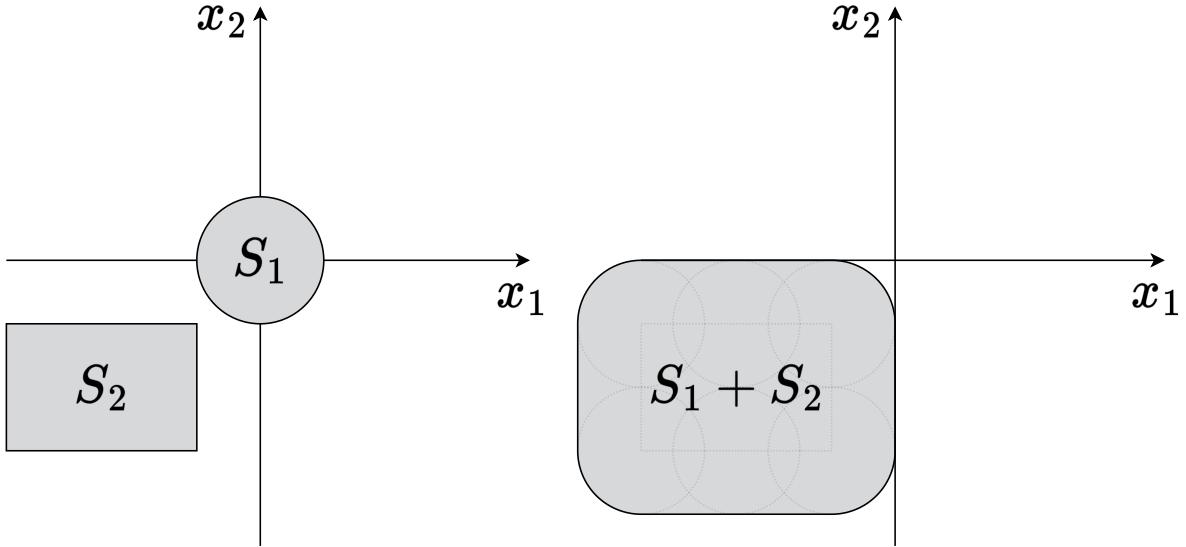
Пример: отрезок на плоскости имеет пустую внутренность, но его относительная внутренность – тот же отрезок без концов.



45. Сумма Минковского.

- 💡 Сумма Минковского – евклидово пространство, формирующееся сложением каждого вектора из S_1 с каждым вектором из S_2 :

$$S_1 + S_2 = \{s_1 + s_2 \mid s_1 \in S_1, s_2 \in S_2\}$$



46. Любые 2 операции с множествами, сохраняющие выпуклость.



1. Линейная комбинация:

$$S = \{s \mid s = c_1x + c_2y, x \in S_x, y \in S_y, c_1, c_2 \in \mathbb{R}\}$$

2. Пересечение любого числа выпуклых множеств

3. Образ множества в аффинном преобразовании:

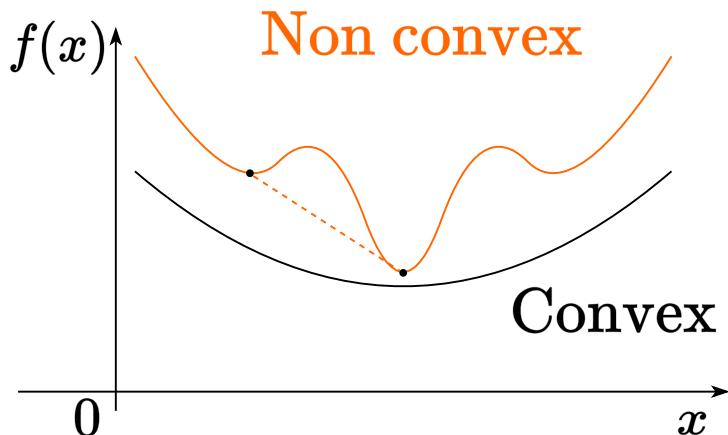
$$S \subseteq \mathbb{R}^n \text{ convex} \rightarrow f(S) = \{f(x) \mid x \in S\} \text{ convex} \quad (f(x) = Ax + b)$$

47. Выпуклая функция.

💡 Функция $f(x)$, определённая на выпуклом множестве $S \subseteq \mathbb{R}^n$ называется выпуклой на S если:

$$\forall x_1, x_2 \in S, \quad \forall \lambda \in [0, 1]$$

$$f(\lambda x_1 + (1 - \lambda)x_2) \leq \lambda f(x_1) + (1 - \lambda)f(x_2)$$



48. Строго выпуклая функция.

💡 Функция $f(x)$, определённая на выпуклом множестве $S \subseteq \mathbb{R}^n$ называется строго выпуклой на S если:

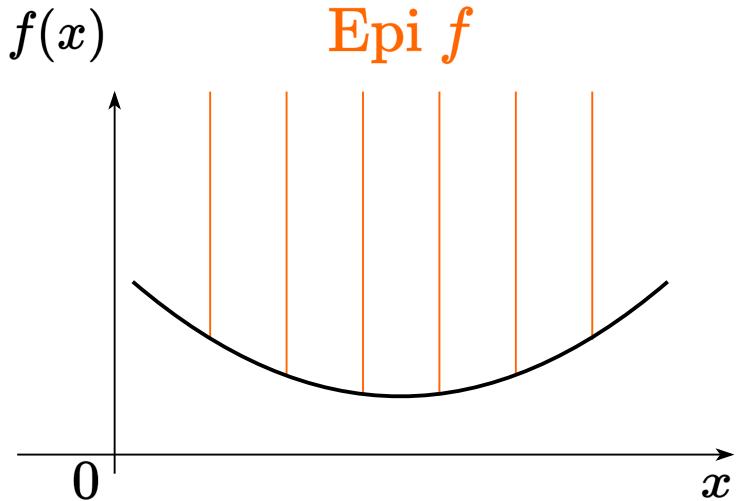
$$\forall x_1, x_2 \in S : x_1 \neq x_2, \quad \forall \lambda \in (0, 1)$$

$$f(\lambda x_1 + (1 - \lambda)x_2) < \lambda f(x_1) + (1 - \lambda)f(x_2)$$

49. Надграфик функции $f(x) : \mathbb{R}^n \rightarrow \mathbb{R}$.

💡 Для функции, определённой на $S \subseteq \mathbb{R}^n$, множество:

$$\text{epi } f = \{[x, \mu] \in S \times \mathbb{R} : f(x) \leq \mu\}$$



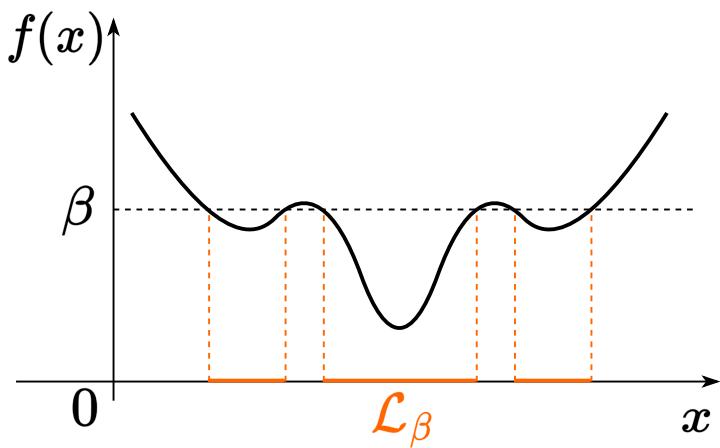
называется надграфиком функции $f(x)$

50. Множество подуровней функции $f(x) : \mathbb{R}^n \rightarrow \mathbb{R}$.

💡 Для функции, определённой на $S \subseteq \mathbb{R}^n$, множество:

$$\mathcal{L}_\beta = \{x \in S : f(x) \leq \beta\}$$

называется множеством подуровней или множеством Лебега функции $f(x)$
Если множество подуровней выпукло \iff функция выпукла.



51. Дифференциальный критерий выпуклости первого порядка.

! Дифференцируемая функция определенная на выпуклом множестве $S \subseteq \mathbb{R}^n$ выпукла тогда и только тогда когда $\forall x, y \in S$:

$$f(y) \geq f(x) + \nabla f^T(x)(y - x)$$

52. Дифференциальный критерий выпуклости второго порядка.

! Дважды дифференцируемая функция определенная на выпуклом множестве $S \subseteq \mathbb{R}^n$ выпукла тогда и только тогда когда для любой внутренней точки $x \in \text{int}(S) \neq \emptyset$:

$$\nabla^2 f(x) \succeq 0$$

53. Связь выпуклости функции и её надграфика.

! Функция выпукла тогда и только тогда, когда её надграфик - выпуклое множество.

54. μ -сильно выпуклая функция.

! Функция определенная на выпуклом множестве $S \subseteq \mathbb{R}^n$ называется сильно выпуклой если $\forall x_1, x_2 \in S, \$ 0 \leq \square \leq 1\$$ и $\mu > 0$:

$$f(\lambda x_1 + (1 - \lambda)x_2) \leq \lambda f(x_1) + (1 - \lambda)f(x_2) - \frac{\mu}{2}\lambda(1 - \lambda)\|x_1 - x_2\|^2$$

55. Дифференциальный критерий сильной выпуклости первого порядка.

! Дифференцируемая функция определенная на выпуклом множестве $S \subseteq \mathbb{R}^n$ является сильно выпуклой тогда и только тогда, когда $\forall x, y \in S$:

$$f(y) \geq f(x) + \nabla f^T(x)(y - x) + \frac{\mu}{2}\|y - x\|^2$$

56. Дифференциальный критерий сильной выпуклости второго порядка.

! Дважды дифференцируемая функция определенная на выпуклом множестве $S \subseteq \mathbb{R}^n$ является сильно выпуклой тогда и только тогда, когда существует $\mu > 0$

$$\nabla^2 f(x) \succeq \mu I$$

57. Любые 2 операции с функциями, сохраняющие выпуклость.

- !
- Сумма выпуклых функций с не отрицательными коэффициентами является выпуклой функцией.
 - Композиция выпуклой функции с афинной выпукла: $g(x) = f(Ax + b)$
 - Поточечный максимум любого числа выпуклых функций есть выпуклая функция.

58. Теорема Тейлора.

💡 $f : \mathbb{R}^n \rightarrow \mathbb{R}$ - непрерывная, дифференцируемая функция и $p \in \mathbb{R}^n$, тогда теорема Тейлора гласит:

$$f(x + p) = f(x) + \nabla f(x + tp)^T p$$

Для некоторого $t \in (0, 1)$ \

Более того, если f - дважды дифференцируема, то:

$$f(x + p) = f(x) + \nabla f(x)^T p + \frac{1}{2} p^T \nabla^2 f(x + tp) p$$

Для некоторого $t \in (0, 1)$

59. Необходимые условия локального экстремума.

💡 Если x^* - локальный экстремум и f непрерывная дифференцируема в открытой окрестности x^* , то:

$$\nabla f(x^*) = 0$$

60. Достаточные условия локального экстремума.

💡 Если $\nabla^2 f$ непрерывна в открытой окрестности x^* и

$$\nabla f(x^*) = 0$$

$$\nabla^2 f(x^*) \succ 0$$

То x^* - локальный минимум $f(x)$. Для локального максимума аналогично, только

$$0 \succ \nabla^2 f(x^*)$$

61. Общая задача математического программирования. Функция Лагранжа.

💡

$$\begin{cases} f_0(x) \rightarrow \min_{x \in \mathbb{R}^d} \\ f_i(x) \leq 0, \quad i = 1, \dots, m, \\ h_i(x) = 0, \quad i = 1, \dots, p. \end{cases}$$

Функция Лагранжа:

$$L(x, \lambda, \nu) = f_0(x) + \sum_{i=1}^m \lambda_i f_i(x) + \sum_{i=1}^p \nu_i h_i(x),$$

62. Теорема Каруша - Куна - Таккера в форме необходимых условий решения задачи математического программирования.

💡 Пусть x_* - решение задачи с нулевым зазором двойственности

$$\begin{cases} f_0(x) \rightarrow \min_{x \in \mathbb{R}^d} \\ f_i(x) \leq 0, \quad i = 1, \dots, m, \\ h_i(x) = 0, \quad i = 1, \dots, p. \end{cases}$$

Функция Лагранжа:

$$L(x, \lambda, \nu) = f_0(x) + \sum_{i=1}^m \lambda_i f_i(x) + \sum_{i=1}^p \nu_i h_i(x),$$

Тогда найдутся такие векторы λ^* и ν^* , что выполнены условия

$$\begin{cases} \nabla f_0(x_*) + \sum_{i=1}^m \lambda^* \nabla f_i(x_*) + \sum_{i=1}^p \nu^* \nabla h_i(x_*) = 0 \\ f_i(x_*) \leq 0, \quad i = 1, \dots, m \\ h_i(x_*) = 0, \quad i = 1, \dots, p \\ \lambda_i^* \geq 0, \quad i = 1, \dots, m \\ \lambda_i^* f_i(x_*) = 0, \quad i = 1, \dots, m \end{cases}$$

63. Условие Слейтера.



- Если задача выпуклая (т.е., говоря о задаче минимизации, оптимизируемая функция f_0 и ограничения вида неравенство f_i - выпуклые, ограничения вида равенства h_i - аффинные)
- И существует точка x такая, что $h(x) = 0$ и $f_i(x) < 0$ (ограничения вида равенства активные, а ограничения вида неравенства выполняются строго)

То тогда задача имеет нулевой зазор двойственности и условия ККТ становятся необходимыми и достаточными.

64. Задача выпуклого программирования.



Задача выпуклого программирования — это задача оптимизации, в которой целевая функция является выпуклой функцией и область допустимых решений выпукла. В форме ниже функции f_0, \dots, f_m - выпуклые, а функции h_i - аффинные.

$$\begin{cases} f_0(x) \rightarrow \min_{x \in \mathbb{R}^d} \\ f_i(x) \leq 0, \quad i = 1, \dots, m, \\ h_i(x) = 0, \quad i = 1, \dots, p. \end{cases}$$

65. Двойственная функция в задаче математического программирования.

Предположим, что $D = \bigcap_{i=0}^m \mathbf{dom} f_i \cap \bigcap_{i=0}^p \mathbf{dom} h_i$ непустое. Определим двойственную функцию $g : \mathbb{R}^m \times \mathbb{R}^p \rightarrow \mathbb{R}$ как минимум лагранжиана по x : для $\lambda \in \mathbb{R}^m, \nu \in \mathbb{R}^p$

$$g(\lambda, \nu) = \inf_{x \in D} L(x, \lambda, \nu) = \inf_{x \in D} f_0(x) + \sum_{i=1}^m \lambda_i f_i(x) + \sum_{i=1}^p \nu_i h_i(x)$$

Так как двойственная функция это поточечный инфинум семейства аффинных функций от (λ, ν) , она вогнутая, даже если изначальная задача не выпуклая.

66. Двойственная задача для задачи математического программирования.

Пусть p^* - оптимальное решение изначальной задачи. Пусть \hat{x} достижимая точка для изначальной задачи, т.е. $f_i(\hat{x}) \leq 0$ and $h_i(\hat{x}) = 0, \lambda \geq 0$. Тогда имеем:

$$L(\hat{x}, \lambda, \nu) = f_0(\hat{x}) + \underbrace{\lambda^T f(\hat{x})}_{\leq 0} + \underbrace{\nu^T h(\hat{x})}_{=0} \leq f_0(\hat{x})$$

Тогда

$$g(\lambda, \nu) = \inf_{x \in D} L(x, \lambda, \nu) \leq L(\hat{x}, \lambda, \nu) \leq f_0(\hat{x})$$

$$g(\lambda, \nu) \leq p^*$$

Двойственной задачей называется

$$g(\lambda, \nu) \rightarrow \max_{\lambda \in \mathbb{R}^m, \nu \in \mathbb{R}^p}$$

$$s.t. \lambda \geq 0$$

67. Сильная двойственность. Зазор двойственности.

Пусть p^* - решение прямой задачи, d^* - решение двойственной задачи. Зазором двойственности называется

$$p^* - d^* \geq 0$$

Сильная двойственность возникает, если зазор равен нулю

$$p^* = d^*$$

68. Локальный анализ чувствительности с помощью множителей Лагранжа.

💡 Перейдем к возмущенной версии задачи:

$$f_0(x) \rightarrow \min_x$$

$$f_i(x) \leq u_i, \quad i = 1, \dots, m$$

$$h_i(x) = v_i, \quad i = 1, \dots, p,$$

Обозначим $p^*(u, v)$ - оптимальное решение этой задачи. Если имеет место сильная двойственность, то выполнено:

$$p^*(u, v) \geq p^*(0, 0) - \lambda^{*T} u - \nu^{*T} v$$

Если множители Лагранжа λ_i^*, ν_i^* большие, то небольшое изменение ограничений приведет к существенному изменению оптимального решения. То есть соответствующие ограничения очень сильно влияют на задачу.\ Если множители Лагранжа маленькие, то соответствующие ограничения мало влияют на задачу.

$$\lambda_i^* = -\frac{\partial p^*(0, 0)}{\partial u_i} \quad \nu_i^* = -\frac{\partial p^*(0, 0)}{\partial v_i}$$

69. Задача линейного программирования. Задача линейного программирования в стандартной форме.

💡 Все задачи с линейным функционалом и линейными ограничениями считаются задачами линейного программирования. Стандартная форма:

$$\begin{aligned} & \min_{x \in \mathbb{R}^n} c^T x \\ & s.t. Ax = b \\ & x_i \geq 0, i = 1, \dots, n \end{aligned}$$

70. Возможные случаи двойственности в задаче линейного программирования.

💡 Двойственная задача:

$$\begin{aligned} & \max_{\nu \in \mathbb{R}^m} -b^T \nu \\ & s.t. -A^T \nu \leq c \end{aligned}$$

1. Если либо у прямой, либо у двойственной задачи есть конечное решение, то и у другой тоже, и целевые переменные равны.
2. Если либо прямая, либо двойственная задача неограничена, то вторая из них невыполнима.

71. Симплекс метод.

💡 Симплекс метод решает следующую задачу:

$$\min_{x \in \mathbb{R}^n} c^\top x$$

$$s.t. Ax \leq b$$

Шаги выполнения симплекс метода:

1. Поиск начальной базисной допустимой точки: Выберем начальную базисную (она является решением системы $A_B x = b_B$, где B - базис размера n пространства, а матрица A обычно имеет больше n ограничений) допустимую ($Ax_0 \leq b$) точку x_0 (искать ее будем через двухфазный симплексметод). Если такая точка не найдена, задача не имеет допустимого решения.

2. Проверка оптимальности:

- Разложение вектора c в данном базисе B с коэффициентами λ_B :

$$\lambda_B^\top A_B = c^\top \quad \text{или} \quad \lambda_B^\top = c^\top A_B^{-1}$$

- Если все компоненты λ_B неположительны, текущий базис является оптимальным. Иначе далее меняем вершину симплекса.

3. Определение переменной для удаления из базиса:

4. Вычисление шага вдоль выбранного направления d :

- Для всех $j \notin B$ считаем шаг:

$$\mu_j = \frac{b_j - a_j^\top x_B}{a_j^\top d}$$

- Новая вершина, которую добавим в базис:

$$t = \arg \min_j \{\mu_j \mid \mu_j > 0\}$$

5. Обновление базиса:

6. Повторение:

- Далее повторяем шаги 2-5 до достижения оптимального решения или установления, что задача не имеет допустимого решения.

72. Нахождение первоначальной угловой точки с помощью двухфазного симплекс метода.



1. Рассмотрим задачу (Phase 1):

$$\min_{\xi \in \mathbb{R}^m, y \in \mathbb{R}^n, z \in \mathbb{R}^n} \sum_{i=1}^m \xi_i$$

$$s.t. Ay - Az \leq b + \xi$$

$$y \geq 0, x \geq 0, \xi \geq 0$$

Для нее есть допустимая угловая точка $z = 0, y = 0, \xi_i = \max(0, -b_i)$. Начиная с нее, решим задачу симплекс методом и получим точку оптимума, в которой $\xi = 0$ и выполнены указанные ограничения.

2. Решение задачи Phase 1 является допустимым базисом задачи Phase 2:

$$\min_{y \in \mathbb{R}^n, z \in \mathbb{R}^n} c^\top (y - z)$$

$$s.t. Ay - Az \leq b$$

$$y \geq 0, x \geq 0$$

3. Заметим, что оно так же будет являться допустимым базисом и угловой точкой для исходной задачи:

$$\min_{x \in \mathbb{R}^n} c^\top x$$

$$s.t. Ax \leq b$$

4. Так и нашли первоначальную угловую точку для исходной задачи.

73. Сходимость симплекс метода.



В худшем случае симплекс метод сходится экспоненциально от размерности задачи, но на практике в среднем алгоритм работает сильно лучше (полиномиально). Задача, на которой симплекс метод работает экспоненциальное время, называется примером Klee Minty.

74. Показать, что направление антиградиента - направление наискорейшего локального убывания функции.



Пусть f дифференцируема, зададим искомое направление локального убывания - h - $\|h\| = 1$. Тогда её аппроксимация: $f(x + \alpha h) = f(x) + \alpha \langle \nabla f(x), h \rangle + o(\alpha)$

$$f(x + \alpha h) < f(x) \Rightarrow \alpha \langle \nabla f(x), h \rangle + o(\alpha) < 0.$$

При $\alpha \rightarrow +0$ получаем: $\alpha \langle \nabla f(x), h \rangle \leq 0$

$$\|\langle \nabla f(x), h \rangle\| \leq \|\nabla f(x)\| \|h\| \leq \|\nabla f(x)\|$$

$$\langle \nabla f(x), h \rangle \geq -\|\nabla f(x)\| \Rightarrow h = \frac{-\nabla f(x)}{\|\nabla f(x)\|}, \text{ ч.т.д.}$$

75. Дифференциальное уравнение градиентного потока.

76. Метод градиентного спуска.

💡 Решаем задачу минимизации

$$f(x) \rightarrow \min_{x \in \mathbb{R}^d}$$

Если f дифференцируема, то тогда для решения этой задачи можно использовать метод градиентного спуска:

$$x_{k+1} = x_k - \alpha \nabla f(x_k)$$

77. Наискорейший спуск.

💡 Решаем задачу минимизации

$$f(x) \rightarrow \min_{x \in \mathbb{R}^d}$$

Если f дифференцируема, то тогда для решения этой задачи можно использовать метод наискорейшего спуска:

$$x_{k+1} = x_k - \alpha_k \nabla f(x_k)$$

$$\alpha_k = \arg \min_{\alpha \in \mathbb{R}^+} f(x_k - \alpha \nabla f(x_k)),$$

т.е. выбираем наилучший шаг спуска на каждой итерации метода.

78. Липшицева парабола для гладкой функции.

💡 Если $f : \mathbb{R}^n \rightarrow \mathbb{R}$ - непрерывно дифференцируема и градиент Липшицев с константой L , то $\forall x, y \in \mathbb{R}^n$:

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

Если зафиксируем $x_0 \in \mathbb{R}^n$, то:

$$\varphi_1(x) = f(x_0) + \langle f(x_0), x - x_0 \rangle - \frac{L}{2} \|x - x_0\|^2$$

$$\varphi_2(x) = f(x_0) + \langle f(x_0), x - x_0 \rangle + \frac{L}{2} \|x - x_0\|^2$$

Это две параболы, и для них верно, что $\varphi_1(x) \leq f(x) \leq \varphi_2(x) \ \forall x$

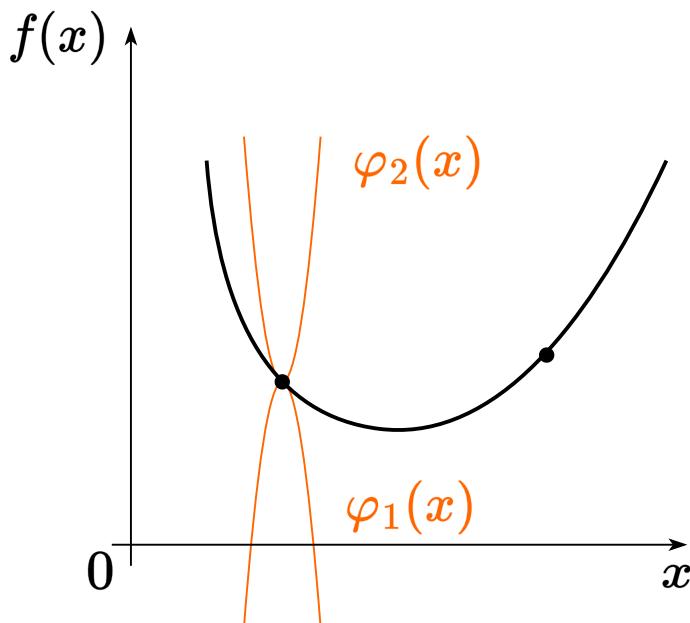


Рисунок 7: Иллюстрация Липшицевых парабол, между которыми зажата гладкая функция. Чаще нас интересует мажорирующая из них.

79. Размер шага наискорейшего спуска для квадратичной функции.

💡 Решаем задачу минимизации методом наискорейшего спуска

$$f(x) = \frac{1}{2}x^T Ax - b^T x + c \rightarrow \min_{x \in \mathbb{R}^d}$$

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

Из условия $\nabla f(x_{k+1})^T \nabla f(x_k) = 0$ получаем:

$$\alpha_k = \frac{2\nabla f(x_k)^T \nabla f(x_k)}{\nabla f(x_k)^T (A + A^T) \nabla f(x_k)} = \frac{\nabla f(x_k)^T \nabla f(x_k)}{\nabla f(x_k)^T \nabla^2 f(x_k) \nabla f(x_k)}.$$

80. Характер сходимости градиентного спуска к локальному экстремуму для гладких невыпуклых функций в терминах \mathcal{O} от числа итераций метода.

💡 $\|\nabla f(x_k)\|^2 \sim \mathcal{O}\left(\frac{1}{k}\right).$

81. Характер сходимости градиентного спуска для гладких выпуклых функций в терминах \mathcal{O} от числа итераций метода.

💡 $f(x_k) - f^* \sim \mathcal{O}\left(\frac{1}{k}\right).$

82. Характер сходимости градиентного спуска для гладких и сильно выпуклых функций в терминах \mathcal{O} от числа итераций метода.

💡 $\|x_k - x^*\|^2 \sim \mathcal{O}\left((1 - \frac{\mu}{L})^k\right).$

83. Связь спектра гессиана с константами сильной выпуклости и гладкости функции.

💡 $\mu = \min_{x \in \text{dom } f} \lambda_{\min}(\nabla^2 f(x)), \quad L = \max_{x \in \text{dom } f} \lambda_{\max}(\nabla^2 f(x)).$

84. Условие Поляка-Лоясиевича (градиентного доминирования) для функций.

💡 $\exists \mu > 0 : \quad \|\nabla f(x)\|^2 \geq 2\mu(f(x) - f^*) \quad \forall x, \text{ где } f^* - \text{минимум функции } f(x).$

85. Сходимость градиентного спуска для сильно выпуклых квадратичных функций. Оптимальные гиперпараметры.

💡 Решаем задачу минимизации методом градиентного спуска. Пусть $A \in \mathbb{S}_{++}^n \Rightarrow \nabla f = Ax - b$.

$$f(x) = \frac{1}{2}x^T Ax - b^T x + c \rightarrow \min_{x \in \mathbb{R}^d}$$

$$x_{k+1} = x_k - \alpha(Ax_k - b)$$

$$\alpha_{opt} = \frac{2}{\mu + L}, \text{ где } \mu = \lambda_{\min}(A), L = \lambda_{\max}(A)$$

$$\kappa = \frac{L}{\mu} \geqslant 1$$

$$\rho = \frac{\kappa - 1}{\kappa + 1}$$

$$\|x_k - x^*\| \leq \rho^k \|x_0 - x^*\|$$

86. Связь PL-функций и сильно выпуклых функций.

💡 Пусть f μ -сильно выпуклая и дифференцируемая $\Rightarrow f \in \text{PL}$.

Обратное неверно - $f(x) = x^2 + 3\sin^2 x \in \text{PL}$, но не сильно выпуклая (она вообще не выпуклая).

Function, that satisfies
Polyak-Lojasiewicz condition

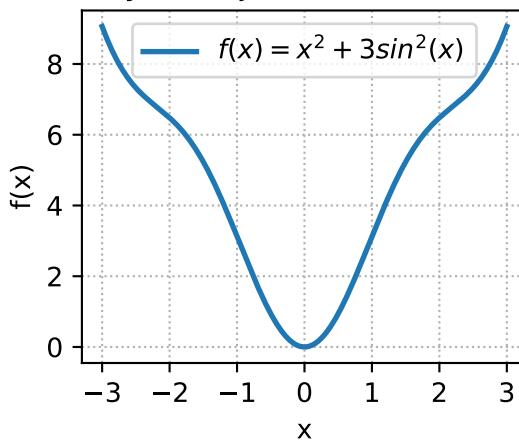


Рисунок 8: Пример невыпуклой PL функции

87. Привести пример выпуклой, но не сильно выпуклой задачи линейных наименьших квадратов (возможно, с регуляризацией).

💡 Рассмотрим задачу минимизации функции:

$$\|Ax - b\|^2 \rightarrow \min_{x \in \mathbb{R}^d},$$

где матрица $A \in \mathbb{R}^{m \times n}$, $b \in \mathbb{R}^m$, $m < n$ (лежачая).

88. Привести пример сильно выпуклой задачи линейных наименьших квадратов (возможно, с регу-

ляризацией).

💡 Рассмотрим задачу минимизации функции:

$$f(x) = \|Ax - b\|_2^2,$$

где $A \in \mathbb{R}^{n \times n}$ (ранг $A = n$). Эта функция сильно выпукла, так как гессиан положительно определен.

89. Привести пример выпуклой негладкой задачи линейных наименьших квадратов (возможно, с регуляризацией).

💡 Рассмотрим задачу минимизации функции:

$$f(x) = \|Ax - b\|_2^2 + \lambda \|x\|_1,$$

где $A \in \mathbb{R}^{n \times n}$, $\lambda > 0$. Эта функция выпукла, но негладка из-за наличия ℓ_1 -регуляризации.

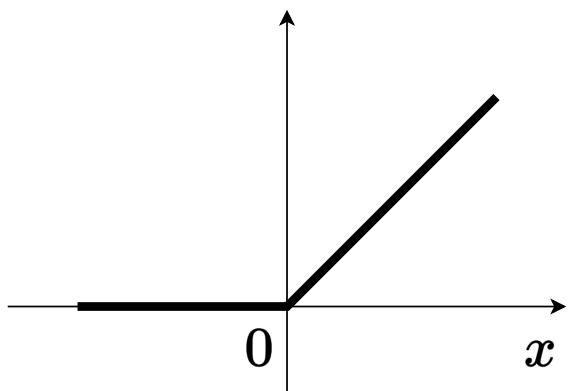
90. Субградиент. Субдифференциал.

💡 Субградиент функции f в точке x — это вектор g , удовлетворяющий условию:

$$f(y) \geq f(x) + g^T(y - x), \quad \forall y.$$

Множество всех субградиентов в точке x называется субдифференциалом и обозначается как $\partial f(x)$.

$$f(x) = \text{ReLU}(x)$$



$$\partial f(x)$$

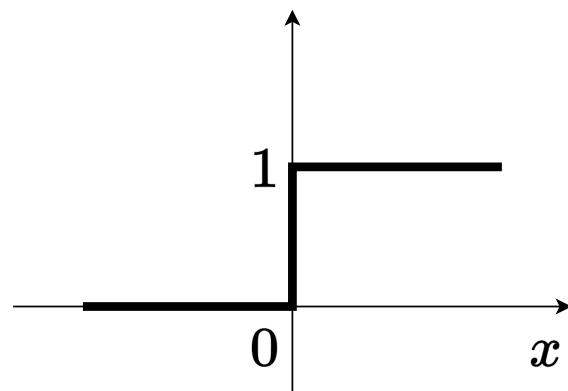


Рисунок 9: Субдифференциал функции ReLU.

91. Субградиентный метод.

💡 Субградиентный метод используется для минимизации выпуклых функций, которые могут быть негладкими. Итерационная формула метода:

$$x_{k+1} = x_k - \alpha_k g_k,$$

где $g_k \in \partial f(x_k)$ — субградиент функции f в точке x_k , α_k — шаг метода на k -й итерации.

92. Характер сходимости субградиентного метода для негладких выпуклых функций в терминах \mathcal{O} от числа итераций метода.

💡 Для негладких выпуклых функций субградиентный метод сходится со скоростью $\mathcal{O}\left(\frac{1}{\sqrt{k}}\right)$, где k — число итераций.

93. Нижние оценки для гладкой выпуклой оптимизации с помощью методов первого порядка в терминах \mathcal{O} от числа итераций метода.



Тип	Нижняя оценка на скорость сходимости
Гладкая и выпуклая	$f(x_k) - f^* \sim \mathcal{O}\left(\frac{1}{k^2}\right)$

94. Отличие ускоренной и неускоренной линейной сходимости для методов первого порядка.



Тип	Неускоренная	Ускоренная
Гладкая и сильно-выпуклая (или PL)	$\mathcal{O}\left((1 - \frac{\mu}{L})^k\right)$	$\mathcal{O}\left((1 - \sqrt{\frac{\mu}{L}})^k\right)$

95. Метод тяжелого шарика (Поляка).

💡 Задача: $f(x) \rightarrow \min_{x \in R^d}$, $f(x)$ - непрерывно дифференцируемая функция

$$x_{k+1} = x_k - \alpha \nabla f(x_k) + \beta(x_k - x_{k-1}), \quad 0 < \beta < 1.$$

96. Ускоренный градиентный метод Нестерова для выпуклых гладких функций.

- 💡 Рассматриваем задачу $f(x) \rightarrow \min_x$, где $f : \mathbb{R}^n \rightarrow \mathbb{R}$ выпуклая и L -гладкая. Алгоритм Нестерова ускоренного градиентного спуска (NAG) имеет вид ($x_0 = y_0$, $\lambda_0 = 0$):

Обновление градиента: $y_{k+1} = x_k - \frac{1}{L} \nabla f(x_k)$

Экстраполяция: $x_{k+1} = (1 - \gamma_k)y_{k+1} + \gamma_k y_k$

Экстраполяция веса: $\lambda_{k+1} = \frac{1 + \sqrt{1 + 4\lambda_k^2}}{2}$

Экстраполяция веса: $\gamma_k = \frac{1 - \lambda_k}{\lambda_{k+1}}$

Метод сходится со скоростью $\mathcal{O}\left(\frac{1}{k^2}\right)$, а именно:

$$f(y_k) - f^* \leq \frac{2L\|x_0 - x^*\|_2^2}{k^2}$$

97. Ускоренный градиентный метод Нестерова для сильно выпуклых гладких функций.

- 💡 Рассматриваем задачу $f(x) \rightarrow \min_x$, где $f : \mathbb{R}^n \rightarrow \mathbb{R}$ – μ -сильно выпуклая и L -гладкая. Алгоритм Нестерова ускоренного градиентного спуска (NAG) имеет вид ($x_0 = y_0$, $\lambda_0 = 0$):

Обновление градиента: $y_{k+1} = x_k - \frac{1}{L} \nabla f(x_k)$

Экстраполяция: $x_{k+1} = (1 - \gamma_k)y_{k+1} + \gamma_k y_k$

Экстраполяция весов: $\gamma_k = \frac{\sqrt{L} - \sqrt{\mu}}{\sqrt{L} + \sqrt{\mu}}$

Метод сходится линейно, а именно:

$$f(y_k) - f^* \leq \frac{\mu + L}{2} \|x_0 - x^*\|_2^2 \exp\left(-\frac{k}{\sqrt{\kappa}}\right), \quad \kappa = \frac{L}{\mu}$$

98. Проекция.

- 💡 Проекция точки $y \in \mathbb{R}^n$ на множество $S \subseteq \mathbb{R}^n$ это точка $\text{proj}_S(y) \in S$:

$$\text{proj}_S(y) = \arg \min_{x \in S} \frac{1}{2} \|x - y\|_2^2$$

99. Достаточное условие существования проекции точки на множество.

- 💡 Если $S \subseteq \mathbb{R}^n$ – замкнутое множество, тогда проекция на множество S существует для любой точки.

100. Достаточное условие единственности проекции точки на множество.

💡 Если $S \subseteq \mathbb{R}^n$ — замкнутое выпуклое множество, тогда проекция на множество S единственна для каждой точки.

101. Метод проекции градиента.

💡 Рассматривается задача $f(x) \rightarrow \min_{x \in S}$, где $S \subseteq \mathbb{R}^n$. Метод проекции градиента — это метод оптимизации с проекцией на бюджетное множество S :

$$x_{k+1} = \text{proj}_S(x_k - \alpha_k \nabla f(x_k)),$$

где α_k — learning rate.

102. Критерий проекции точки на выпуклое множество (Неравенство Бурбаки-Чейни-Гольдштейна).

💡 Проекция $\text{proj}_S(x)$ точки x на выпуклое множество S удовлетворяет:

$$\langle x - \text{proj}_S(x), y - \text{proj}_S(x) \rangle \leq 0 \quad \forall y \in S.$$

103. Проекция как нерастягивающий оператор.

💡 Проекция на выпуклое множество S является нерастягивающим оператором:

$$\|\text{proj}_S(x) - \text{proj}_S(y)\| \leq \|x - y\| \quad \forall x, y.$$

104. Метод Франк-Вульфа.

💡 Рассматриваем задачу $f(x) \rightarrow \min_{x \in S}$. Метод Франк-Вульфа имеет вид:

$$y_k = \arg \min_{x \in S} f_{x_k}^I(x) = \arg \min_{x \in S} \langle \nabla f(x_k), x \rangle$$
$$x_{k+1} = \gamma_k x_k + (1 - \gamma_k) y_k$$

где γ_k — гиперпараметр.

105. Характер сходимости метода проекции градиента для гладких выпуклых функций в терминах \mathcal{O} от числа итераций метода.

💡 Для гладких выпуклых функций метод проекции градиента имеет сходимость порядка $\mathcal{O}\left(\frac{1}{k}\right)$, где k — число итераций. То есть сходимость такая же, как и для безусловной задачи, но стоимость итерации может быть выше из-за проекции.

106. Характер сходимости метода проекции градиента для гладких сильно выпуклых функций в терминах \mathcal{O} от числа итераций метода.

💡 Для гладких сильно выпуклых функций метод проекции градиента имеет линейную сходимость порядка $\mathcal{O}\left((1 - \frac{\mu}{L})^k\right)$, где k — число итераций. То есть сходимость такая же, как и для безусловной задачи, но стоимость итерации может быть выше из-за проекции.

107. Характер сходимости метода Франк-Вульфа для гладких выпуклых функций в терминах \mathcal{O} от числа итераций метода.

💡 Метод Франк-Вульфа для гладких выпуклых функций имеет сходимость порядка $\mathcal{O}\left(\frac{1}{k}\right)$, где k — число итераций.

108. Характер сходимости метода Франк-Вульфа для гладких сильно выпуклых функций в терминах \mathcal{O} от числа итераций метода.

💡 Для гладких сильно выпуклых функций метод Франк-Вульфа имеет сходимость порядка $\mathcal{O}\left(\frac{1}{k}\right)$, где k — число итераций.

109. A -сопряженность двух векторов. A -ортогональность. Скалярное произведение $\langle \cdot, \cdot \rangle_A$.

💡 A -ортогональность (сопряженность):

$$x \perp_A y \iff x^T A y = 0.$$

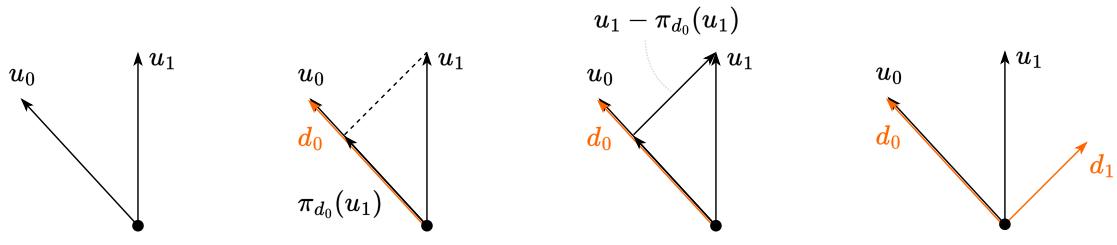
110. Процедура ортогонализации Грама-Шмидта.

💡 Пусть a_1, \dots, a_n - ЛНЗ векторы и $\text{proj}_b a$ - оператор проекции a на b , определенный как

$$\text{proj}_b a = \frac{\langle a, b \rangle}{\langle b, b \rangle} b,$$

Ортогонализация Грама-Шмидта:

$$\begin{aligned} b_1 &= a_1 \\ b_2 &= a_2 - \text{proj}_{b_1} a_2 \\ b_3 &= a_3 - \text{proj}_{b_1} a_3 - \text{proj}_{b_2} a_3 \\ &\dots \\ b_n &= a_n - \sum_{i=1}^{n-1} \text{proj}_{b_i} a_n \end{aligned}$$



111. Метод сопряженных направлений.

💡 Рассматриваем задачу

$$f(x) = \frac{1}{2}x^T Ax - b^T x + c \rightarrow \min_{x \in \mathbb{R}^d}$$

Идея

- В изотропном $A = I$ мире, наискорейший спуск стартующий из произвольной точки в любом пространстве натянутом на линейную оболочку из n ортогональных ЛН векторов будет сходится за n шагов в точной арифметике. Мы попытаемся в случае $A \neq I$ провести A -ортогонализацию, чтобы “наискорейшим” образом спускаться в измененном базисе.
- Предположим имеется набор из n линейно независимых A -ортогональных векторов(направлений) d_0, \dots, d_{n-1} (которые, например, были получены в ходе A -ортогонализации Г-Ш).
- Мы хотим создать метод, который переходит от x_0 к x^* по указанным ортогональным направлениям с некоторыми шагами, т.е. $x_0 - x^* = \sum_{i=0}^{d-1} \alpha_i d_i$, где α_i - из решения задачи линейного поиска.

Алгоритм

- $k = 0$ и $x_k = x_0$, $d_k = d_0 = -\nabla f(x_0)$.
- Пока $k < n$
 - Линейный поиск шага α : $f(x_k + \alpha_k d_k) \rightarrow \min_{\alpha} \Rightarrow$
$$\alpha_k = -\frac{d_k^\top (Ax_k - b)}{d_k^\top Ad_k}$$
 - Шаг алгоритма:
$$x_{k+1} = x_k + \alpha_k d_k$$
 - Обновление направления: $d_{k+1} = -\nabla f(x_{k+1}) + \beta_k d_k$, где β_k определяется из требований A - ортогональности d_{k+1} всем предыдущим направлениям.
 - $k := k + 1$

112. Метод сопряженных градиентов.

💡 Рассматриваем задачу

$$f(x) = \frac{1}{2}x^T Ax - b^T x + c \rightarrow \min_{x \in \mathbb{R}^d}$$

Метод сопряженных градиентов:

- $r_0 := b - Ax_0$
- if r_0 sufficiently small, then return x_0 as result
- $d_0 := r_0$
- $k := 0$
- while r_{k+1} is not sufficiently small :

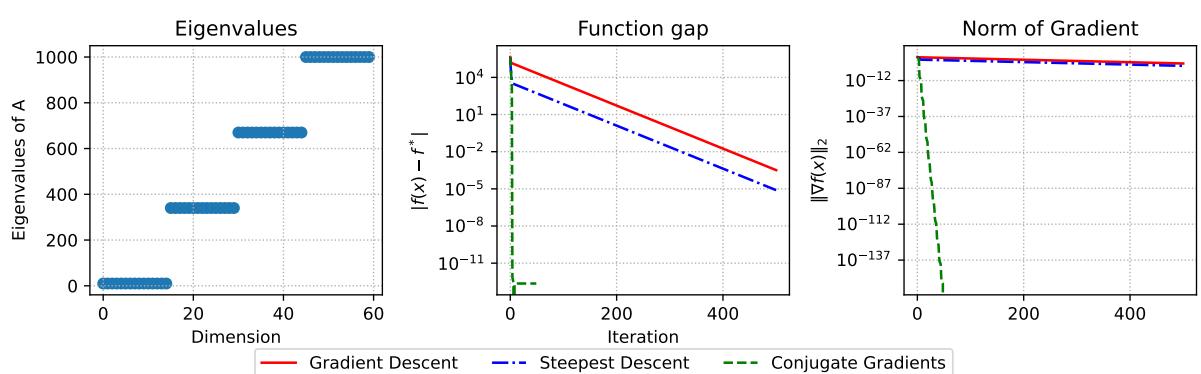
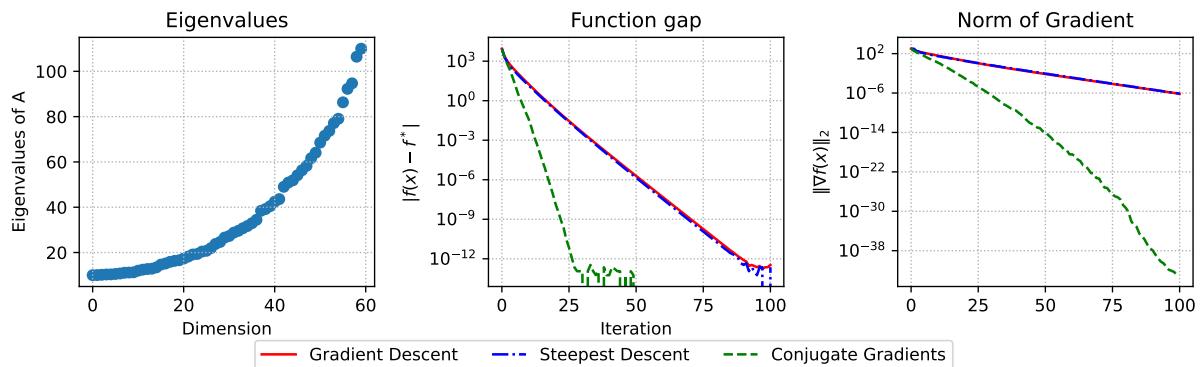
$$\begin{aligned} & - \alpha_k := \frac{r_k^T r_k}{d_k^T A d_k} \\ & - x_{k+1} := x_k + \alpha_k d_k \\ & - r_{k+1} := r_k - \alpha_k A d_k \\ & - \beta_k := \frac{r_{k+1}^T r_{k+1}}{r_k^T r_k} \\ & - d_{k+1} := r_{k+1} + \beta_k d_k \\ & - k := k + 1 \end{aligned}$$

- return x_{k+1} as result.

113. Зависимость сходимости метода сопряженных градиентов от спектра матрицы.

💡 Если матрица A имеет только r различных собственных чисел, тогда метод сопряжённых градиентов сходится за r итераций.

Strongly convex quadratics. n=60, random matrix.



114. Характер сходимости метода сопряженных градиентов в терминах \mathcal{O} от числа итераций метода.



$$\|x_k - x^*\|_A \leq 2 \left(\frac{\sqrt{\kappa(A)} - 1}{\sqrt{\kappa(A)} + 1} \right)^k \|x_0 - x^*\|_A$$

Имеет место оценка числа итераций при заданной точности ε : $\|x_k - x^*\|_A \leq \varepsilon \|x_0 - x^*\|_A$

$$k \leq \left\lceil \frac{1}{2} \sqrt{\kappa(A)} \ln\left(\frac{2}{\varepsilon}\right) \right\rceil$$

115. Метод Поляка-Рибьера.

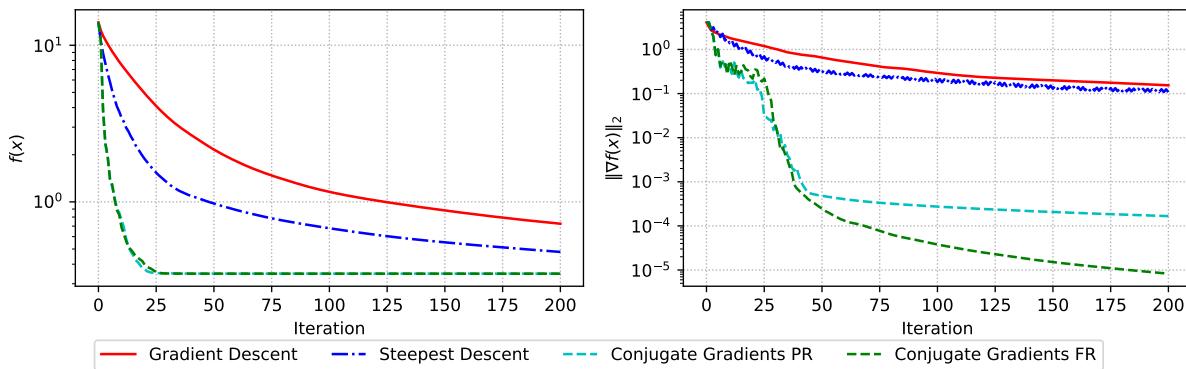


Используется для минимизации неквадратичных выпуклых функций.

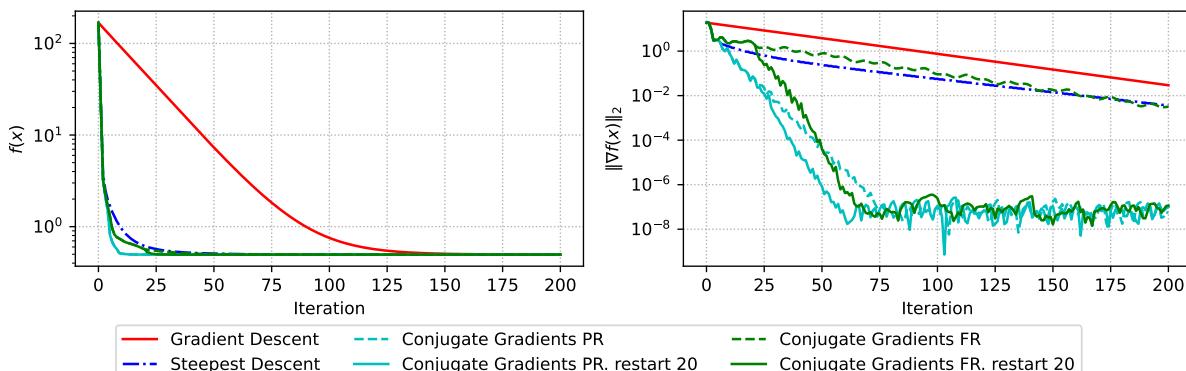
Без знания аналитического выражения шаг 2 алгоритма метода сопряжённых направлений вместо подсчёта α из минимизации $f(x_k + \alpha_k d_k)$ находит α обычным линейным поиском.

$$\beta_k = \frac{\nabla f(x_{k+1})^T (\nabla f(x_{k+1}) - \nabla f(x_k))}{d_k^T (\nabla f(x_{k+1}) - \nabla f(x_k))}$$

Regularized binary logistic regression. n=300. m=1000. $\mu=0$



Regularized binary logistic regression. n=300. m=1000. $\mu=1$



116. Метод Ньютона.

💡 Рассматривается задача минимизации функции с невырожденным гессианом.

$$x_{k+1} = x_k - [\nabla^2 f(x_k)]^{-1} \nabla f(x_k)$$

117. Сходимость метода Ньютона для квадратичной функции.

💡 Метод Ньютона сходится для квадратичной функции за одну итерацию. Следует из метода Ньютона квадратичной тейлоровской аппроксимации:

$$f(x) = f(x_k) + \nabla f(x_k)^T (x - x_k) + \frac{1}{2} (x - x_k)^T \nabla^2 f(x_k) (x - x_k), \quad \nabla f(x_{k+1}) = 0$$

118. Характер сходимости метода Ньютона для сильно выпуклых гладких функций - куда и как сходится.

💡 Пусть $f(x)$ сильно выпукла дважды непрерывно дифференцируемая на \mathbb{R}^n и выполняются неравенства: $\mu I_n \preceq \nabla^2 f(x) \preceq L I_n$. Тогда метод Ньютона с постоянным шагом локально сходится к решению со сверхлинейной скоростью. Если вдобавок, Гессиан M -Липшицев, тогда метод сходится локально к x^* с квадратичной скоростью.

119. Демпфированный метод Ньютона.

💡

$$x_{k+1} = x_k - \alpha_k [\nabla^2 f(x_k)]^{-1} \nabla f(x_k), \quad \alpha_k \in [0, 1]$$

где α_k находят с помощью линейного поиска. Сходимость глобальная.

120. Идея квазиньютоновских методов. Метод SR-1.

💡

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

Пусть $x_0 \in \mathbb{R}^n$, $B_0 \succ 0$. Для $k = 1, 2, 3, \dots$, повторим:

1. Решить $B_k d_k = -\nabla f(x_k)$ относительно d_k .
2. Обновить $x_{k+1} = x_k + \alpha_k d_k$.
3. Вычислить B_{k+1} из B_k .

$$B_{k+1} = B_k + \frac{(\Delta y_k - B_k d_k)(\Delta y_k - B_k d_k)^T}{(\Delta y_k - B_k d_k)^T d_k}, \quad \Delta y_k = \nabla f(x_{k+1}) - \nabla f(x_k).$$

121. Нижние оценки для негладкой выпуклой оптимизации с помощью методов первого порядка в терминах \mathcal{O} от числа итераций метода.

💡

Тип

Нижняя оценка на скорость сходимости

Негладкая и выпуклая

$$f(x_k) - f^* \sim \mathcal{O}\left(\frac{1}{\sqrt{k}}\right)$$

122. Проксимальный оператор.

💡 $\text{prox}_{\alpha f}(x_k) = \arg \min_{x \in \mathbb{R}^n} [\alpha f(x) + \frac{1}{2} \|x - x_k\|_2^2]$

123. Оператор проекции как частный случай проксимального оператора.



$$\text{proj}_S(y) := \arg \min_{x \in S} \frac{1}{2} \|x - y\|_2^2$$

Введём индикаторную функцию:

$$\mathbb{I}_S(x) = \begin{cases} 0, & \text{if } x \in S, \\ \infty, & \text{else.} \end{cases}$$

Перепишем оператор:

$$\text{proj}_S(y) = \arg \min_{x \in S} \left[\frac{1}{2} \|x - y\|_2^2 + \mathbb{I}_S(x) \right]$$

И, для сравнения, вспомним $\text{prox}_r(x_k) = \arg \min_{x \in \mathbb{R}^n} [\frac{1}{2} \|x - x_k\|_2^2 + r(x)]$.

124. Характер сходимости проксимального градиентного метода для гладких выпуклых функций f в терминах \mathcal{O} от числа итераций метода.

💡 Рассматривается задача: $\varphi(x) \rightarrow \min_{x \in \mathbb{R}^n}$, где $\varphi(x) = f(x) + r(x)$, $f(x)$ - гладкая выпуклая, $r(x)$ - негладкая выпуклая, проксимально дружественная.

$$x_{k+1} = \text{prox}_{r,\alpha}(x_k - \alpha \nabla f(x_k))$$

Сходится за $\mathcal{O}\left(\frac{1}{k}\right)$.

125. Характер сходимости проксимального градиентного метода для гладких сильно выпуклых функций f в терминах \mathcal{O} от числа итераций метода.

💡 Рассматривается задача: $\varphi(x) \rightarrow \min_{x \in \mathbb{R}^n}$, где $\varphi(x) = f(x) + r(x)$, $f(x)$ - гладкая выпуклая, $r(x)$ - негладкая выпуклая, проксимально дружественная.

$$\|x_k - x^*\|^2 \sim \mathcal{O}\left(\left(1 - \frac{\mu}{L}\right)^k\right)$$

где μ - константа сильной выпуклости функции f , L - константа гладкости функции f .

126. Аналитическое выражение для $\text{prox}_{\lambda \|x\|_1}$.



$$r(x) = \lambda \|x\|_1, \quad \lambda > 0$$
$$[\text{prox}_r(x)]_i = [|x_i - \lambda|_+ \cdot \text{sign}(x_i)]$$

127. Аналитическое выражение для $\text{prox}_{\frac{\mu}{2}\|x\|_2^2}$.



$$r(x) = \frac{\mu}{2} \|x\|_2^2$$
$$\text{prox}_r(x) = \frac{x}{1 - \mu}$$

128. Проксимальный оператор как нерастягивающий оператор.



Проксимальный оператор $\text{prox}_r(x)$ строго нерастягивающий (FNE - firmly non-expansive):

$$\|\text{prox}_r(x) - \text{prox}_r(y)\|_2^2 \leq \langle \text{prox}_r(x) - \text{prox}_r(y), x - y \rangle$$

и нерастягивающий:

$$\|\text{prox}_r(x) - \text{prox}_r(y)\|_2 \leq \|x - y\|_2$$

129. Характер сходимости ускоренного проксимального градиентного метода для гладких выпуклых функций f в терминах \mathcal{O} от числа итераций метода.



$$\varphi(x) = f(x) + r(x), \quad f(x) \text{ - выпуклая, } L\text{-гладкая, } r(x) \text{ - выпуклая и определен } \text{prox}_{\alpha r}(x_k) \Rightarrow$$
$$\varphi(x_k) - \varphi^* \leq \frac{L\|x_0 - x^*\|^2}{2k^2} \sim \mathcal{O}\left(\frac{1}{k^2}\right)$$

130. Метод стохастического градиентного спуска.



Решаемая задача: $f(x) \rightarrow \min_{x \in \mathbb{R}^p}$, где $f(x) = \frac{1}{n} \sum_{i=1}^n f_i(x)$

$$\text{SGD: } x_{k+1} = x_k - \alpha_k \nabla f_{i_k}(x),$$

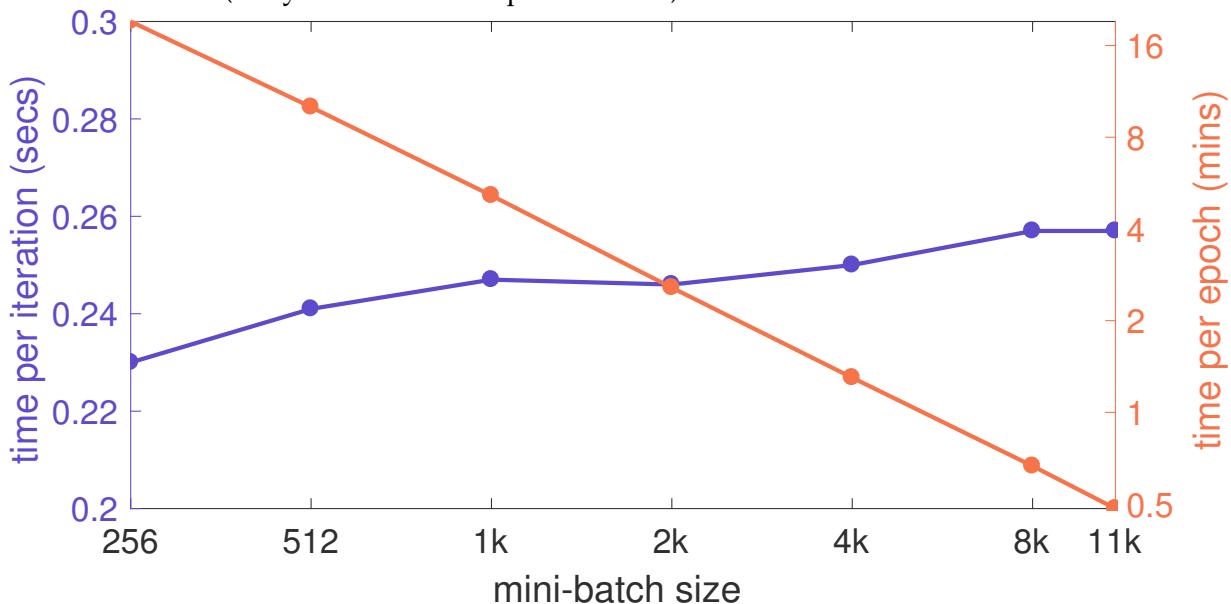
где i_k - случайно выбранный индекс. Если $\mathbb{P}(i_k = i) = \frac{1}{n}$, то $\mathbb{E}[\nabla f_{i_k}(x)] = \nabla f(x)$

131. Идея мини-батча для метода стохастического градиентного спуска. Эпоха.

💡 Разделим данные размера N на k мини-батчей (выборок) размера B_k , на каждой итерации посчитаем градиент мини-батча с использованием параллелизма. За $\frac{N}{k}$ итераций пройдёмся по всей выборке. Эпоха - набор k итераций с батчем размера $B_k = \frac{N}{k}$.

$$x_{k+1} = x_k - \frac{1}{|B_k|} \sum_{i \in B_k} -\text{шаг мини-батча}.$$

С увеличением размера мини-батча время на эпоху уменьшается до тех пор, пока нам хватает памяти (в случае наличия параллелизма).



132. Характер сходимости стохастического градиентного спуска для гладких выпуклых функций в терминах \mathcal{O} от числа итераций метода.

💡 f - гладкая и выпуклая $\Rightarrow \mathcal{O}\left(\frac{1}{\varepsilon^2}\right), \mathcal{O}\left(\frac{1}{\sqrt{k}}\right)$

133. Характер сходимости стохастического градиентного спуска для гладких PL-функций в терминах \mathcal{O} от числа итераций метода.

💡 $f \in \text{PL} \Rightarrow \mathcal{O}\left(\frac{1}{k}\right), \mathcal{O}\left(\frac{1}{\varepsilon}\right)$

134. Характер работы стохастического градиентного спуска с постоянным шагом для гладких PL-функций.

💡 Пусть $\min_{x \in \mathbb{R}^p} f(x) = \min_{x \in \mathbb{R}^p} \frac{1}{n} \sum_{i=1}^n f_i(x)$ при использовании стохастического градиентного спуска с постоянным шагом α

$$x_{k+1} = x_k - \alpha \nabla f_{i_k}(x_k)$$

имеем следующую оценку $\mathbb{E}[f(x_{k+1}) - f^*] \leq (1 - 2\alpha\mu)^k [f(x_0) - f^*] + \frac{L\sigma^2\alpha}{4\mu}$. Характер сходимости - линейный до некоторого шара несходимости, в котором будут происходить осцилляции и сходимости не будет.

135. Основная идея методов уменьшения дисперсии.

💡 Рассматриваем случаную величину X . Хотим уменьшить у неё дисперсию. Пусть Y - тоже случайная величина с известным мат. ожиданием. Рассмотрим новую с.в $Z_\alpha = \alpha(X - Y) + \mathbb{E}[Y]$

- $\mathbb{E}[Z_\alpha] = \alpha\mathbb{E}[X] + (1 - \alpha)\mathbb{E}[Y]$
- $\text{var}(Z_\alpha) = \alpha^2 (\text{var}(X) + \text{var}(Y) - 2\text{cov}(X, Y))$
 - $\alpha = 1$: нет смещения мат.ожидания
 - $\alpha < 1$: потенциальное смещение (но уменьшение дисперсии).
- Полезно, если Y коррелирует с X .

136. Метод SVRG.

- 💡
- Пусть $X = \nabla f_{i_k}(x_{m-1})$ - стох. градиент, а $Y = \nabla f_{i_k}(\tilde{x})$, с $\alpha = 1$ и \tilde{x} хранятся в памяти.
 - $\mathbb{E}[Y] = \frac{1}{n} \sum_{i=1}^n \nabla f_i(\tilde{x})$ полный градиент в \tilde{x} ;
 - $X - Y = \nabla f_{i_k}(x^{(m-1)}) - \nabla f_{i_k}(\tilde{x})$

Получаем алгоритм:

- **Initialize:** $\tilde{x} \in \mathbb{R}^d$
- **For** $i_{epoch} = 1$ **to** # of epochs
- Compute all gradients $\nabla f_i(\tilde{x})$; store $\nabla f(\tilde{x}) = \frac{1}{n} \sum_{i=1}^n \nabla f_i(\tilde{x})$
- Initialize $x_0 = \tilde{x}$
- **For** $t = 1$ **to** length of epochs (m)

$$x_t = x_{t-1} - \alpha \left[\nabla f(\tilde{x}) + \left(\nabla f_{i_t}(x_{t-1}) - \nabla f_{i_t}(\tilde{x}) \right) \right]$$

- Update $\tilde{x} = x_t$

137. Метод SAG.

Задача: $f(x) = \frac{1}{n} \sum_{i=1}^n f_i(x)$

1. Initialize $x^{(0)}$ and $g_i^{(0)} = \nabla f_i(x^{(0)})$
2. At steps $k = 1, 2, 3, \dots$ pick random $i_k \in \{1, \dots, n\}$
3. $g_{i_k}^{(k)} = \nabla f_{i_k}(x^{(k-1)})$
4. Set all other $g_i^{(k)} = g_i^{(k-1)}, i \neq i_k$
5. Update: $g^{(k)} = g^{(k-1)} + \frac{1}{n} (g_{i_k}^{(k)} - g_{i_k}^{(k-1)}) = \frac{1}{n} \sum_{i=1}^n g_i^{(k)}$
6. $x^{(k)} = x^{(k-1)} - \alpha^k g^{(k)}$

PS: стоимость итерации как в обычном SGD, но платим за это памятью.

Сходимость в выпуклом случае: $f(x_{\text{mean}}^{(k)}) - f^* \leq \frac{48n|f(x^{(0)}) - f^*| + 128L\|x^{(0)} - x^*\|^2}{k} = O(\frac{1}{k})$

Скорость в сильновыпуклом случае с параметром μ :

$$\mathbb{E}[f(x^{(k)}) - f^*] \leq (1 - \min(\frac{\mu}{16L}, \frac{1}{8n}))^k (\frac{3}{2}(f(x^{(0)}) - f^*) + \frac{4L}{n}\|x^{(0)} - x^*\|^2) = O(\gamma^k)$$

138. Метод Adagrad.

Задача: $f(x) = \frac{1}{n} \sum_{i=1}^n f_i(x)$. Пусть $g^{(k)} = \nabla f_{i_k}(x^{(k-1)})$, и обновляем for $j = 1, \dots, p$:

$$v_j^{(k)} = v_j^{k-1} + (g_j^{(k)})^2$$

$$x_j^{(k)} = x_j^{(k-1)} - \alpha \frac{g_j^{(k)}}{\sqrt{v_j^{(k)}} + \varepsilon}$$

Постоянная ε обычно устанавливается равным 10^{-6} чтобы гарантировать, что мы не будем иметь проблемы от деления на ноль или чрезмерно больших размеров шага.

139. Метод RMSProp.

Задача: $f(x) = \frac{1}{n} \sum_{i=1}^n f_i(x)$

Усовершенствование AdaGrad, учитывающее его агрессивную, монотонно снижающуюся скорость обучения. Использует скользящее среднее квадратов градиентов для корректировки скорости обучения для каждого веса. Пусть $g^{(k)} = \nabla f_{i_k}(x^{(k-1)})$ and update rule for $j = 1, \dots, p$:

$$v_j^{(k)} = \gamma v_j^{(k-1)} + (1 - \gamma)(g_j^{(k)})^2$$

$$x_j^{(k)} = x_j^{(k-1)} - \alpha \frac{g_j^{(k)}}{\sqrt{v_j^{(k)}} + \varepsilon}$$

140. Метод Adadelta.

Задача: $f(x) = \frac{1}{n} \sum_{i=1}^n f_i(x)$

Расширение RMSProp, направленное на снижение его зависимости от глобальной скорости обучения, устанавливаемой вручную. Вместо накопления всех прошлых квадратов градиентов, Adadelta ограничивает окно накопленных прошлых градиентов некоторым фиксированным размером w . Механизм обновления не требует скорости обучения α :

$$\begin{aligned} v_j^{(k)} &= \gamma v_j^{(k-1)} + (1 - \gamma) (g_j^{(k)})^2 \\ \tilde{g}_j^{(k)} &= \frac{\sqrt{\Delta x_j^{(k-1)} + \varepsilon}}{\sqrt{v_j^{(k)} + \varepsilon}} g_j^{(k)} \\ x_j^{(k)} &= x_j^{(k-1)} - \tilde{g}_j^{(k)} \\ \Delta x_j^{(k)} &= \rho \Delta x_j^{(k-1)} + (1 - \rho) (\tilde{g}_j^{(k)})^2 \end{aligned}$$

141. Метод Adam.

Задача: $f(x) = \frac{1}{n} \sum_{i=1}^n f_i(x)$

$$\begin{aligned} m_j^{(k)} &= \beta_1 m_j^{(k-1)} + (1 - \beta_1) g_j^{(k)} \\ v_j^{(k)} &= \beta_2 v_j^{(k-1)} + (1 - \beta_2) (g_j^{(k)})^2 \\ \tilde{m}_j &= \frac{m_j^{(k)}}{1 - \beta_1^k}, \quad \hat{v}_j = \frac{v_j^{(k)}}{1 - \beta_2^k} \\ x_j^{(k)} &= x_j^{(k-1)} - \alpha \frac{\tilde{m}_j}{\sqrt{\hat{v}_j} + \varepsilon} \end{aligned}$$

142. Идея проекции функции потерь нейронной сети на прямую, плоскость.

Пусть $L(w)$ - функция от $w \in \mathbb{R}^n$. Введем проекцию на линию:

$$L(\alpha) = L(w_0 + \alpha w_1)$$

для некоторого $w_1 \in \mathbb{R}^n$. Аналогично можно ввести проекцию на плоскость

$$L(\alpha, \beta) = L(w_0 + \alpha w_1 + \beta w_2)$$

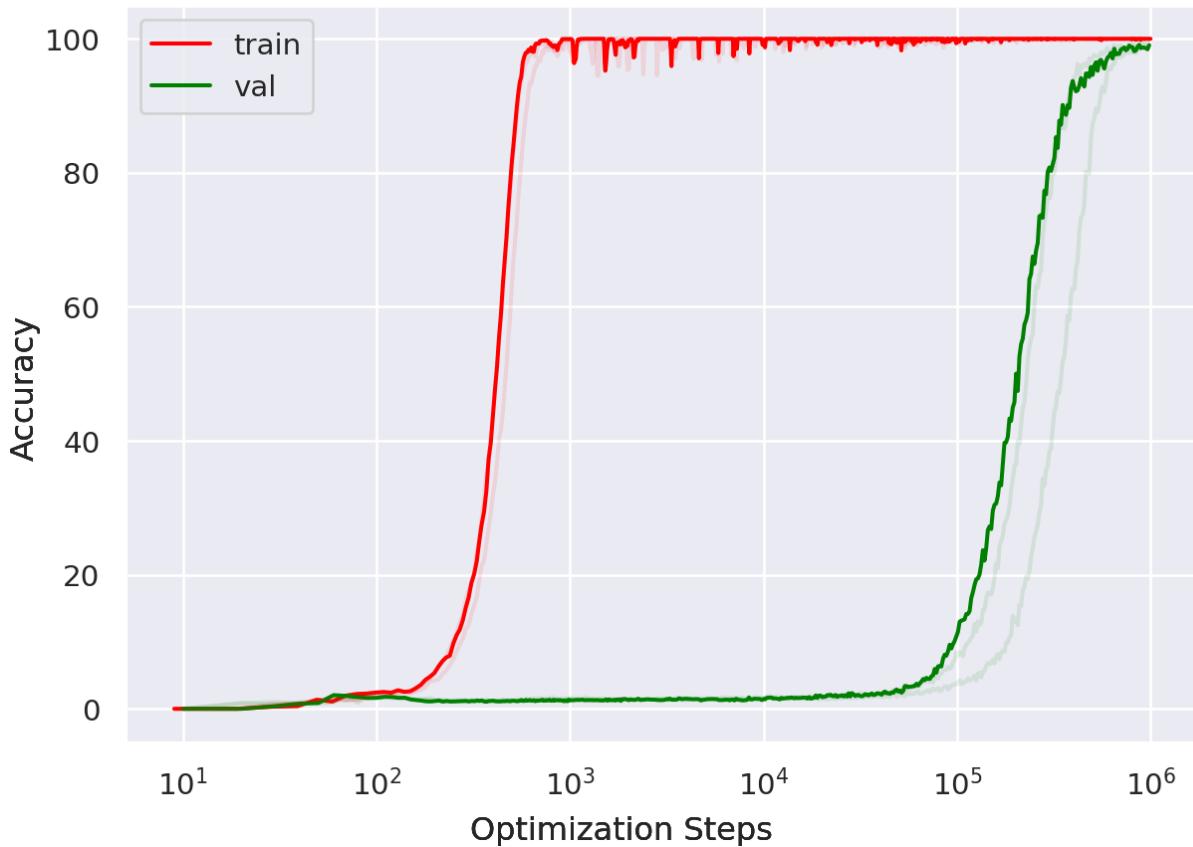
для некоторых $w_1, w_2 \in \mathbb{R}^n$.

- Два случайных вектора большой размерности с высокой вероятностью ортогональны друг другу.
- Если проекция функции невыпукла, то и исходная функция невыпукла. Таким образом можно заглянуть на устройство функции от многих переменных.

143. Grokking.

💡 Grokking при обучении нейронных сетей — это явление, когда модель после продолжительного обучения сначала демонстрирует плохую обобщающую способность на новых данных, несмотря на хорошее качество на обучающем наборе. Затем, после дальнейшего обучения, модель неожиданно начинает показывать значительно лучшую производительность и на тестовых данных. Это подразумевает, что модель в конечном итоге находит более глубокие и универсальные закономерности, которые позволяют ей лучше обобщать на неизвестные данные.

Modular Division (training on 50% of data)



144. Double Descent.

Double descent — это явление, наблюдаемое при обучении нейронных сетей, когда увеличение количества параметров модели сначала приводит к снижению ошибки на обучающем и тестовом наборах (классическое поведение bias-variance tradeoff), затем происходит резкое увеличение ошибки (первая точка перегиба, связанная с переобучением), после чего, с дальнейшим увеличением количества параметров, ошибка снова начинает уменьшаться, формируя вторую “волну” улучшения. Это поведение отличается от традиционной U-образной кривой, и его понимание важно для эффективной настройки гиперпараметров и выбора архитектуры модели.

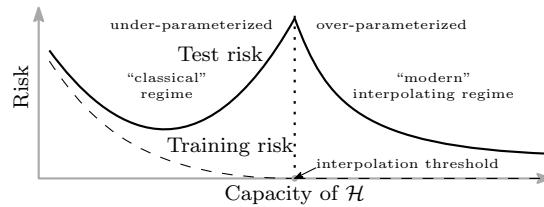


Рисунок 10: Иллюстрация зависимости обобщающей способности модели от размера.

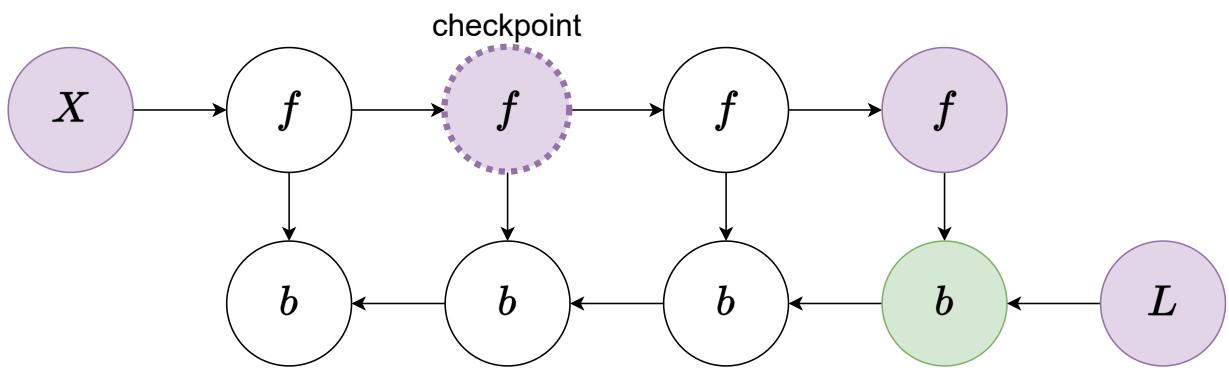
145. Взрыв/Затухание градиентов при обучении глубоких нейронных сетей.

При обучении глубоких нейронных сетей часто возникают проблемы взрыва и затухания градиентов, что приводит к медленной или нестабильной сходимости модели. Эти явления можно описать с помощью производной функции ошибки L по весам сети W . Пусть L - функция потерь, а $\frac{\partial L}{\partial W}$ - градиенты, используемые для обновления весов. Когда сеть имеет много слоев, градиенты вычисляются как произведение матриц Якоби каждого слоя: $\frac{\partial L}{\partial W} = \frac{\partial L}{\partial z^{(n)}} \cdot \frac{\partial z^{(n)}}{\partial z^{(n-1)}} \dots \frac{\partial z^{(2)}}{\partial z^{(1)}} \cdot \frac{\partial z^{(1)}}{\partial W}$, где $z^{(i)}$ - активации i -го слоя. Если значения производных $\frac{\partial z^{(i+1)}}{\partial z^{(i)}}$ в среднем больше единицы, градиенты начинают экспоненциально увеличиваться при обратном распространении, вызывая взрыв градиентов. Напротив, если значения производных меньше единицы, градиенты экспоненциально уменьшаются, что приводит к их затуханию.

146. Идея gradient checkpointing.

💡 Gradient checkpointing — это техника, которая позволяет значительно снизить потребление памяти при обучении глубоких нейронных сетей за счет стратегического пересчета промежуточных активаций во время обратного распространения ошибки. В стандартном процессе обучения с использованием обратного распространения ошибка вычисляется для каждого слоя и промежуточные активации сохраняются в памяти, что требует $O(N)$ памяти, где N — количество слоев в сети.

При gradient checkpointing вместо сохранения активаций для всех слоев, мы сохраняем их только для некоторых стратегически выбранных слоев, называемых чекпоинтами. Активации для остальных слоев пересчитываются на этапе обратного распространения, что снижает общее потребление памяти. Если мы сохраняем активации через каждые k слоев, то потребление памяти уменьшается до $O(\frac{N}{k})$. Однако, это приводит к дополнительным вычислительным затратам, так как активации некоторых слоев пересчитываются несколько раз.



147. Идея аккумуляции градиентов.

💡 Аккумуляция градиентов — это метод, используемый для эффективного обучения больших нейросетевых моделей, когда ограничен объем доступной видеопамяти. Вместо обновления весов модели после каждого батча данных, как это происходит в стандартном стохастическом градиентном спуске (SGD), градиенты накапливаются в течение нескольких батчей. Затем обновление весов происходит только после накопления градиентов от нескольких батчей, эквивалентных одному большому батчу. Этот подход позволяет использовать меньший объем памяти, так как не требуется хранить большие батчи данных в видеопамяти, при этом достигается сходный с большим батчем эффект на обновление весов, что способствует более стабильному и эффективному обучению модели.

148. Зачем увеличивать батч при обучении больших нейросетевых моделей. Warmup.

💡 Если увеличивать размер батча, то, при наличии параллелизма, время прохождения эпохи уменьшается. Эмпирическое правило: когда размер минибатча увеличился в k раз, learning rate также необходимо увеличить в k раз (linear scaling rule). Для адаптивных методов эмпирически используется шкалирование базового learning rate в \sqrt{k} раз (square root scaling rule).

Warmup — это техника, применяемая к процессу обучения моделей, чтобы стабилизировать и улучшить обучение в ранних этапах. В процессе Warmup начальное значение скорости обучения постепенно увеличивается от низкого значения до целевого значения в течение нескольких первых эпох или шагов. Эта техника помогает избежать проблем, связанных с нестабильностью градиентов и резкими изменениями параметров модели в самом начале обучения.

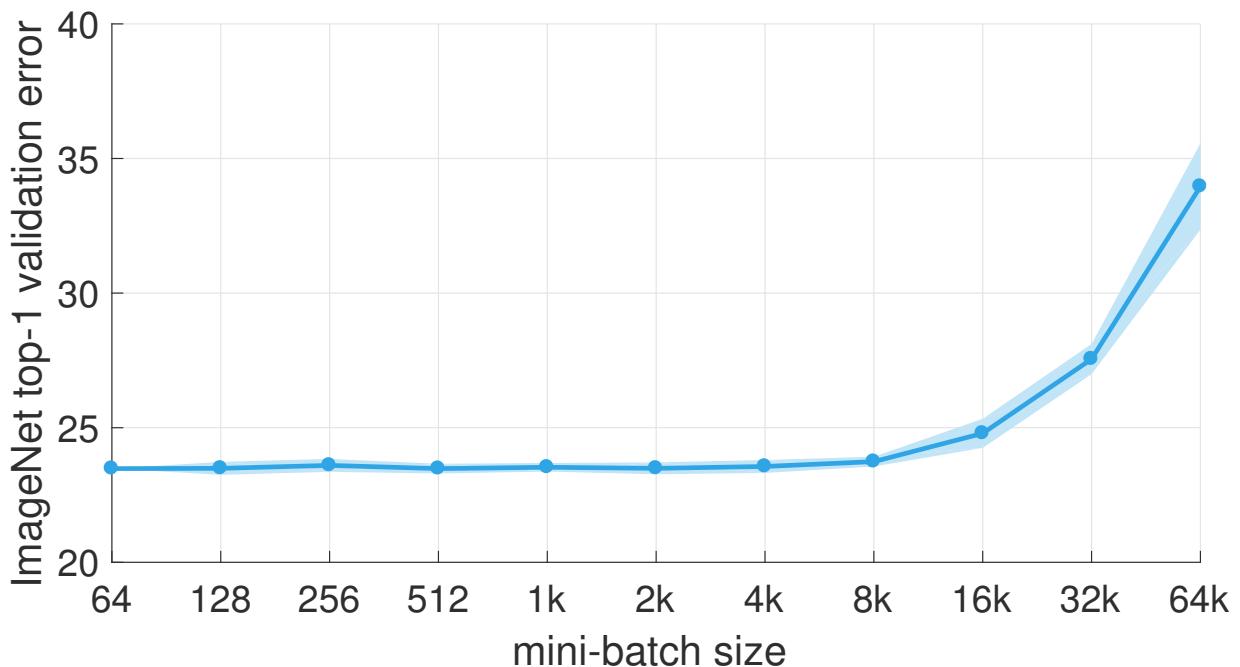
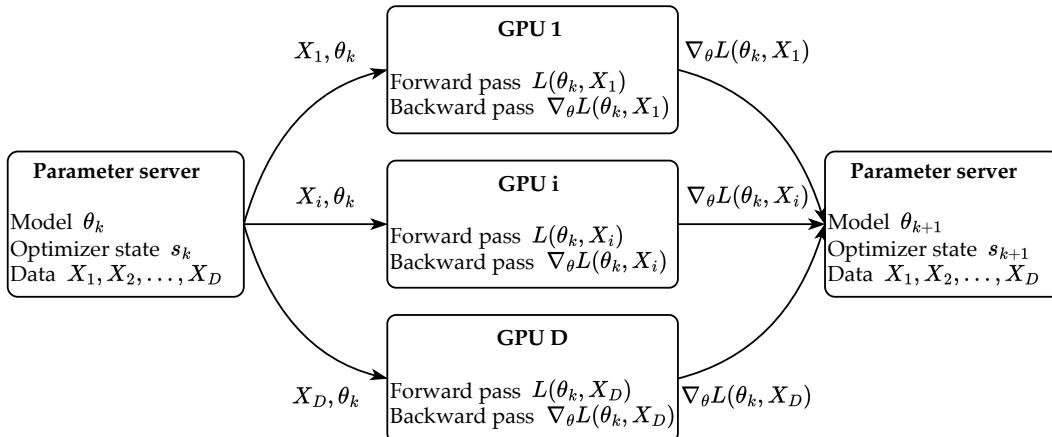


Рисунок 11: Увеличение размера батча приводит к росту ошибки из-за сложности минимизируемой функции.

149. Data Parallel обучение на нескольких видеокартах.



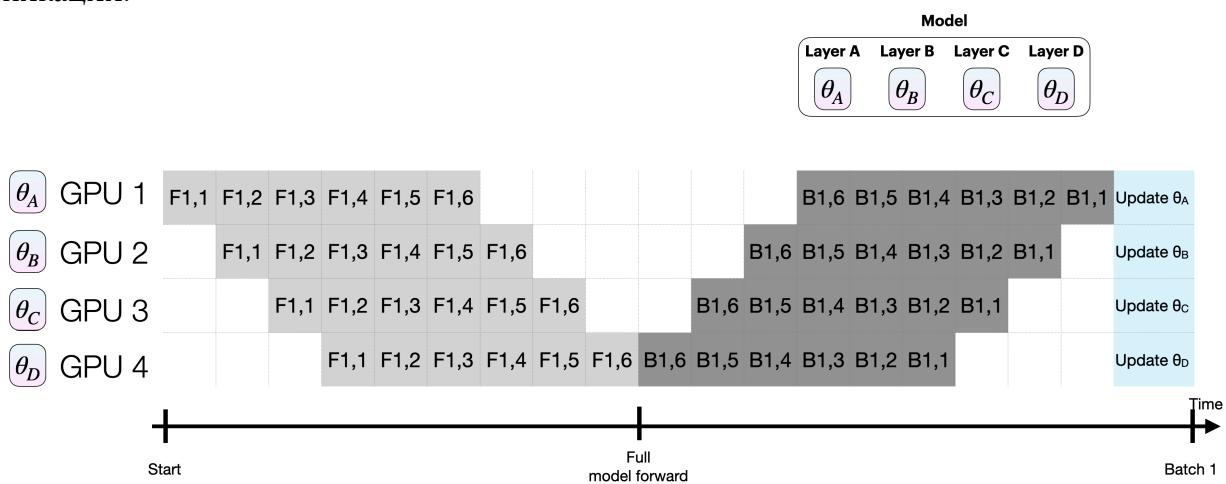
1. Сервер параметров отправляет полную копию модели на каждое устройство
2. Каждое устройство выполняет прямой и обратный проходы
3. Сервер параметров собирает градиенты
4. Сервер параметров обновляет модель



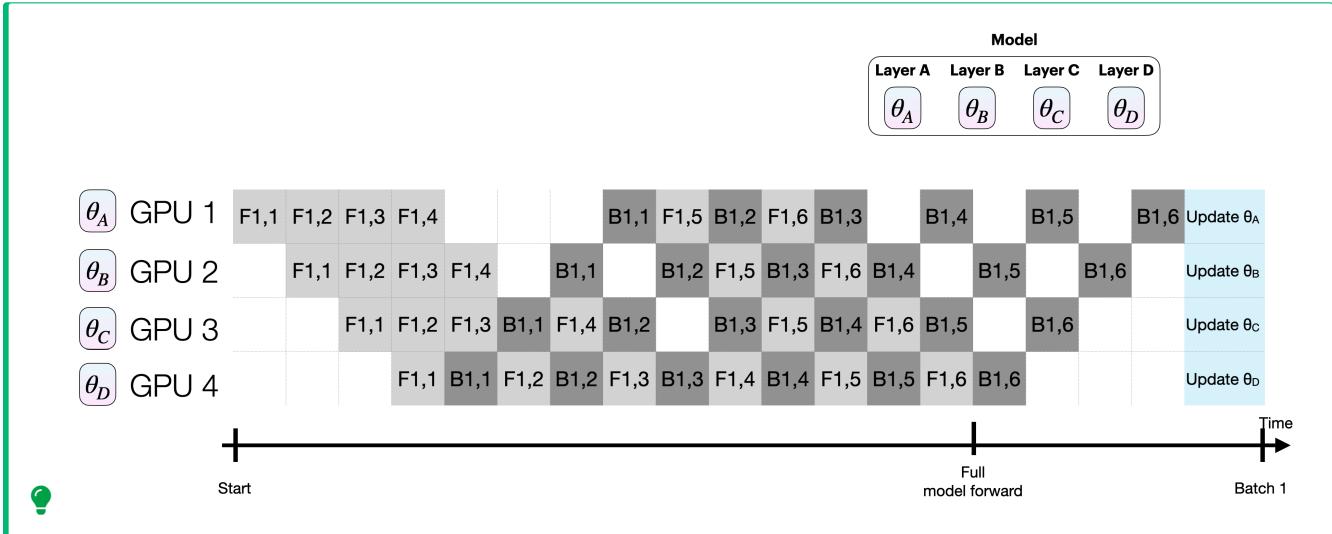
150. GPipe Pipeline параллелизм.



GPipe разделяет модель на этапы, каждый из которых обрабатывается последовательно. Микро-батчи проходят через пайплайн, что позволяет перекрывать вычисления и коммуникации.



151. PipeDream Pipeline параллелизм.



152. Дообучение больших моделей с помощью LoRA адаптеров.

💡 LoRA предполагает дообучать модель в виде поиска добавки к весам малого ранга

$$W_{\text{new}} = W + \Delta W$$

где $\Delta W = AB^T$, причем A и B являются матрицами низкого ранга. Это уменьшает вычислительные затраты, сохраняя производительность модели.

- A инициализируется как обычно, тогда как B инициализируется нулями, чтобы начать с тождественного отображения
 - r обычно выбирается в диапазоне от 2 до 64
 - Обычно применяется к модулям внимания

$$h = W_{\text{new}}x = Wx + \Delta Wx = Wx + AB^Tx$$

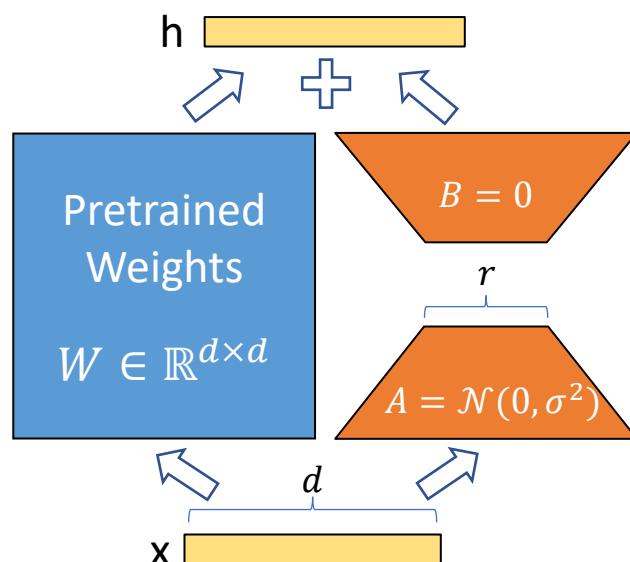


Рисунок 12: Схема работы

153. Метод двойственного градиентного подъема.

💡 Рассматривается задача:

$$f(x) \rightarrow \min_{Ax=b} .$$

Двойственная задача:

$$-f^*(-A^T u) - b^T u \rightarrow \max_u,$$

где $f^*(y) = \max_x [y^T x - f(x)]$ - сопряженная функция. Определим $g(u) = -f^*(-A^T u) - b^T u$, тогда $\partial g(u) = A\partial f^*(-A^T u) - b$. Перепишем это в виде $\partial g(u) = Ax - b$, где $x \in \arg \min_z [f(z) + u^T Az]$. Тогда определим метод двойственного градиентного подъема:

$$\begin{aligned} x_k &\in \arg \min_x [f(x) + (u_{k-1})^T Ax] \\ u_k &= u_{k-1} + \alpha_k (Ax_k - b). \end{aligned}$$

154. Связь константы сильной выпуклости f и гладкости f^* .

💡 Пусть f - замкнутая и выпуклая. Тогда f - сильно выпуклая с константой выпуклости $\mu \Leftrightarrow \nabla f^*$ - липшицев с параметром $\frac{1}{\mu}$.

155. Идея dual decomposition.

💡 Рассматриваем задачу $\sum_{i=1}^B f_i(x_i) \rightarrow \min_{Ax=b}$. Здесь $x = (x_1, \dots, x_B)^T \in \mathbb{R}^n$ разделены на B блоков переменных, каждый $x_i \in \mathbb{R}^{n_i}$. Разделим аналогично матрицу A : $A = [A_1, \dots, A_B]$, где $A_i \in \mathbb{R}^{m \times n_i}$. Тогда

$$x^{\text{new}} \in \arg \min_x \left(\sum_{i=1}^B f_i(x_i) + u^T Ax \right) \Rightarrow x_i^{\text{new}} \in \arg \min_{x_i} (f_i(x_i) + u^T A_i x_i), \quad i = \overline{1, B}$$

Тогда метод двойственного подъема запишется следующим образом:

$$\begin{aligned} x_i^k &\in \arg \min_{x_i} (f_i(x_i) + u^T A_i x_i), \quad i = \overline{1, B} \\ u^k &= u^{k-1} + \alpha_k \left(\sum_{i=1}^B A_i x_i^k - b \right). \end{aligned}$$

156. Метод двойственного градиентного подъема для линейных ограничений-неравенств.

💡 Рассматриваем задачу $\sum_{i=1}^B f_i(x_i) \rightarrow \min_{\sum_{i=1}^B A_i x_i \leq b}$.
 $x_i^k \in \arg \min_{x_i} [f_i(x_i) + (u^{k-1})^T A_i x_i], \quad i = \overline{1, B}$
 $u^k = \left(u^{k-1} + \alpha_k \left[\sum_{i=1}^B A_i x_i^k - b \right] \right)_+$,
где $(u)_+$ обозначает $(u_+)_i = \max\{0, u_i\}, i = \overline{0, m}$.

157. Метод модифицированной функции Лагранжа.

💡 Рассматриваем задачу $f(x) + \frac{\rho}{2} \|Ax - b\|^2 \rightarrow \min_{Ax = b}$, где $\rho > 0$ - параметр. Тогда метод двойственного градиентного подъема имеет вид:

$$\begin{aligned} x_k &= \arg \min_x \left[f(x) + (u_{k-1})^T Ax + \frac{\rho}{2} \|Ax - b\|^2 \right] \\ u_k &= u_{k-1} + \rho(Ax_k - b). \end{aligned}$$

В этом случае имеет место следующее:

$$\begin{aligned} L &= f(x) + u^T(Ax - b) + \frac{\rho}{2} \|Ax - b\|^2 \\ x_k &= \arg \min_x \left[f(x) + (u_{k-1})^T Ax + \frac{\rho}{2} \|Ax - b\|^2 \right] \\ 0 &\in \partial f(x_k) + A^T(u_{k-1} + \rho(Ax_k - b)) \\ 0 &\in \partial f(x_k) + A^T u_k. \end{aligned}$$

158. Метод ADMM.

💡 Рассматриваем задачу

$$\begin{aligned} \min_{x,z} \quad & f(x) + g(z) \\ \text{s.t.} \quad & Ax + Bz = c \end{aligned}$$

После добавления штрафа за выход из бюджетного множества имеем $f(x) + g(z) + \|Ax + Bz - c\|^2 \rightarrow \min_{Ax + Bz = c}$, где $\rho > 0$ - параметр. Тогда функция Лагранжа имеет вид:

$$L_\rho(x, z, u) = f(x) + g(z) + u^T(Ax + Bz - c) + \frac{\rho}{2} \|Ax + Bz - c\|^2.$$

И шаг ADMM записывается как:

$$\begin{aligned} x_k &= \arg \min_x L_\rho(x, z_{k-1}, u_{k-1}) \\ z_k &= \arg \min_z L_\rho(x_k, z, u_{k-1}) \\ u_k &= u_{k-1} + \rho(Ax_k + Bz_k - c). \end{aligned}$$

159. Формулировка задачи линейных наименьших квадратов с ℓ_1 регуляризацией в форме ADMM.

💡 Пусть имеются $b \in \mathbb{R}^n$, $A \in \mathbb{R}^{n \times p}$ и рассматривается задача lasso: $\frac{1}{2}\|Ax - b\|_2^2 + \lambda\|x\|_1$.
Преобразуем проблему к ADMM виду: $\frac{1}{2}\|Ax - b\|_2^2 + \lambda\|z\|_1 \rightarrow \min_{x-z=0}$.

160. Формулировка задачи поиска точки на пересечении двух выпуклых множеств в форме ADMM.

💡 Пусть имеются выпуклые множества $U, V \subseteq \mathbb{R}^n$. Рассматриваем задачу $\mathbb{I}_U(x) + \mathbb{I}_V(x) \rightarrow \min_x$. Преобразуем проблему к ADMM виду: $\mathbb{I}_U(x) + \mathbb{I}_V(z) \rightarrow \min_{x-z=0}$

Теоремы с доказательствами

1. Критерий положительной определенности матрицы через знаки собственных значений матрицы.

i $A \succeq (\succ)0 \iff$ все собственные значения матрицы $A \geq (>)0$

→ Пусть некоторые собственные значения λ отрицательны, и x - соответствующий ему собственный вектор. Тогда:

$$Ax = \lambda x, x^T Ax \geq 0 \rightarrow x^T Ax = \lambda x^T x, x^T x \geq 0 \rightarrow \lambda \geq 0 \text{ - противоречие}$$

← Помним, что положительная определённость задаётся для симметричных матриц. Для симметричной матрицы можем выбрать собственные векторы v_i , образующие ортогональный базис ($i \neq j : v_i^T v_j = 0$ - выкидываем часть слагаемых из суммы в доказательстве). Тогда для $x \in \mathbb{R}^n$

$$x^T Ax = (\alpha_1 v_1 + \dots + \alpha_n v_n)^T A (\alpha_1 v_1 + \dots + \alpha_n v_n) = \sum \alpha_i^2 v_i^T A v_i = \sum \alpha_i^2 v_i^T \lambda v_i$$

Так как $\lambda_i \geq 0$, то и вся сумма неотрицательна. :::

2. Автоматическое дифференцирование. Вычислительный граф. Forward/ Backward mode (в этом вопросе нет доказательств, но необходимо подробно описать алгоритмы).

- i** Для функции $L(w_1, w_2, \dots, w_d)$ вычислительный граф представляет собой ориентированный ациклический граф с двумя типами вершин: w_1, w_2, \dots, w_d и f_1, f_2, \dots . Здесь w_1, w_2, \dots, w_d обозначают входные переменные, а v_1, v_2, \dots обозначают промежуточные функции. В w_1, w_2, \dots, w_d нет входящих ребер, а входящие ребра в v_1, v_2, \dots обозначают значения, подающиеся на вход функциям. Также есть одна единственная вершина, соответствующая итоговому значению L , из которой не выходит ребер. Это описание характерно для функции, используемой в обучении нейросети, где может быть множество вершин-стоков L_1, L_2, \dots .

$$L(w_1, w_2) = w_2 \log w_1 + \sqrt{w_2 \log w_1}$$

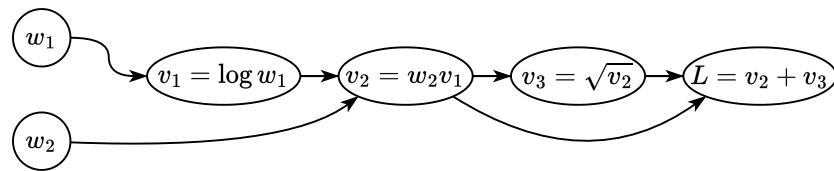


Рисунок 13: Вычислительный график

Для вычисления $\frac{\partial L}{\partial w_k}$ для всех k существуют две конкурирующие процедуры.

Forward mode

Для каждой w_1, w_2, \dots, w_d выполняется следующее: для всех i вычисляются $\dot{v}_i = \frac{\partial v_i}{\partial w_k}$. Для этого выполняется:

- Вычисляется v_i как функция от её родителей (входов) x_1, \dots, x_{t_i} :

$$v_i = v_i(x_1, \dots, x_{t_i})$$

- Вычисляется производная \dot{v}_i с использованием chain rule:

$$\dot{v}_i = \sum_{j=1}^{t_i} \frac{\partial v_i}{\partial x_j} \frac{\partial x_j}{\partial w_k}$$

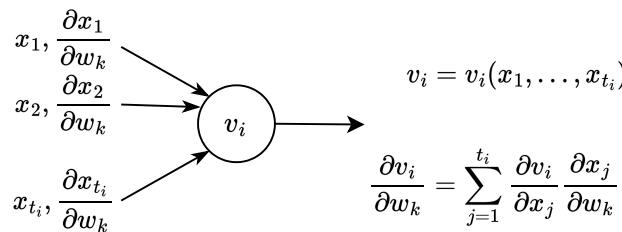


Рисунок 14: Forward mode

Заметим, что для вычисления $\frac{\partial L}{\partial w_k}$ требуется $O(T)$ операций умножения и сложения градиентов, где $T \sim$ — число элементов в вычислительном графе. Итого $O(dT)$.

Backward mode

Forward pass

- Вычисляется v_i как функция от её родителей (входов) x_1, \dots, x_{t_i} :

$$v_i = v_i(x_1, \dots, x_{t_i})$$

Backward pass

Обрабатывая вершины в обратном порядке топологического упорядочения:

- Вычисляется производная \dot{v}_i , используя правило обратной цепочки и информацию от всех его детей (выходов) (x_1, \dots, x_{t_i}) , используя предподсчитанные значения на forward pass:

$$\dot{v}_i = \frac{\partial L}{\partial v_i} = \sum_{j=1}^{t_i} \frac{\partial L}{\partial x_j} \frac{\partial x_j}{\partial v_i}$$

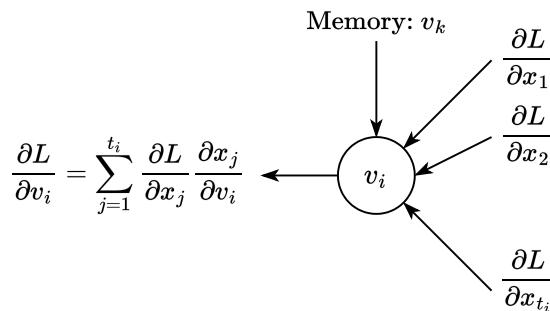


Рисунок 15: Backward mode

Backward mode быстрее для единственного выхода L . Работает аналогично за линейное время.

- Метод дихотомии и золотого сечения для унимодальных функций. Скорость сходимости.

i Методы локализации решения для скалярной минимизации. Сходятся линейно.

Метод дихотомии

Решаем следующую задачу:

$$f(x) \rightarrow \min_{x \in [a, b]}$$

где $f(x)$ — унимодальная функция.

Мы хотим на каждом шаге вдвое сокращать область, в которой ищем минимум. Для этого будем пользоваться основным свойством унимодальных функций:

$$\forall a \leq x_1 < x_2 \leq b :$$

$$f(x_1) \leq f(x_2) \Rightarrow x_* \in [a, x_2]$$

$$f(x_1) \geq f(x_2) \Rightarrow x_* \in [x_1, b]$$

где x_* — точка, в которой достигается минимум

Алгоритм:

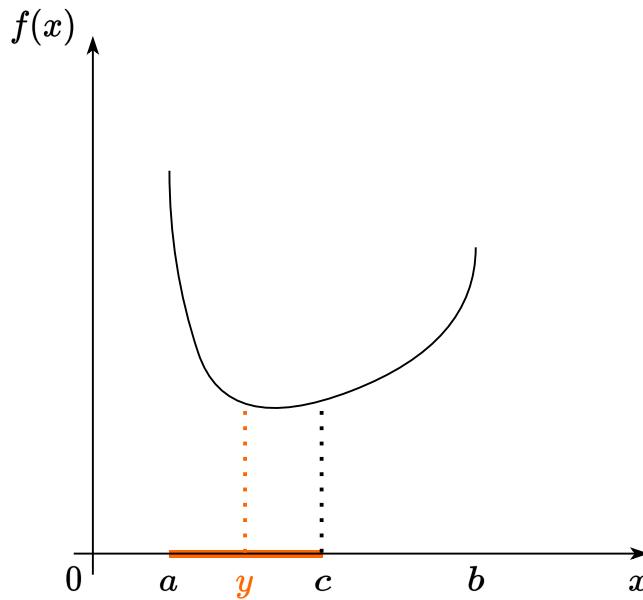


Рисунок 16: Алгоритм дихотомии

Можно заметить, что на каждой итерации требуется не более 2-х вычислений значения функции.

Сходимость метода дихотомии

Длина отрезка на $k + 1$ итерации:

$$\Delta_{k+1} = b_{k+1} - a_{k+1} = \frac{1}{2^k}(b - a)$$

Если будем выбирать середину отрезка как выход $k + 1$ итерации:

$$|x_{k+1} - x_*| \leq \frac{\Delta_{k+1}}{2}$$

Подставим полученное ранее выражение для длины отрезка:

$$|x_{k+1} - x_*| \leq \frac{1}{2^{k+1}}(b - a)$$

$$|x_{k+1} - x_*| \leq (0.5)^{k+1}(b - a)$$

Получили выражение для сходимости по итерациям. Отсюда также можно выразить необходимое количество итераций для достижения точности ε :

$$K = \left\lceil \log_2 \frac{b-a}{\varepsilon} - 1 \right\rceil$$

Теперь получим выражение для сходимости по количеству вычислений значения функции. Знам, что на каждой итерации вычисляем значение не более 2-х раз, значит количество вычислений значения функции возьмём $N = 2k$:

$$\begin{aligned}|x_{k+1} - x_*| &\leq (0.5)^{\frac{N}{2}+1}(b-a) \\ |x_{k+1} - x_*| &\leq (0.707)^N \frac{b-a}{2}\end{aligned}$$

Метод золотого сечения

Идея такая же, как и в методе дихотомии, но хотим уменьшить количество вычислений значения функции. Для этого будем вычислять значения в точках золотого сечения. Так на каждой итерации нам нужно будет вычислять значение только в одной точке, так как для нового отрезка в одной из точек золотого сечения значение будет уже посчитано:

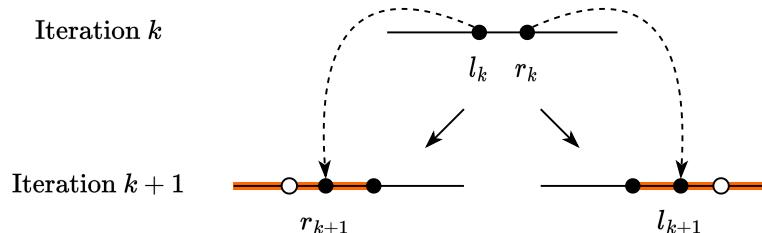


Рисунок 17: Золотое сечение

Алгоритм:

```
def golden_search(f, a, b, epsilon):
    tau = (sqrt(5) + 1) / 2
    y = a + (b - a) / tau**2
    z = a + (b - a) / tau
    while b - a > epsilon:
        if f(y) <= f(z):
            b = z
            z = y
            y = a + (b - a) / tau**2
        else:
            a = y
            y = z
            z = a + (b - a) / tau
    return (a + b) / 2
```

Сходимость метода золотого сечения

На каждой итерации длина отрезка будет уменьшаться в $\tau = \frac{\sqrt{5}+1}{2}$ раз. Тогда оценка сходимости (и по итерациям, и по вычислениям значений функции):

$$|x_{k+1} - x_*| \leq \frac{b_{k+1} - a_{k+1}}{2} = \left(\frac{1}{\tau}\right)^{N-1} \frac{b-a}{2} \approx 0.618^k \frac{b-a}{2}$$

Получили сходимость по итерациям хуже, чем у дихотомии, так как отрезки уменьшаются слабее на каждой итерации. Но по количеству вычислений значения функции сходимость у метода золотого сечения быстрее.

4. Базовые операции, сохраняющие выпуклость множеств: пересечение бесконечного числа множеств, линейная комбинация множеств, образ афинного отображения.



- Пересечение любого (!) количества выпуклых множеств — выпуклое множество.
- Линейная комбинация выпуклых множеств выпукла.
- Образ выпуклого множества после применения афинного отображения — выпуклое множество.

Пересечение бесконечного числа множеств

Пересечение любого (!) количества выпуклых множеств — выпуклое множество.

Если итоговое пересечение пустое или содержит одну точку, то свойство выпуклости выполняется по определению. Иначе возьмем 2 точки и отрезок между ними. Эти точки должны лежать во всех пересекаемых множествах. Так как все пересекаемые множества выпуклы, отрезок между этими двумя точками лежит во всех множествах. А значит, отрезок лежит и в их пересечении.

Линейная комбинация множеств

Линейная комбинация выпуклых множеств выпукла.

Пусть есть 2 выпуклых множества S_x, S_y , рассмотрим их линейную комбинацию

$$S = \{s \mid s = c_1 x + c_2 y, x \in S_x, y \in S_y, c_1, c_2 \in \mathbb{R}\}$$

Возьмем две точки из S : $s_1 = c_1 x_1 + c_2 y_1, s_2 = c_1 x_2 + c_2 y_2$ и докажем, что отрезок между ними $\theta s_1 + (1-\theta)s_2, \theta \in [0, 1]$ также принадлежит S

$$\begin{aligned} & \theta s_1 + (1-\theta)s_2 \\ & \theta(c_1 x_1 + c_2 y_1) + (1-\theta)(c_1 x_2 + c_2 y_2) \\ & c_1(\theta x_1 + (1-\theta)x_2) + c_2(\theta y_1 + (1-\theta)y_2) \\ & c_1 x + c_2 y \in S \end{aligned}$$

Образ афинного отображения

Образ выпуклого множества после применения афинного отображения — выпуклое множество.

$$S \subseteq \mathbb{R}^n \text{ выпукло} \rightarrow f(S) = \{f(x) \mid x \in S\} \text{ выпукло} \quad (f(x) = \mathbf{A}x + \mathbf{b})$$

Доказательство

При $\theta \in [0, 1]; x, y \in S, S$ — выпуклое. Тогда и $\theta x + (1-\theta)y \in S$. В то же время $f(\theta x + (1-\theta)y) = \theta Ax + \theta b + (1-\theta)Ay + (1-\theta)b = \theta Ax + (1-\theta)Ay + b = \theta f(x) + (1-\theta)f(y)$. В итоге мы доказали,

что образ $f(S)$ — тоже выпуклый, так как $\forall \theta \in [0, 1], x, y \in S$ выполняется $f(x) + (1 - \theta)f(y) \leq f(S)$.

Примеры афинных функций: растяжение, сжатие, проекция, транспонирование, множество решений линейного матричного неравенства $\{x \mid x_1 A_1 + \dots + x_m A_m \leq B\}$. Здесь $A_i, B \in \mathbf{S}^p$ — симметричные матрицы $p \times p$.

Заметим также, что прообраз выпуклого множества при аффинном отображении также является выпуклым.

$$S \subseteq \mathbb{R}^m \text{ convex } \rightarrow f^{-1}(S) = \{x \in \mathbb{R}^n \mid f(x) \in S\} \text{ convex } (f(x) = \mathbf{A}x + \mathbf{b})$$

5. Неравенство Йенсена для выпуклой функции и выпуклой комбинации точек.

i Пусть $f(x)$ — выпуклая функция, определённая на выпуклом множестве $S \subseteq \mathbb{R}^n$. Тогда для точек $x_1, \dots, x_m \in S$ выполнено неравенство:

$$f\left(\sum_{i=1}^m \lambda_i x_i\right) \leq \sum_{i=1}^m \lambda_i f(x_i)$$

$$\lambda = [\lambda_1, \dots, \lambda_m] \in \Delta_m.$$

1. Заметим, что $\sum_{i=1}^m \lambda_i x_i$ является выпуклой комбинацией элементов S и лежит в S .

2. Доказательство по индукции. Для $m = 1$ очевидно, для $m = 2$ следует из определения выпуклой функции.

3. Пусть неравенство верно для $m = 1, \dots, k$, докажем для $m = k + 1$. Пусть $\lambda \in \Delta_{k+1}$, $x = \sum_{i=1}^{k+1} \lambda_i x_i = \lambda_{k+1} x_{k+1} + \sum_{i=1}^k \lambda_i x_i$. При $\lambda_i = 0$ либо 1 выражение сводится к уже рассмотренным случаям, далее полагаем $0 < \lambda_i < 1$:

$$x = \lambda_{k+1} x_{k+1} + (1 - \lambda_{k+1}) \sum_{i=1}^k \frac{\lambda_i}{1 - \lambda_{k+1}} x_i = \lambda_{k+1} x_{k+1} + (1 - \lambda_{k+1}) \hat{x}$$

где $\hat{x} = \sum_{i=1}^k \gamma_i x_i$ и $\gamma_i = \frac{\lambda_i}{1 - \lambda_{k+1}} \geq 0$, $1 \geq i \geq k$.

4. Так как $\lambda \in \Delta_{k+1}$, то $\gamma = [\gamma_1, \dots, \gamma_k] \in \Delta_k$. Значит, $\hat{x} \in S$, из выпуклости $f(x)$ и предположения индукции следует:

$$f\left(\sum_{i=1}^{k+1} \lambda_i x_i\right) = f(\lambda_{k+1} x_{k+1} + (1 - \lambda_{k+1}) \hat{x}) \leq \lambda_{k+1} f(x_{k+1}) + (1 - \lambda_{k+1}) f(\hat{x}) \leq \sum_{i=1}^{k+1} \lambda_i f(x_i)$$

6. Выпуклость надграфика как критерий выпуклости функции.

i Чтобы функция $f(x)$, определенная на выпуклом множестве X , была выпуклой на X , необходимо и достаточно чтобы надграфик f был выпуклым множеством.

Для функции $f(x)$, определенной на $S \subseteq \mathbb{R}^n$, множество:

$$\text{epi } f = \{[x, \mu] \in S \times \mathbb{R} : f(x) \leq \mu\}$$

называется **надграфиком** функции $f(x)$ (здесь $\mu \in \mathbb{R}, x \in S$).

Необходимость

Предположим, что $f(x)$ выпукла на X . Возьмем две произвольные точки $[x_1, \mu_1] \in \text{epi } f$ и $[x_2, \mu_2] \in \text{epi } f$. Также возьмем $0 \leq \lambda \leq 1$ и обозначим $x_\lambda = \lambda x_1 + (1 - \lambda)x_2, \mu_\lambda = \lambda\mu_1 + (1 - \lambda)\mu_2$. Тогда,

$$\lambda \begin{bmatrix} x_1 \\ \mu_1 \end{bmatrix} + (1 - \lambda) \begin{bmatrix} x_2 \\ \mu_2 \end{bmatrix} = \begin{bmatrix} x_\lambda \\ \mu_\lambda \end{bmatrix}.$$

Из выпуклости X следует, что $x_\lambda \in X$. Более того, так как $f(x)$ – выпуклая функция, то

$$f(x_\lambda) \leq \lambda f(x_1) + (1 - \lambda)f(x_2) \leq \lambda\mu_1 + (1 - \lambda)\mu_2 = \mu_\lambda$$

Из неравенства выше по определению надграфика следует, что $\begin{bmatrix} x_\lambda \\ \mu_\lambda \end{bmatrix} \in \text{epi } f$. Следовательно, надграфик f – выпуклое множество.

Достаточность

Предположим, что надграфик f , $\text{epi } f$, выпуклое множество. Тогда, исходя из того что $[x_1, \mu_1] \in \text{epi } f$ и $[x_2, \mu_2] \in \text{epi } f$, получаем

$$\begin{bmatrix} x_\lambda \\ \mu_\lambda \end{bmatrix} = \lambda \begin{bmatrix} x_1 \\ \mu_1 \end{bmatrix} + (1 - \lambda) \begin{bmatrix} x_2 \\ \mu_2 \end{bmatrix} \in \text{epi } f$$

для любого $0 \leq \lambda \leq 1$.

Следовательно, из определения надграфика, подставив значение μ_λ , получаем, что $f(x_\lambda) \leq \mu_\lambda = \lambda\mu_1 + (1 - \lambda)\mu_2$.

$$f(x_\lambda) = f(\lambda x_1 + (1 - \lambda)x_2) \leq \mu_\lambda = \lambda\mu_1 + (1 - \lambda)\mu_2$$

Но это верно для всех $\mu_1 \geq f(x_1)$ и $\mu_2 \geq f(x_2)$, в том числе и при $\mu_1 = f(x_1)$ и $\mu_2 = f(x_2)$. Тогда мы получаем неравенство:

$$f(x_\lambda) \leq \lambda f(x_1) + (1 - \lambda)f(x_2)$$

Так как $x_1 \in X$ и $x_2 \in X$ выбирались произвольно, $f(x)$ – выпуклая функция на X .

7. Дифференциальный критерий сильной выпуклости первого порядка.

i Пусть $f(x)$ – дифференцируемая функция на выпуклом множестве $X \subseteq \mathbb{R}^n$. Тогда $f(x)$ сильно выпукла на X с константой $\mu > 0$ тогда и только тогда, когда

$$f(x) - f(x_0) \geq \langle \nabla f(x_0), x - x_0 \rangle + \frac{\mu}{2} \|x - x_0\|^2$$

для всех $x, x_0 \in X$.

Необходимость

Пусть $0 < \lambda \leq 1$. Согласно определению сильно выпуклой функции,

$$f(\lambda x + (1 - \lambda)x_0) \leq \lambda f(x) + (1 - \lambda)f(x_0) - \frac{\mu}{2}\lambda(1 - \lambda)\|x - x_0\|^2$$

или эквивалентно,

$$\begin{aligned} f(x) - f(x_0) - \frac{\mu}{2}(1 - \lambda)\|x - x_0\|^2 &\geq \frac{1}{\lambda}[f(\lambda x + (1 - \lambda)x_0) - f(x_0)] = \\ &= \frac{1}{\lambda}[f(x_0 + \lambda(x - x_0)) - f(x_0)] = \frac{1}{\lambda}[\lambda\langle\nabla f(x_0), x - x_0\rangle + o(\lambda)] = \\ &= \langle\nabla f(x_0), x - x_0\rangle + \frac{o(\lambda)}{\lambda}. \end{aligned}$$

Таким образом, переходя к пределу при $\lambda \rightarrow 0$, мы приходим к первоначальному утверждению.

Достаточность

Предположим, что неравенство в теореме выполнено для всех $x, x_0 \in X$. Возьмем $x_0 = \lambda x_1 + (1 - \lambda)x_2$, где $x_1, x_2 \in X$, $0 \leq \lambda \leq 1$. Согласно неравенству из условия теоремы, выполняются следующие неравенства:

$$f(x_1) - f(x_0) \geq \langle\nabla f(x_0), x_1 - x_0\rangle + \frac{\mu}{2}\|x_1 - x_0\|^2,$$

$$f(x_2) - f(x_0) \geq \langle\nabla f(x_0), x_2 - x_0\rangle + \frac{\mu}{2}\|x_2 - x_0\|^2.$$

Умножая первое неравенство на λ и второе на $1 - \lambda$ и складывая их, учитывая, что

$$x_1 - x_0 = (1 - \lambda)(x_1 - x_2), \quad x_2 - x_0 = \lambda(x_2 - x_1),$$

и что $\lambda(1 - \lambda)^2 + \lambda^2(1 - \lambda) = \lambda(1 - \lambda)$, получаем:

$$\begin{aligned} \lambda f(x_1) + (1 - \lambda)f(x_2) - f(x_0) - \frac{\mu}{2}\lambda(1 - \lambda)\|x_1 - x_2\|^2 &\geq \\ &\geq \langle\nabla f(x_0), \lambda x_1 + (1 - \lambda)x_2 - x_0\rangle = 0. \end{aligned}$$

Таким образом, неравенство из определения сильно выпуклой функции выполнено. Важно отметить, что при $\mu = 0$ получаем случай выпуклости и соответствующий дифференциальный критерий.

8. Дифференциальный критерий сильной выпуклости второго порядка.

i Пусть $X \subseteq \mathbb{R}^n$ — выпуклое множество с непустой внутренностью. Пусть также $f(x)$ — дважды непрерывно дифференцируемая функция на X . Тогда $f(x)$ сильно выпукла на X с константой $\mu > 0$ тогда и только тогда, когда

$$\langle y, \nabla^2 f(x)y \rangle \geq \mu \|y\|^2$$

для всех $x \in X$ и $y \in \mathbb{R}^n$.

Другая форма записи:

$$\nabla^2 f(x) \succcurlyeq \mu I$$

Целевое неравенство тривиально, когда $y = 0_n$, поэтому предположим, что $y \neq 0_n$.

Необходимость

Пусть x является внутренней точкой X . Тогда $x + \alpha y \in X$ для всех $y \in \mathbb{R}^n$ и достаточно малых α . Поскольку $f(x)$ дважды дифференцируема,

$$f(x + \alpha y) = f(x) + \alpha \langle \nabla f(x), y \rangle + \frac{\alpha^2}{2} \langle y, \nabla^2 f(x)y \rangle + o(\alpha^2).$$

Основываясь на критерии первого порядка сильной выпуклости, имеем

$$\frac{\alpha^2}{2} \langle y, \nabla^2 f(x)y \rangle + o(\alpha^2) = f(x + \alpha y) - f(x) - \alpha \langle \nabla f(x), y \rangle \geq \frac{\mu}{2} \alpha^2 \|y\|^2.$$

Это неравенство сводится к целевому неравенству после деления обеих частей на α^2 и перехода к пределу при $\alpha \rightarrow 0$.

Если $x \in X$, но $x \notin \text{int } X$, рассмотрим последовательность $\{x_k\}$ такую, что $x_k \in \text{int } X$ и $x_k \rightarrow x$ при $k \rightarrow \infty$. Тогда мы приходим к целевому неравенству после перехода к пределу.

Достаточность

Формула Тейлора с остаточным членом Лагранжа второго порядка $\forall x, y : x, x + y \in X$ найдется α такая, что:

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

где $0 < \alpha < 1$.

Используя формулу Тейлора с остаточным членом Лагранжа и неравенство из условия, получаем для $x + y \in X$:

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

где $0 \leq \alpha \leq 1$. Следовательно,

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

Таким образом, по критерию первого порядка сильной выпуклости, функция $f(x)$ является сильно выпуклой с константой μ . Важно отметить, что $\mu = 0$ соответствует случаю выпуклости и соответствующему дифференциальному критерию.

9. Необходимые условия безусловного экстремума.

- i** Если в x^* достигается локальный минимум и f непрерывно дифференцируема в открытой окрестности, то

$$\nabla f(x^*) = 0$$

Предположим обратное. Пусть $\nabla f(x^*) \neq 0$. Рассмотрим вектор $p = -\nabla f(x^*)$ и заметим, что

$$p^T \nabla f(x^*) = -\|\nabla f(x^*)\|^2 < 0$$

Так как ∇f непрерывна в окрестности x^* , то существует скаляр $T > 0$ такой, что

$$p^T \nabla f(x^* + tp) < 0, \text{ для любого } t \in [0, T]$$

Для любого $\bar{t} \in (0, T]$, мы можем воспользоваться теоремой Тейлора:

$$f(x^* + \bar{t}p) = f(x^*) + \bar{t}p^T \nabla f(x^* + tp), \text{ для некоторого } t \in (0, \bar{t})$$

Следовательно, $f(x^* + \bar{t}p) < f(x^*)$ для любого $\bar{t} \in (0, T]$. Мы нашли направление, идя вдоль которого из x^* функция f убывает. Тогда x^* – не точка локального минимума. Получили противоречие.

10. Достаточные условия безусловного экстремума.

- i** Пусть $\nabla^2 f$ непрерывна в открытой окрестности x^* и

$$\nabla f(x^*) = 0 \quad \nabla^2 f(x^*) \succ 0.$$

Тогда x^* – точка локального минимума f .

Так как гессиан непрерывен и положительно определен в x^* , то мы можем выбрать радиус $r > 0$ такой, что $\nabla^2 f(x)$ остается положительно определенной для всех x в открытом шаре $B = \{z \mid \|z - x^*\| < r\}$. Взяв любой ненулевой вектор p , для которого выполняется $\|p\| < r$, мы получаем $x^* + p \in B$, а также по формуле Тейлора:

$$\begin{aligned} f(x^* + p) &= f(x^*) + p^T \nabla f(x^*) + \frac{1}{2} p^T \nabla^2 f(z)p \\ &= f(x^*) + \frac{1}{2} p^T \nabla^2 f(z)p \end{aligned}$$

где $z = x^* + tp$ для некоторого $t \in (0, 1)$. Так как $z \in B$, мы получаем $p^T \nabla^2 f(z)p > 0$, и следовательно $f(x^* + p) > f(x^*)$. Таким образом x^* – точка локального минимума.

11. Формулировка симплекс метода для задачи линейного программирования в стандартной форме. Теорема о проверке оптимальности решения.

- i** Если все элементы λ_B неположительны и базис B допустимый, тогда базис B оптимальен. Здесь λ_B это коэффициенты при разложении c по базису B : $\lambda_B^T A_B = c^T \Rightarrow \lambda_B^T = c^T A_B^{-1}$.

Формулировка симплекс метода для задачи линейного программирования в стандартной форме. Теорема о проверке оптимальности решения

Задача линейного программирования:

Пусть $c \in \mathbb{R}^n$, $b \in \mathbb{R}^m$, $A \in \mathbb{R}^{m \times n}$, тогда задача формулируется так:

$$\begin{aligned} & \min_{x \in \mathbb{R}^n} c^T x \\ & \text{s.t. } Ax \leq b \end{aligned}$$

Идейное описание симплекс метода:

1. Убедится, что точка, в которой мы находимся, является угловой
2. Проверить оптимальность точки
3. Если необходимо, сменить угол (то есть сменить базис)
4. Повторять до схождения

Шаги выполнения симплекс метода:

1. Поиск начальной базисной допустимой точки:

- Выберем начальную базисную (она является решением системы $A_B x = b_B$, где B - базис размера n пространства, а матрица A обычно имеет больше n ограничений) допустимую ($Ax_0 \leq b$) точку x_0 (искать ее будем через двухфазный симплексметод). Если такая точка не найдена, задача не имеет допустимого решения.

2. Проверка оптимальности:

- Разложение вектора c в данном базисе B с коэффициентами λ_B :

$$\lambda_B^\top A_B = c^\top \quad \text{или} \quad \lambda_B^\top = c^\top A_B^{-1}$$

- Если все компоненты λ_B неположительны, текущий базис является оптимальным. Иначе далее меняем вершину симплекса.

3. Определение переменной для удаления из базиса:

- Если в разложении λ_B есть положительные координаты, продолжаем оптимизацию. Пусть $\lambda_B^k > 0$. Необходимо исключить k из базиса. Рассчитаем направляющий вектор d , идя вдоль которого изменим вершину следующим образом: во-первых, для вектором всех ограничений из базиса, которые мы оставляем, направление должно быть им ортогонально, и, во-вторых, вдоль него значение, связанное с нашим ограничением, должно убывать:

$$\begin{cases} A_{B \setminus \{k\}} d = 0 \\ a_k^\top d < 0 \end{cases}$$

4. Вычисление шага вдоль выбранного направления d :

- Для всех $j \notin B$ считаем шаг:

$$\mu_j = \frac{b_j - a_j^\top x_B}{a_j^\top d}$$

- Новая вершина, которую добавим в базис:

$$t = \arg \min_j \{\mu_j \mid \mu_j > 0\}$$

5. Обновление базиса:

- Обновляем базис и текущее решение:

$$\begin{aligned} B' &= B \setminus \{k\} \cup \{t\}, \\ x_{B'} &= x_B + \mu_t d = A_{B'}^{-1} b_{B'} \end{aligned}$$

- Изменение базиса приводит к уменьшению значения целевой функции:

$$c^\top x_{B'} = c^\top (x_B + \mu_t d) = c^\top x_B + \mu_t c^\top d$$

6. Повторение:

- Далее повторяем шаги 2-5 до достижения оптимального решения или установления, что задача не имеет допустимого решения.

Теорема о проверке оптимальности решения:

Если все элементы λ_B неположительны и базис B достижим, тогда базис B оптimalен.

Здесь λ_B это коэффициенты при разложении c по базису B : $\lambda_B^T A_B = c^T \Rightarrow \lambda_B^T = c^T A_B^{-1}$.

Доказательство:

Предположим противное (что этот базис не оптimalен), пусть $\exists x^* : Ax^* \leq b$ и при этом $c^T x^* < c^T x_B$. Так как для всей матрицы A и вектора b неравенство верно, то и для подматрицы оно верно:

$$A_B x^* \leq b_B$$

Так как все элементы λ_B неположительны, то домножим строки на соответствующие элементы и сложим:

$$\lambda_B^T A_B x^* \geq \lambda_B^T b_B$$

$$c^T x^* \geq \lambda_B^T b_B = \lambda_B^T A_B x_B = c^T x_B$$

Противоречие.

12. Теорема сходимости градиентного спуска для гладких выпуклых функций.

i Рассматриваем задачу

$$f(x) \rightarrow \min_{x \in \mathbb{R}^d}$$

и предполагаем, что f - выпуклая, L -гладкая, $L > 0$.

Пусть $(x_T)_{T \in \mathbb{N}}$ это последовательность, созданная градиентным спуском с постоянным шагом α , $0 < \alpha \leq \frac{1}{L}$. Тогда градиентный спуск сходится сублинейно, то есть:

$$f(x_T) - f^* \leq \frac{L \|x^0 - x^*\|^2}{2T}.$$

$$x_{k+1} = x_k - \alpha \nabla f(x_k) \Rightarrow x_{k+1} - x_k = -\alpha \nabla f(x_k)$$

L -гладкость: $\forall x, y : f(y) \leq f(x) + \langle \nabla f(x), y - x \rangle + \frac{L}{2} \|y - x\|^2$

$$y := x_{k+1}, x := x_k \Rightarrow f(x_{k+1}) \leq f(x_k) + \langle \nabla f(x_k), -\alpha \nabla f(x_k) \rangle + \frac{L}{2} \alpha^2 \|\nabla f(x_k)\|^2$$

$$f(x_{k+1}) \leq f(x_k) - \alpha \|\nabla f(x_k)\|^2 + \frac{L}{2} \alpha^2 \|\nabla f(x_k)\|^2 \quad (1)$$

$(\frac{L}{2} \alpha^2 - \alpha) \rightarrow \min_{\alpha}$. Получаем оптимальный шаг: $\alpha = \frac{1}{L}$ и $f(x_k) - f(x_{k+1}) \geq \frac{1}{2L} \|\nabla f(x_k)\|^2$

Выпуклость: $f(y) \geq f(x) + \nabla f(x)^T (y - x)$

$$\begin{aligned} y &:= x^*, x := x_k \Rightarrow f(x^*) \geq f(x_k) + \nabla f(x_k)^T (x^* - x_k) \Rightarrow \\ &\Rightarrow f(x_k) \leq f(x^*) + \nabla f(x_k)^T (x_k - x^*) \Rightarrow f(x_k) - f(x^*) \leq \nabla f(x_k)^T (x_k - x^*) \end{aligned}$$

Подставим $f(x_k)$ в (1):

$$f(x_{k+1}) \leq f(x_k) - \frac{1}{2L} \|\nabla f(x_k)\|^2 \leq f(x^*) + \nabla f(x_k)^T (x_k - x^*) - \frac{1}{2L} \|\nabla f(x_k)\|^2. \quad (2)$$

Заметим, что $\forall a, b \in \mathbb{R}^d$:

$$\begin{aligned} (a - b)^T (a + b) &= a^T a - b^T a + a^T b - b^T b = a^T a - b^T b = \|a\|^2 - \|b\|^2. \\ a &:= x^* - x_k, b := x_k - x^* - \frac{1}{L} \nabla f(x_k) \Rightarrow \nabla f(x_k)^T (x_k - x^*) - \frac{1}{2L} \|\nabla f(x_k)\|^2 = \\ &= \langle \nabla f(x_k), x_k - x^* \rangle - \frac{1}{2L} \langle \nabla f(x_k), \nabla f(x_k) \rangle = \\ &= \frac{L}{2} \left(\left\langle \frac{\nabla f(x_k)}{L}, 2x_k + 2x^* - \frac{\nabla f(x_k)}{L} \right\rangle \right) = \end{aligned}$$

$$= \frac{L}{2} (b - a)^T (-a - b) = \frac{L}{2} (a - b)^T (a + b) = \frac{L}{2} (\|a\|^2 - \|b\|^2) = \frac{L}{2} (\|x_k - x^*\|^2 - \|x_k - x^* - \frac{1}{L} \nabla f(x_k)\|^2)$$

Подставим в (2):

$$f(x_{k+1}) \leq f(x^*) + \frac{L}{2} \left(\|x_k - x^*\|^2 - \|x_k - x^* - \frac{1}{L} \nabla f(x_k)\|^2 \right) = f(x^*) + \frac{L}{2} (\|x_k - x^*\|^2 - \|x_{k+1} - x^*\|^2)$$

Просуммируем ($R^2 = \|x_0 - x^*\|^2$):

$$\begin{aligned} \sum_{k=0}^{T-1} (f(x_{k+1}) - f(x^*)) &\leq \frac{L}{2} (\|x_0 - x^*\|^2 - \|x_T - x^*\|^2) = \frac{L}{2} (R^2 - \|x_T - x^*\|^2) \leq \frac{LR^2}{2} \\ \frac{1}{T} \sum_{k=0}^{T-1} f(x_{k+1}) - f(x^*) &\leq \frac{LR^2}{2T} \end{aligned}$$

Заметим, что $f(x_T) \leq f(x_i) \forall i = \overline{1, T-1} \Rightarrow f(x_T) \leq \frac{1}{T} \sum_{k=0}^{T-1} f(x_{k+1})$

Итого имеем:

$$f(x_T) - f(x^*) \leq \frac{LR^2}{2T}$$

То есть сходимость сублинейная.

13. Теорема сходимости градиентного спуска для гладких PL функций.

i Рассматриваем задачу

$$f(x) \rightarrow \min_{x \in \mathbb{R}^d}$$

и предполагаем, что f - μ -PL-функция, L -гладкая, для некоторых $L \geq \mu > 0$.

Пусть $(x_T)_{T \in \mathbb{N}}$ это последовательность, созданная градиентным спуском с постоянным шагом $\alpha : 0 < \alpha \leq \frac{1}{L}$. Тогда:

$$f(x_T) - f^* \leq (1 - \alpha\mu)^T (f(x^0) - f^*).$$

$$x_{T+1} = x_T - \alpha \nabla f(x_T) \Rightarrow x_{T+1} - x_T = -\alpha \nabla f(x_T)$$

$$L\text{-гладкость: } \forall x, y : f(y) \leq f(x) + \langle \nabla f(x), y - x \rangle + \frac{L}{2} \|y - x\|^2$$

$$y := x_{T+1}, x := x_T \Rightarrow f(x_{T+1}) \leq f(x_T) + \langle \nabla f(x_T), -\alpha \nabla f(x_T) \rangle + \frac{L}{2} \alpha^2 \|\nabla f(x_T)\|^2$$

$$f(x_{T+1}) \leq f(x_T) - \alpha \|\nabla f(x_T)\|^2 + \frac{L}{2} \alpha^2 \|\nabla f(x_T)\|^2$$

$$f(x_{T+1}) \leq f(x_T) - \frac{\alpha}{2} (2 - L\alpha) \|\nabla f(x_T)\|^2 \leq f(x^T) - \frac{\alpha}{2} \|\nabla f(x^T)\|^2 \quad (L\alpha \leq 1)$$

$$\text{Условие PL: } \|\nabla f(x_T)\|^2 \geq 2\mu(f(x_T) - f^*)$$

$$f(x_{T+1}) - f^* \leq f(x_T) - f^* - \alpha\mu(f(x_T) - f^*) = (1 - \alpha\mu)(f(x_T) - f^*) = \dots = (1 - \alpha\mu)^{T+1}(f(x_0) - f^*)$$

То есть $f(x_T) - f^* \leq (1 - \alpha\mu)^T (f(x^0) - f^*)$ и характер сходимости - линейный.

14. Теорема сходимости градиентного спуска для сильно выпуклых квадратичных функций. Оптимальные гиперпараметры.

i

$$f(x) \rightarrow \min_{x \in \mathbb{R}^d}$$

$$f(x) = \frac{1}{2} x^T A x - b^T x + c, \quad A \in \mathbb{S}_{++}$$

Тогда градиентный спуск с шагом $\alpha = \frac{2}{\mu+L}$ сходится линейно с показателем $\frac{L-\mu}{L+\mu}$

$$f(x_k) - f^* \leq \left(\frac{L-\mu}{L+\mu} \right)^k (f(x_0) - f^*).$$

$$\nabla f(x) = Ax - b \stackrel{\nabla f(x^*)=0}{\Rightarrow} Ax^* = b$$

Тогда шаг градиентного спуска имеет вид

$$x_{k+1} = x_k - \alpha(Ax - b)$$

Найдем α^* . Воспользуемся $A = Q\Lambda Q^T$, где $\Lambda = \text{diag}\{\lambda_1, \dots, \lambda_n\}$, $Q = [q_1, \dots, q_n]$, λ_i, q_i - собственное значение и собственный вектор соответственно.

$$x_{k+1} = (I - \alpha A)x_k + \alpha A x^* \mid -x^*$$

$$\begin{aligned}x_{k+1} - x^* &= (I - \alpha A)(x_k - x^*) \\x_{k+1} - x^* &= (I - \alpha Q \Lambda Q^T)(x_k - x^*) | \cdot Q^T \\Q^T(x_{k+1} - x^*) &= (Q^T - \alpha \Lambda Q^T)(x_k - x^*) = (I - \alpha \Lambda)Q^T(x_k - x^*)\end{aligned}$$

Замена: $\tilde{x} = Q^T(x - x^*) \Rightarrow \tilde{x}_{k+1} = (I - \alpha \Lambda)\tilde{x}_k \Leftrightarrow \tilde{x}_i^{(k+1)} = (1 - \alpha \lambda_i)\tilde{x}_i^{(k)} i = \overline{1, d}$

$$\lambda_{\min} = \mu, \quad \lambda_{\max} = L$$

Сходимость есть $\Leftrightarrow \max_i |1 - \alpha \lambda_i| < 1$

$$\left\{ \begin{array}{l} |1 - \lambda \mu| < 1 \Rightarrow 1 - \lambda \mu < 1 \Rightarrow \alpha > 0 \\ \alpha \mu - 1 < 1 \Rightarrow \alpha < \frac{2}{\mu} \\ |1 - \alpha L| < 1 \Rightarrow 1 - \alpha L < 1 \Rightarrow \alpha > 0 \\ \alpha L - 1 < 1 \Rightarrow \alpha < \frac{2}{L} \end{array} \right\} \Rightarrow \alpha < \frac{2}{L}$$

Радиус сходимости $\rho = \max(|1 - \alpha \mu|, |1 - \alpha L|)$ и $\rho \rightarrow \min \Leftrightarrow \alpha^* L - 1 = 1 - \alpha^* \mu \Rightarrow \alpha^* = \frac{2}{\mu + L}$ и $\rho^* = \frac{L - \mu}{L + \mu}$

Итого получаем, что для градиентного спуска выполняется $f(x_k) - f^* \leq \left(1 - \frac{\mu}{\mu + L}\right)^k (f(x_0) - f^*)$.

15. Теорема сходимости субградиентного метода для выпуклых функций. Сходимость метода для разных стратегий выбора шага: постоянный размер шага $\alpha_k = \alpha$; Обратный квадратный корень $\frac{R}{G\sqrt{k}}$; Обратный $\frac{1}{k}$; Размер шага Поляка: $\alpha_k = \frac{f(x^k) - f^*}{\|g_k\|_2^2}$.

i Пусть f - выпуклая функция G -Липшица. Для фиксированного размера шага $\alpha = \frac{\|x_0 - x^*\|_2}{G} \sqrt{\frac{1}{K}}$, субградиентный метод достигает

$$f(\bar{x}) - f^* \leq \frac{G\|x_0 - x^*\|_2}{\sqrt{K}} \quad \bar{x} = \frac{1}{K} \sum_{k=0}^{K-1} x_i$$

Для фиксированных стратегий выбора шага можно получить оценки сходимости вида $\mathcal{O}\left(\frac{1}{\sqrt{k}}\right)$

Рассматривается задача

$$f(x) \rightarrow \min_{x \in \mathbb{R}^d},$$

где f - выпуклая. Рассмотрим сходимости метода субградиента для разных шагов.

$$x_{k+1} = x_k - \alpha_k g_k, \quad f(x_k) \geq f(x_0) + g_k^T(x_k - x_0) \Rightarrow g_k^T(x_k - x^*) \geq f(x_k) - f(x^*) \quad (1)$$

$$\|x_{k+1} - x^*\|^2 = \|x_k - x^* - \alpha_k g_k\|^2 = \|x_k - x^*\|^2 + \alpha_k^2 \|g_k\|^2 - 2\alpha_k g_k^T(x_k - x^*)$$

$$2\alpha_k g_k^T(x_k - x^*) = \|x_k - x^*\|^2 + \alpha_k^2 \|g_k\|^2 - \|x_{k+1} - x^*\|^2$$

$$\sum_{i=1}^{k-1} 2\alpha_i g_i^T(x_i - x^*) = \|x_0 - x^*\|^2 - \|x_k - x^*\|^2 + \sum_{i=0}^{k-1} \alpha_i^2 \|g_i\|^2 \leq \|x_0 - x^*\|^2 + \sum_{i=0}^{k-1} \alpha_i \|g_i\|^2 \leq R^2 + G \sum_{i=0}^{k-1} \alpha_i^2$$

Здесь мы предположили, что $\exists G : \|g_k\|^2 \leq G^2 \forall k$. Обозначим $f_k^{\text{best}} = \min_{i \leq k} f(x_i)$. Тогда

$$(1) \Rightarrow f_k^{\text{best}} - f^* \leq \frac{R^2 + G^2 \sum_{i=0}^{k-1} \alpha_i^2}{2 \sum_{i=0}^{k-1} \alpha_i}$$

1. $\alpha_k = \alpha$.

$$f_k^{\text{best}} - f^* \leq \frac{R^2 + G^2 k \alpha^2}{2 k \alpha} = \frac{R^2}{2 k \alpha} + G^2 \alpha \xrightarrow[k \rightarrow \infty]{} G^2 \alpha$$

То есть при таком шаге о сходимости сказать ничего нельзя.

2. $\alpha_k = \frac{R}{G\sqrt{k}}$.

$$f_k^{\text{best}} - f^* \leq \frac{R^2 + G^2 \sum_{i=1}^k \frac{R^2}{G^2 i}}{2 \sum_{i=1}^k \frac{R}{G\sqrt{i}}} = \frac{RG}{2} \frac{1 + \sum_{i=1}^k \frac{1}{i}}{\sum_{i=1}^k \frac{1}{\sqrt{i}}} \underset{k \gg 1}{\approx} \frac{RG}{2} \frac{1 + \int_1^k \frac{dx}{x}}{\int_1^k \frac{dx}{\sqrt{x}}} = \frac{RG}{2} \frac{1 + \ln k}{\sqrt{k} - 1} \xrightarrow[k \rightarrow \infty]{} 0.$$

То есть при таком шаге есть сублинейная сходимость.

3. $\alpha_k = \frac{1}{k}$.

$$f_k^{\text{best}} - f^* \leq \frac{R^2 + G^2 \sum_{i=1}^k \frac{1}{i^2}}{2 \sum_{i=1}^k \frac{1}{i}} \underset{k \gg 1}{\approx} \frac{R^2 + G^2 \frac{\pi}{6}}{2 \ln k} \xrightarrow[k \rightarrow \infty]{} 0.$$

То есть при таком шаге есть сублинейная сходимость.

4. $\alpha_k = \frac{f(x^k) - f^*}{\|g_k\|_2^2}$.

$$\|x_{i+1} - x^*\|_2^2 \leq \|x_i - x^*\|_2^2 - 2 \frac{f(x_i) - f^*}{\|g_i\|_2^2} (f(x_i) - f^*) + \frac{(f(x_i) - f^*)^2}{\|g_i\|_2^4} \|g_i\|_2^2$$

$$\|x_{i+1} - x^*\|_2^2 - \|x_i - x^*\|_2^2 \leq -\frac{(f(x_i) - f^*)^2}{\|g_i\|_2^2}$$

$$\frac{(f(x_i) - f^*)^2}{\|g_i\|_2^2} \leq \|x_i - x^*\|_2^2 - \|x_{i+1} - x^*\|_2^2 \Rightarrow (f_k^{\text{best}} - f^*)^2 \frac{k}{G^2} \leq R^2$$

$$f_k^{\text{best}} - f^* \leq R \sqrt{\frac{G}{k}} \xrightarrow[k \rightarrow \infty]{} 0.$$

То есть при таком шаге есть сублинейная сходимость.

16. Теорема о сходимости метода тяжелого шарика для сильно выпуклой квадратичной задачи.

i Рассматривается задача

$$f(x) \rightarrow \min_{x \in \mathbb{R}^d}$$

$$f(x) = \frac{1}{2}x^T Ax - b^T x + c, \quad A \in \mathbb{S}_{++} \Rightarrow \nabla f(x) = Ax - b \stackrel{\nabla f(x^*)=0}{\Rightarrow} Ax^* = b.$$

Не умаляя общности, $c = 0$, так как решение от него не зависит.

Метод тяжелога шарика имеет вид:

$$x_{k+1} = x_k - \alpha \nabla f(x_k) + \beta(x_k - x_{k-1})$$

Тогда скорость сходимости (ρ) не зависит от шага (при допустимых его значениях), $\rho \sim \sqrt{\beta^*}$, где β^* - оптимальный гиперпараметр и выполняется

$$\|x_k - x^*\| \leq \left(\frac{\sqrt{\kappa} - 1}{\sqrt{\kappa} + 1} \right)^k \|x_0 - x^*\|.$$

Воспользуемся $A = Q\Lambda Q^T$, где $\Lambda = \text{diag}\{\lambda_1, \dots, \lambda_n\}$, $Q = [q_1, \dots, q_n]$, λ_i , q_i - собственное значение и собственный вектор соответственно.

$$\begin{aligned} \tilde{x} &= Q^T(x - x^*), \\ f(\tilde{x}) &= \frac{1}{2}(Q\tilde{x} + x^*)^T A(Q\tilde{x} + x^*) - b^T(Q\tilde{x} + x^*) \\ &= \frac{1}{2}\tilde{x}^T A Q^T A Q \tilde{x} + (x^*)^T A Q^T A Q \tilde{x} + \frac{1}{2}(x^*)^T A (x^*)^T - b^T Q \tilde{x} - b^T x^* \\ &= \frac{1}{2}\tilde{x}^T \Lambda \tilde{x} \\ \nabla f(\tilde{x}) &= \Lambda \tilde{x} \end{aligned}$$

Тогда можем переписать правило обновления:

$$\begin{cases} \tilde{x}_{k+1} = (I - \alpha \Lambda + \beta I)\tilde{x}_k - \beta \tilde{x}_{k-1} \\ \tilde{x}_k = \tilde{x}_k \end{cases}$$

$$\text{Рассмотрим } \tilde{z}_k = \begin{pmatrix} \tilde{x}_{k+1} \\ \tilde{x}_k \end{pmatrix}$$

Тогда правило обновления имеет вид:

$$\begin{aligned} \tilde{z}_{k+1} &= M \tilde{z}_k \\ M &= \begin{pmatrix} I - \alpha \Lambda + \beta I & -\beta I \\ I & 0 \end{pmatrix} \in \mathbb{R}^{d \times d} \end{aligned}$$

Сделаем reshape:

$$\begin{pmatrix} \tilde{x}_k^{(1)} \\ \tilde{x}_k^{(2)} \\ \vdots \\ \tilde{x}_k^{(d)} \\ \tilde{x}_{k+1}^{(1)} \\ \tilde{x}_{k+1}^{(2)} \\ \vdots \\ \tilde{x}_{k+1}^{(d)} \end{pmatrix} \rightarrow \begin{pmatrix} \tilde{x}_k^{(1)} \\ \tilde{x}_{k+1}^{(1)} \\ \tilde{x}_k^{(2)} \\ \tilde{x}_{k+1}^{(2)} \\ \vdots \\ \tilde{x}_k^{(d)} \\ \tilde{x}_{k+1}^{(d)} \end{pmatrix} \quad M = \begin{pmatrix} M_1 & & & & \\ & M_2 & & & \\ & & M_3 & & \\ & & & \ddots & \\ & & & & M_d \end{pmatrix}$$

Для i -й координаты:

$$M_i = \begin{pmatrix} 1 - \alpha\lambda_i + \beta & -\beta \\ 1 & 0 \end{pmatrix}$$

Метод будет сходиться, если $\rho(M) < 1$ и оптимальные параметры могут быть подобраны через оптимизацию спектрального радиуса:

$$\alpha^*, \beta^* = \arg \min_{\alpha, \beta} \max_{i=1, d} \rho(M_i)$$

$$\alpha^* = \frac{4}{(\sqrt{L} + \sqrt{\mu})^2}, \quad \beta^* = \left(\frac{\sqrt{L} - \sqrt{\mu}}{\sqrt{L} + \sqrt{\mu}} \right)^2$$

Собственные значения M_i имеют вид:

$$\lambda_1^M, \lambda_2^M = \frac{1 + \beta - \alpha\lambda_i \pm \sqrt{(1 + \beta - \alpha\lambda_i)^2 - 4\beta}}{2} \quad (1)$$

При (α^*, β^*) собственные значения являются комплексно сопряженными $\Rightarrow (1 + \beta - \alpha\lambda_i)^2 - 4\beta \leq 0 \Rightarrow \beta \geq (1 + \sqrt{\alpha\lambda_i})^2$

$$(\alpha^*, \beta^*) \rightarrow (1) \Rightarrow |\lambda_1^M| = \frac{L - \mu}{(\sqrt{L} + \sqrt{\mu})^2} = \frac{\sqrt{L} - \sqrt{\mu}}{\sqrt{L} + \sqrt{\mu}} = \sqrt{\beta^*}$$

То есть скорость сходимости не зависит от α^* и равна $\sqrt{\beta^*}$. Тогда получаем оценку:

$$\|\tilde{z}_k - \tilde{z}^*\| \leq \sqrt{\beta^*} \|\tilde{z}_{k-1} - \tilde{z}^*\| \Rightarrow \|\tilde{z}_k - \tilde{z}^*\| \leq (\sqrt{\beta})^k \|\tilde{z}_0 - \tilde{z}^*\|$$

Итого получаем оценку

$$\|x_k - x^*\| \leq \|z_k - z^*\| \leq \left(\frac{\sqrt{\kappa} - 1}{\sqrt{\kappa} + 1} \right)^k \|x_0 - x^*\|$$

17. Теорема о сходимости метода проекции градиента для выпуклой гладкой функции.

i Рассматриваем задачу

$$f(x) \rightarrow \min_{x \in S},$$

где f - выпуклая и L -гладкая. Пусть $S \subseteq \mathbb{R}^n$ замкнутое выпуклое множество. Тогда метод проекции градиента с шагом $\alpha = \frac{1}{L}$:

$$x_{k+1} = \text{proj}_S \left(x_k - \frac{1}{L} \nabla f(x_k) \right)$$

сходится со скоростью $\mathcal{O}\left(\frac{1}{T}\right)$ и $\forall T$ выполняется неравенство:

$$f(x_T) - f^* \leq \frac{L\|x_0 - x^*\|_2^2}{2T} = \frac{LR^2}{2T}.$$

Докажем лемму, предполагая, что $y_k = x_k - \frac{1}{L} \nabla f(x_k)$ и используя равенство

$$2x^T y = \|x\|^2 + \|y\|^2 - \|x - y\|^2. \quad (1)$$

L-гладкость:

$$\begin{aligned}
 f(x_{k+1}) &\leq f(x_k) + \langle \nabla f(x_k), x_{k+1} - x_k \rangle + \frac{L}{2} \|x_{k+1} - x_k\|_2^2 = \\
 &= f(x_k) - L \langle y_k - x_k, x_{k+1} - x_k \rangle + \frac{L}{2} \|x_{k+1} - x_k\|_2^2 = \\
 &= f(x_k) - \frac{L}{2} (\|y_k - x_k\|^2 + \|x_{k+1} - x_k\|^2 - \|y_k - x_{k+1}\|^2) + \frac{L}{2} \|x_{k+1} - x_k\|_2^2 = \\
 &= f(x_k) - \frac{1}{2L} \|\nabla f(x_k)\|^2 + \frac{L}{2} \|y_k - x_{k+1}\|^2. \\
 (1) \Rightarrow \quad \left\langle \frac{1}{L} \nabla f(x_k), x_k - x^* \right\rangle &= \frac{1}{2} \left(\frac{1}{L^2} \|\nabla f(x_k)\|^2 + \|x_k - x^*\|^2 - \|x_k - x^* - \frac{1}{L} \nabla f(x_k)\|^2 \right) \\
 \langle \nabla f(x_k), x_k - x^* \rangle &= \frac{L}{2} \left(\frac{1}{L^2} \|\nabla f(x_k)\|^2 + \|x_k - x^*\|^2 - \|x_k - x^* - \frac{1}{L} \nabla f(x_k)\|^2 \right)
 \end{aligned}$$

Воспользуемся свойством проекции:

$$\begin{aligned}
 \|x - \text{proj}_S(y)\|^2 + \|y - \text{proj}_S(y)\|^2 &\leq \|x - y\|^2 \quad \forall x \in S \\
 x := x^*, y := y_k \Rightarrow \|x^* - \text{proj}_S(y_k)\|^2 + \|y_k - \text{proj}_S(y_k)\|^2 &\leq \|x^* - y_k\|^2 \\
 \|y_k - x^*\|^2 &\geq \|x^* - x_{k+1}\|^2 + \|y_k - x_{k+1}\|^2
 \end{aligned}$$

Выпуклость:

$$f(x_k) - f^* \leq \langle \nabla f(x_k), x_k - x^* \rangle \leq \frac{L}{2} \left(\frac{1}{L^2} \|\nabla f(x_k)\|^2 + \|x_k - x^*\|^2 - \|x_{k+1} - x^*\|^2 - \|y_k - x_{k+1}\|^2 \right)$$

Суммируем от 0 до $T - 1$:

$$\begin{aligned}
 \sum_{k=0}^{T-1} [f(x_k) - f^*] &\leq \sum_{k=0}^{T-1} \left[f(x_k) - f(x_{k+1}) + \frac{L}{2} \|y_k - x_{k+1}\|^2 \right] + \frac{L}{2} \|x_0 - x^*\|^2 - \frac{L}{2} \sum_{k=0}^{T-1} \|y_k - x_{k+1}\|^2 \leq \\
 &\leq f(x_0) - f(x_T) + \frac{L}{2} \sum_{k=0}^{T-1} \|y_k - x_{k+1}\|^2 + \frac{L}{2} \|x_0 - x^*\|^2 - \frac{L}{2} \sum_{k=0}^{T-1} \|y_k - x_{k+1}\|^2 \leq \\
 &\leq f(x_0) - f(x_T) + \frac{L}{2} \|x_0 - x^*\|^2 \\
 \sum_{k=0}^{T-1} f(x_k) - T f^* &\leq f(x_0) - f(x_T) + \frac{L}{2} \|x_0 - x^*\|^2 \\
 \sum_{k=1}^T [f(x_k) - f^*] &\leq \frac{L}{2} \|x_0 - x^*\|^2
 \end{aligned}$$

Заметим, что $f(x_T) \leq f(x_i) \forall i = \overline{1, T-1} \Rightarrow f(x_T) \leq \frac{1}{T} \sum_{k=0}^{T-1} f(x_{k+1})$ Итого имеем:

$$f(x_T) - f(x^*) \leq \frac{LR^2}{2T}$$

То есть сходимость сублинейная.

18. Теорема о сходимости метода проекции градиента для сильно выпуклой гладкой функции.

i Рассматриваем задачу

$$f(x) \rightarrow \min_{x \in S},$$

где f - μ -сильно выпуклая и L -гладкая. Пусть $S \subseteq \mathbb{R}^n$ замкнутое выпуклое множество. Тогда метод проекции градиента с постоянным шагом $\alpha \leq \frac{1}{L}$:

$$x_{k+1} = \text{proj}_S(x_k - \alpha \nabla f(x_k))$$

сходится со линейно и $\forall T$ выполняется неравенство:

$$f(x_T) - f^* \leq (1 - \alpha\mu)^T (f(x_0) - f^*).$$

Повторяет доказательство теоремы 13, с заменой оператора prox на proj.

19. Теорема о сходимости метода Франк-Вульфа для выпуклой гладкой функции.

i Рассматриваем задачу

$$f(x) \rightarrow \min_{x \in S},$$

где f -выпуклая и L -гладкая. Метод Франк-Вульфа имеет вид:

$$\begin{cases} x_{k+1} = \gamma_k x_k + (1 - \gamma_k) s_k \\ s_k = \arg \min_{x \in S} f_{x_k}^I(x) = \arg \min_{x \in S} \langle \nabla f(x_k), x \rangle \end{cases},$$

где $f_{x_k}^I(x)$ - тейлоровская аппроксимация 1-го порядка в точке x_k . И для $\gamma_k = \frac{k-1}{k+1}$ выполняется

$$f(x_k) - f(x^*) \leq \frac{2LR^2}{k+1},$$

где $R = \max_{x,y \in S} \|x - y\|$. То есть имеет место сублинейная сходимость.

L -гладкость:

$$f(x) - f(y) - \langle \nabla f(y), x - y \rangle \leq \frac{L}{2} \|x - y\|^2, \quad \forall x, y \in S$$

$$\begin{aligned} f(x_{k+1}) - f(x_k) &\leq \langle \nabla f(x_k), x_{k+1} - x_k \rangle + \frac{L}{2} \|x_{k+1} - x_k\|^2 \\ &= (1 - \gamma_k) \langle \nabla f(x_k), s_k - x_k \rangle + \frac{L(1 - \gamma_k)^2}{2} \|s_k - x_k\|^2 \end{aligned}$$

Выпуклость:

$$f(x) - f(y) - \langle \nabla f(y), x - y \rangle \geq 0 \quad \forall x, y \in S \Rightarrow x := x^*, y := x_k \Rightarrow \langle \nabla f(x_k), x^* - x_k \rangle \leq f(x^*) - f(x_k)$$

$$f(x_{k+1}) - f(x_k) \leq (1 - \gamma_k) \langle \nabla f(x_k), x^* - x_k \rangle + \frac{L(1 - \gamma_k)^2}{2} R^2 \leq (1 - \gamma_k) (f(x^*) - f(x_k)) + (1 - \gamma_k)^2 \frac{LR^2}{2}$$

$$f(x_{k+1}) - f(x^*) \leq \gamma_k (f(x_k) - f(x^*)) + (1 - \gamma_k)^2 \frac{LR^2}{2}$$

Обозначим $\delta_k = \frac{f(x_k) - f(x^*)}{LR^2}$. Тогда неравенство перепишется в виде

$$\delta_{k+1} \leq \gamma_k \delta_k + \frac{(1 - \gamma_k)^2}{2} = \frac{k-1}{k+1} \delta_k + \frac{2}{(k+1)^2}.$$

Начиная с неравенства $\delta_2 \leq \frac{1}{2}$, применением индукции по k получаем желаемый результат.

20. Доказательство сходимости метода сопряженных градиентов и вывод формулы.

i Рассматриваем задачу

$$f(x) = \frac{1}{2}x^T Ax - b^T x + c \rightarrow \min_{x \in \mathbb{R}^d}$$

Метод сопряженных градиентов:

- $r_0 := b - Ax_0$
- if r_0 sufficiently small, then return x_0 as result
- $d_0 := r_0$
- $k := 0$
- while r_{k+1} is not sufficiently small :
 - $\alpha_k := \frac{r_k^T r_k}{d_k^T A d_k}$
 - $x_{k+1} := x_k + \alpha_k d_k$
 - $r_{k+1} := r_k - \alpha_k A d_k$
 - $\beta_k := \frac{r_{k+1}^T r_{k+1}}{r_k^T r_k}$
 - $d_{k+1} := r_{k+1} + \beta_k d_k$
 - $k := k + 1$
- return x_{k+1} as result.

Лемма: Пусть d_1, d_2, \dots, d_m попарно A —ортогональные вектора. Тогда они линейно независимы. $A \in S_{++}^n$

Доказательство: Пусть они ЛНЗ, т.е. $\sum_{i=0}^n \alpha_i d_i = 0$. Домножим слева на $d_j^T A$:

$$0 = d_j^T A \sum_{i=0}^n \alpha_i d_i = \sum_{i=0}^n \alpha_i d_j^T A d_i = \alpha_j d_j^T A d_j + 0 + \dots + 0 \Rightarrow \alpha_j = 0$$

В силу A -ортогональности. Повторим рассуждение $\forall j = \overline{1, n} \Rightarrow$ противоречие \Rightarrow ЛНЗ.

Справка: $r_k = b - Ax^k$ —невязка, $e_k = x^k - x^*$ —ошибка, $r_k = Ae_k$.

Лемма: Метод сопряженных градиентов сходится за n шагов, т.е. $e_0 = x_0 - x^* = \sum_{i=0}^{n-1} \delta_i d_i$

Доказательство:

Пусть есть n A -ортогональных векторов: d_0, \dots, d_{n-1} . $x_{k+1} = x_0 + \sum_{i=0}^k \alpha_i d_i$. α_i подбирается

LineSearch. $\alpha_i = \frac{d_i^T r_i}{d_i^T A d_i}$ и необходимо показать, что $\delta_i = -\alpha_i$, $x_0 + \sum_{i=0}^{n-1} \alpha_i d_i = x^*$.

1. Фиксируем индекс k . Домножим ошибку e_0 на $d_k^T A$:

$$d_k^T A e_0 = \sum_{i=0}^{n-1} \delta_i d_k^T A d_i \stackrel{\perp_A}{=} \delta_k d_k^T A d_k$$

Подставим умный ноль $\sum_{i=0}^{k-1} \alpha_i d_k^T A d_i = 0$ (в силу предыдущей леммы):

$$d_k^T A (e_0 + \sum_{i=0}^{k-1} \alpha_i d_i) = \delta_k d_k^T A d_k$$

$$e_k = e_0 + \sum_{i=0}^{k-1} \alpha_i d_i \Rightarrow \delta_k = \frac{d_k^T A e_k}{d_k^T A d_k} = -\frac{d_k^T r_k}{d_k^T A d_k} = -\alpha_k \text{ ч.т.д.}$$

Лемма:

- В методе сопряженных градиентов мы рассматриваем ортогонализацию Грамма-Шмидта для невязок, т.е. $u_i = r_i$. Формула Грамма-Шмидта: $d_i = u_i + \sum_{j=0}^{i-1} \beta_{ij} d_j$, $\beta_{ij} = -\frac{u_i^T A d_j}{d_j^T A d_j}$.
- Рассмотрим ошибку на i -ой итерации:

$$e_i = e_0 + \sum_{j=0}^{i-1} \alpha_j d_j = \left\{ e_0 = -\sum_{j=0}^{n-1} \alpha_j d_j \right\} = -\sum_{j=0}^{n-1} \alpha_j d_j + \sum_{j=0}^{i-1} \alpha_j d_j = -\sum_{j=i}^{n-1} \alpha_j d_j$$

Теперь зафиксируем индекс k : $e_k = -\sum_{j=k}^{n-1} \alpha_j d_j$ и для некоторого l помножим e_k на $d_l^T A$:

$$d_l^T A e_k = -\sum_{j=k}^{n-1} \alpha_j d_l^T A d_j$$

Если $l < k$, то $d_l^T A d_j = 0 \Rightarrow d_l^T r_k = 0$. А значит r_k перпендикулярна всем предыдущим направлениям d_k .

- Теперь покажем, что r_k перпендикулярны друг другу: Пользуемся формулой Грамма-Шмидта:

$$r_k^T d_i = r_k^T (u_i + \sum_{j=0}^{i-1} \beta_{ij} d_j) = r_k^T u_i + \sum_{j=0}^{i-1} \beta_{ij} r_k^T d_j$$

По предыдущему пункту, если $i < k$: $r_k^T d_i = r_k^T u_i = r_k^T r_i = 0$. А значит r_k ортогонально всем предыдущем r_i . Если же $i = k$, то $r_k^T d_k = r_k^T u_k = r_k^T r_k$

- Посчитаем теперь коэффициенты β_{ij} : $r_{i+1} = -Ae_{i+1} = -A(e_i + \alpha_i d_i) = -Ae_i - \alpha_i Ad_i = r_i - \alpha_i Ad_i$.

Оказывается, что $\beta_{ij} = -\frac{u_i^T A d_j}{d_j^T A d_j} = -\frac{r_i^T A d_j}{d_j^T A d_j}$ почти всегда 0, кроме случаев соседних i, j . Для доказательства рассмотрим:

$$\langle r_i, r_{j+1} \rangle = \langle r_i, r_j - \alpha_j Ad_j \rangle = \langle r_i, r_j \rangle - \alpha_j \langle r_i, Ad_j \rangle$$

$$\alpha_j \langle r_i, Ad_j \rangle = \langle r_i, r_j \rangle - \langle r_i, r_{j+1} \rangle$$

Если $i = j$: $\alpha_i \langle r_i, Ad_i \rangle = \langle r_i, r_i \rangle - \langle r_i, r_{i+1} \rangle = \langle r_i, r_i \rangle$ по предыдущим пунктам.

Если $i = j + 1$ или $i = j - 1$: $\alpha_i \langle r_i, Ad_i \rangle = \langle r_i, r_i \rangle$

Если $i \neq j + 1$ или $i \neq j - 1$: $\langle r_i, Ad_i \rangle = 0$ из ортогональности невязок.

5. Осталось посчитать:

$$\begin{aligned}\beta_{ij} &= -\frac{r_i^T Ad_j}{d_j^T Ad_j} = \frac{1}{\alpha_j} \frac{r_i^T r_i}{d_j^T Ad_j} = \left\{ \alpha_j = \frac{d_j^T r_j}{d_j^T Ad_j} \right\} = \\ &= \frac{d_j^T Ad_j}{d_j^T r_j} \cdot \frac{r_i^T r_j}{d_j^T Ad_j} = \frac{r_i^T r_i}{d_j^T r_j} = \frac{r_i^T r_j}{d_j^T r_i} = \frac{r_i^T r_j}{d_j^T r_j} = \frac{r_i^T r_j}{r_j^T r_j} = [i = j - 1] = \frac{r_i^T r_j}{r_{i-1}^T r_{i-1}}\end{aligned}$$

21. Теорема сходимости метода Ньютона для сильно выпуклых функций с липшицевым гессианом.

i Рассматриваем задачу

$$f(x) \rightarrow \min_{x \in \mathbb{R}^d},$$

где f - μ -сильно выпуклая дважды непрерывно дифференцируемая функция на \mathbb{R}^d , причем для второй производной которой выполняются неравенства: $\mu I_d \preceq \nabla^2 f(x) \preceq L I_d$. Тогда метод Ньютона с постоянным шагом локально сходится к решению задачи со сверхлинейной скоростью. Если, кроме того, гессиан M -Липшицев, то этот метод локально сходится к x^* с квадратичной скоростью.

1. Воспользуемся формулой Ньютона-Лейбница

$$\nabla f(x_k) - \nabla f(x^*) = \int_0^1 \nabla^2 f(x^* + \tau(x_k - x^*)) (x_k - x^*) d\tau$$

2. Рассмотрим расстояние до решения:

$$\begin{aligned}x_{k+1} - x^* &= x_k - x^* - [\nabla^2 f(x_k)]^{-1} \nabla f(x_k) = \\ &= x_k - x^* - [\nabla^2 f(x_k)]^{-1} \int_0^1 \nabla^2 f(x^* + \tau(x_k - x^*)) (x_k - x^*) d\tau = \\ &= \left(I - [\nabla^2 f(x_k)]^{-1} \int_0^1 \nabla^2 f(x^* + \tau(x_k - x^*)) d\tau \right) (x_k - x^*) = \\ &= [\nabla^2 f(x_k)]^{-1} \left(\nabla^2 f(x_k) - \int_0^1 \nabla^2 f(x^* + \tau(x_k - x^*)) d\tau \right) (x_k - x^*) = \\ &= [\nabla^2 f(x_k)]^{-1} \left(\int_0^1 (\nabla^2 f(x_k) - \nabla^2 f(x^* + \tau(x_k - x^*))) d\tau \right) (x_k - x^*) = \\ &= [\nabla^2 f(x_k)]^{-1} G_k (x_k - x^*)\end{aligned}$$

3. Замена:

$$G_k = \int_0^1 (\nabla^2 f(x_k) - \nabla^2 f(x^* + \tau(x_k - x^*)) d\tau.$$

$$\|G_k\| = \left\| \int_0^1 (\nabla^2 f(x_k) - \nabla^2 f(x^* + \tau(x_k - x^*)) d\tau \right\| \leq \int_0^1 \|\nabla^2 f(x_k) - \nabla^2 f(x^* + \tau(x_k - x^*))\| d\tau \leq$$

$$(\text{Гессиан - } M\text{-Липшицев}) \leq \int_0^1 M \|x_k - x^* - \tau(x_k - x^*)\| d\tau = \int_0^1 M \|x_k - x^*\| (1 - \tau) d\tau = \frac{r_k}{2} M,$$

где $r_k = \|x_k - x^*\|$.

4. Итак:

$$r_{k+1} \leq \left\| [\nabla^2 f(x_k)]^{-1} \right\| \frac{r_k}{2} M r_k$$

5. Из-за непрерывности и симметрии Липшица Гессиана:

$$\nabla^2 f(x_k) - \nabla^2 f(x^*) \succcurlyeq -M r_k I_n$$

$$\nabla^2 f(x_k) \succcurlyeq \nabla^2 f(x^*) - M r_k I_n$$

$$\nabla^2 f(x_k) \succcurlyeq \mu I_n - M r_k I_n$$

$$\nabla^2 f(x_k) \succcurlyeq (\mu - M r_k) I_n$$

$$\left\| [\nabla^2 f(x_k)]^{-1} \right\| \leq (\mu - M r_k)^{-1}$$

$$r_{k+1} \leq \frac{r_k^2 M}{2(\mu - M r_k)}$$

6. Условие сходимости: $r_{k+1} < r_k \Rightarrow r_k < \frac{2\mu}{3M}$, то есть метод Ньютона для функции с липшицевым положительно определенным гессианом вблизи x^* ($\|x_0 - x^*\| < \frac{2\mu}{3M}$) сходится квадратично к решению.

22. Вывод формул обновления оценок обратного гессиана и гессиана квазиньютоновских методов SR-1, DFP, BFGS.

$$\begin{aligned} x_{k+1} &= x_k + \alpha_k d_k, \\ B_k d_k &= -\nabla f(x_k) \\ B_k &= \nabla^2 f(x_k) \end{aligned}$$

То есть на каждой итерации необходимо вычислять гессиан и решать систему линейных уравнений.

В квазиньютоновских методах мы рассматриваем последовательность матриц B_k , сходящихся в каком-то смысле к настоящему значению обратного Гессиана в локальном оптимуме: $[\nabla^2 f(x^*)]^{-1}$.

Общая схема:

1. Решить $B_k d_k = -\nabla f(x_k)$
2. Обновить $x_{k+1} = x_k + \alpha_k d_k$ (уравнения секущих)

3. Вычислить B_{k+1} из B_k

Требования к B_{k+1} из ур-я секущих:

$$\begin{aligned}\nabla f(x_{k+1}) - \nabla f(x_k) &= B_{k+1}(x_{k+1} - x_k) = B_{k+1}d_k \\ \Delta y_k &= B_{k+1}d_k\end{aligned}$$

Также требуем:

- B_{k+1} - симметрична
- B_{k+1} "близка" к B_k
- $B_k \succ 0 \Rightarrow B_{k+1} \succ 0$

1. Symmetric Rank-One (Broyden) Update

Поробуем такой вид обновления:

$$B_{k+1} = B_k + auu^T$$

уравнение секущих $B_{k+1}d_k = \Delta y_k$ приводит к:

$$(au^T d_k)u = \Delta y_k - B_k d_k$$

Это справедливо только в том случае, если u кратно $\Delta y_k - B_k d_k$. Полагая $u = \Delta y_k - B_k d_k$, мы решаем приведенную выше задачу,

$$a = \frac{1}{(\Delta y_k - B_k d_k)^T d_k},$$

Что приводит к:

$$B_{k+1} = B_k + \frac{(\Delta y_k - B_k d_k)(\Delta y_k - B_k d_k)^T}{(\Delta y_k - B_k d_k)^T d_k}$$

Называется симметричный одноранговый апдейт (SR1) или метод Бройдена.

2. Davidon-Fletcher-Powell Update (DFP)

Как мы можем решить

$$B_{k+1}d_{k+1} = -\nabla f(x_{k+1}),$$

для того, чтобы сделать следующий шаг? В дополнение к приведению B_k к B_{k+1} , давайте будем приводить обратные, т.е. $C_k = B_k^{-1}$ to $C_{k+1} = (B_{k+1})^{-1}$.

Sherman-Morrison Formula: Формула Шермана-Моррисона утверждает:

$$(A + uv^T)^{-1} = A^{-1} - \frac{A^{-1}uv^TA^{-1}}{1 + v^TA^{-1}u}$$

$$C_{k+1} = C_k + auu^T + bvv^T.$$

Умножая на Δy_k и используя уравнение секущих $d_k = C_{k+1}\Delta y_k$ имеет:

$$d_k = C_k \Delta y_k + (au^T \Delta y_k)u + (bv^T \Delta y_k)v$$

Полагая $u = C_k \Delta y_k, v = d_k$ и решая для a, b получаем:

$$(1 + a \Delta y_k^T C \Delta y_k) C_k \Delta y_k + (bd_k^T \Delta y_k - 1) d_k \Leftrightarrow a = -\frac{1}{\Delta y_k^T C \Delta y_k}, b = \frac{1}{\Delta y_k^T d_k}$$

$$C_{k+1} = C_k - \frac{C_k \Delta y_k \Delta y_k^T C_k}{\Delta y_k^T C_k \Delta y_k} + \frac{d_k d_k^T}{\Delta y_k^T d_k}$$

Woodbury Formula Application Формула показывает:

$$B_{k+1} = \left(I - \frac{\Delta y_k d_k^T}{\Delta y_k^T d_k} \right) B_k \left(I - \frac{d_k \Delta y_k^T}{\Delta y_k^T d_k} \right) + \frac{\Delta y_k \Delta y_k^T}{\Delta y_k^T d_k}$$

Это обновление Davidon-Fletcher-Powell (DFP). Также дешево: $\mathcal{O}(n^2)$, сохраняет положительную определенность. Не так популярно, как BFGS.

3. Broyden-Fletcher-Goldfarb-Shanno update

Давайте теперь попробуем обновление второго ранга:

$$B_{k+1} = B_k + auu^T + bvv^T.$$

Умножая на Δy_k и используя уравнение секущих $\Delta y_k = B_{k+1} d_k$ имеем:

$$\Delta y_k - B_k d_k = (au^T d_k)u + (bv^T d_k)v$$

Полагая $u = \Delta y_k, v = B_k d_k$, и решая для a, b мы получаем:

$$(1 - a \Delta y_k^T d_k) \Delta y_k - (1 + bd_k^T B_k d_k) B_k d_k \Leftrightarrow a = \frac{1}{y_k^T d_k}, b = -\frac{1}{d_k^T B_k d_k}$$

$$B_{k+1} = B_k - \frac{B_k d_k d_k^T B_k}{d_k^T B_k d_k} + \frac{\Delta y_k \Delta y_k^T}{d_k^T \Delta y_k}$$

называется обновлением Брайдена-Флетчера-Гольдфарба-Шанно (BFGS).

23. Теорема о сходимости проксимального градиентного метода для выпуклой гладкой функции f .

i Рассматриваем задачу

$$\varphi(x) \rightarrow \min_{x \in \mathbb{R}^d} .$$

Причем $\varphi(x) = f(x) + r(x)$, и

- f -выпуклая и L -гладкая, $\text{dom } f = \mathbb{R}^n$
- r - выпуклая и $\text{prox}_{\alpha r}(x_k) = \arg \min_{x \in \mathbb{R}^n} [\alpha r(x) + \frac{1}{2}\|x - x_k\|^2]$ может быть вычислен

Тогда для проксимального метода с фиксированным шагом $\alpha = \frac{1}{L}$

$$x_{k+1} = \text{prox}_{\alpha,r} \left(x_k - \frac{1}{L} \nabla f(x_k) \right)$$

выполняется

$$\varphi(x_k) - \varphi^* \leq \frac{L\|x_0 - x^*\|^2}{2k},$$

то есть имеет место сублинейная сходимость.

1. Представим отображение градиента, обозначаемое как $G_\alpha(x)$, действует как "градиентоподобный объект":

$$\begin{aligned} x_{k+1} &= \text{prox}_{\alpha r}(x_k - \alpha \nabla f(x_k)) \\ x_{k+1} &= x_k - \alpha G_\alpha(x_k). \end{aligned}$$

где $G_\alpha(x)$ имеет вид:

$$G_\alpha(x) = \frac{1}{\alpha} (x - \text{prox}_{\alpha r}(x - \alpha \nabla f(x)))$$

$G_\alpha(x) = 0 \Leftrightarrow x = x^* \Rightarrow G_\alpha$ аналогичен ∇f .

2. L -гладкость:

$$f(x_{k+1}) \leq f(x_k) + \langle \nabla f(x_k), x_{k+1} - x_k \rangle + \frac{L}{2} \|x_{k+1} - x_k\|_2^2$$

Выпуклость:

$$\begin{aligned} f(x) &\geq f(x_k) + \langle \nabla f(x_k), x - x_k \rangle \\ f(x_{k+1}) &\leq f(x) - \langle \nabla f(x_k), x - x_k \rangle + \langle \nabla f(x_k), x_{k+1} - x_k \rangle + \frac{\alpha^2 L}{2} \|G_\alpha(x_k)\|_2^2 \leq \\ &\leq f(x) + \langle \nabla f(x_k), x_{k+1} - x \rangle + \frac{\alpha^2 L}{2} \|G_\alpha(x_k)\|_2^2 \quad (1) \end{aligned}$$

3. Воспользуемся свойством проксимального оператора:

$$\begin{aligned} x_{k+1} = \text{prox}_{\alpha r}(x_k - \alpha \nabla f(x_k)) &\Leftrightarrow x_k - \alpha \nabla f(x_k) - x_{k+1} \in \partial \alpha r(x_{k+1}) \\ x_k - x_{k+1} = \alpha G_\alpha(x_k) &\Rightarrow \alpha G_\alpha(x_k) - \alpha \nabla f(x_k) \in \partial \alpha r(x_{k+1}) \\ G_\alpha(x_k) - \nabla f(x_k) &\in \partial r(x_{k+1}) \end{aligned}$$

4. По определению субградиента:

$$r(x) \geq r(x_{k+1}) + \langle g, x - x_{k+1} \rangle, \quad g \in \partial r(x_{k+1})$$

$$r(x) \geq r(x_{k+1}) + \langle G_\alpha(x_k) - \nabla f(x_k), x - x_{k+1} \rangle$$

$$r(x) \geq r(x_{k+1}) + \langle G_\alpha(x_k), x - x_{k+1} \rangle - \langle \nabla f(x), x - x_{k+1} \rangle$$

$$\langle \nabla f(x), x_{k+1} - x \rangle \leq r(x) - r(x_{k+1}) - \langle G_\alpha(x_k), x - x_{k+1} \rangle$$

5. Подставляем полученные результаты в (1):

$$f(x_{k+1}) \leq f(x) + \langle \nabla f(x_k), x_{k+1} - x \rangle + \frac{\alpha^2 L}{2} \|G_\alpha(x_k)\|_2^2$$

$$f(x_{k+1}) \leq f(x) + r(x) - r(x_{k+1}) - \langle G_\alpha(x_k), x - x_{k+1} \rangle + \frac{\alpha^2 L}{2} \|G_\alpha(x_k)\|_2^2$$

$$f(x_{k+1}) + r(x_{k+1}) \leq f(x) + r(x) - \langle G_\alpha(x_k), x - x_k + \alpha G_\alpha(x_k) \rangle + \frac{\alpha^2 L}{2} \|G_\alpha(x_k)\|_2^2$$

6. Используя $\varphi(x) = f(x) + r(x)$ доказываем монотонное уменьшение итерации:

$$\varphi(x_{k+1}) \leq \varphi(x) - \langle G_\alpha(x_k), x - x_k \rangle - \langle G_\alpha(x_k), \alpha G_\alpha(x_k) \rangle + \frac{\alpha^2 L}{2} \|G_\alpha(x_k)\|_2^2$$

$$\varphi(x_{k+1}) \leq \varphi(x) + \langle G_\alpha(x_k), x_k - x \rangle + \frac{\alpha}{2} (\alpha L - 2) \|G_\alpha(x_k)\|_2^2$$

$$\left(\alpha \leq \frac{1}{L} \Rightarrow \frac{\alpha}{2} (\alpha L - 2) \leq -\frac{\alpha}{2} \right) \Rightarrow \varphi(x_{k+1}) \leq \varphi(x) + \langle G_\alpha(x_k), x_k - x \rangle - \frac{\alpha}{2} \|G_\alpha(x_k)\|_2^2$$

$$x := x_k \Rightarrow \varphi(x_{k+1}) \leq \varphi(x_k) - \frac{\alpha}{2} \|G_\alpha(x_k)\|_2^2$$

7. Рассмотрим теперь $x = x^*$:

$$\varphi(x_{k+1}) \leq \varphi(x^*) + \langle G_\alpha(x_k), x_k - x^* \rangle - \frac{\alpha}{2} \|G_\alpha(x_k)\|_2^2$$

$$\varphi(x_{k+1}) - \varphi(x^*) \leq \langle G_\alpha(x_k), x_k - x^* \rangle - \frac{\alpha}{2} \|G_\alpha(x_k)\|_2^2 \leq \frac{1}{2\alpha} [2\langle \alpha G_\alpha(x_k), x_k - x^* \rangle - \|\alpha G_\alpha(x_k)\|_2^2] \leq$$

$$\leq \frac{1}{2\alpha} [2\langle \alpha G_\alpha(x_k), x_k - x^* \rangle - \|\alpha G_\alpha(x_k)\|_2^2 - \|x_k - x^*\|_2^2 + \|x_k - x^*\|_2^2] \leq$$

$$\leq \frac{1}{2\alpha} [-\|x_k - x^* - \alpha G_\alpha(x_k)\|_2^2 + \|x_k - x^*\|_2^2] \leq \frac{1}{2\alpha} [\|x_k - x^*\|_2^2 - \|x_{k+1} - x^*\|_2^2]$$

8. Суммируем $i = \overline{0, k-1}$ и суммируем их:

$$\sum_{i=0}^{k-1} [\varphi(x_{i+1}) - \varphi(x^*)] \leq \frac{1}{2\alpha} [\|x_0 - x^*\|_2^2 - \|x_k - x^*\|_2^2] \leq \frac{1}{2\alpha} \|x_0 - x^*\|_2^2$$

9. Поскольку $\varphi(x_{\{k\}})$ является убывающей последовательностью, из этого следует, что:

$$k\varphi(x_k) \leq \sum_{i=0}^{k-1} \varphi(x_{i+1}) \Rightarrow \varphi(x_k) \leq \frac{1}{k} \sum_{i=0}^{k-1} \varphi(x_{i+1})$$

$$\varphi(x_k) - \varphi(x^*) \leq \frac{1}{k} \sum_{i=0}^{k-1} [\varphi(x_{i+1}) - \varphi(x^*)] \leq \frac{\|x_0 - x^*\|_2^2}{2\alpha k} = \frac{L\|x_0 - x^*\|_2^2}{2k}$$

То есть имеет место сублинейная сходимость.

24. Теорема о сходимости проксимального градиентного метода для сильно выпуклой гладкой функции f .

i Рассматриваем задачу

$$\varphi(x) \rightarrow \min_{x \in \mathbb{R}^d}$$

Пусть $\varphi(x) = f(x) + r(x)$, причем

- f - μ -сильно выпуклая, L -гладкая, $\text{dom } f = \mathbb{R}^n$
- r - выпуклая и $\text{prox}_{\alpha r}(x_k) = \arg \min_{x \in \mathbb{R}^n} [\alpha r(x) + \frac{1}{2} \|x - x_k\|^2]$ может быть вычислен

Тогда проксимальный градиентный спуск с фиксированным шагом $\alpha \leq \frac{1}{L}$

$$x_{k+1} = \text{prox}_{\alpha r}(x_k - \alpha \nabla f(x_k))$$

сходится линейно, то есть имеет место

$$\|x_{k+1} - x^*\|_2^2 \leq (1 - \alpha\mu)^k \|x_0 - x^*\|_2^2$$

$$\begin{aligned} \|x_{k+1} - x^*\|_2^2 &= \|\text{prox}_{\alpha f}(x_k - \alpha \nabla f(x_k)) - x^*\|_2^2 \stackrel{1}{=} \|\text{prox}_{\alpha f}(x_k - \alpha \nabla f(x_k)) - \text{prox}_{\alpha f}(x^* - \alpha \nabla f(x^*))\|_2^2 \stackrel{2}{\leq} \\ &\stackrel{2}{\leq} \|x_k - \alpha \nabla f(x_k) - x^* + \alpha \nabla f(x^*)\|_2^2 = \|x_k - x^*\|^2 - 2\alpha \langle \nabla f(x_k) - \nabla f(x^*), x_k - x^* \rangle + \alpha^2 \|\nabla f(x_k) - \nabla f(x^*)\|_2^2 \end{aligned}$$

Воспользуемся L -гладкостью и сильной выпуклостью

$$\begin{aligned} \|\nabla f(x_k) - \nabla f(x^*)\|_2^2 &\leq 2L(f(x_k) - f(x^*) - \langle \nabla f(x^*), x_k - x^* \rangle)^3 \\ - \langle \nabla f(x_k) - \nabla f(x^*), x_k - x^* \rangle &\leq - \left(f(x_k) - f(x^*) + \frac{\mu}{2} \|x_k - x^*\|_2^2 \right) - \langle \nabla f(x^*), x_k - x^* \rangle \end{aligned}$$

Подставляем

$$\begin{aligned} \|x_{k+1} - x^*\|_2^2 &\leq \|x_k - x^*\|^2 - 2\alpha \left(f(x_k) - f(x^*) + \frac{\mu}{2} \|x_k - x^*\|_2^2 \right) - 2\alpha \langle \nabla f(x^*), x_k - x^* \rangle + \\ &\quad + \alpha^2 2L(f(x_k) - f(x^*) - \langle \nabla f(x^*), x_k - x^* \rangle) \\ &\leq (1 - \alpha\mu) \|x_k - x^*\|^2 + 2\alpha(\alpha L - 1) (f(x_k) - f(x^*) - \langle \nabla f(x^*), x_k - x^* \rangle) \end{aligned}$$

Так как f выпуклая: $f(x_k) - f(x^*) - \langle \nabla f(x^*), x_k - x^* \rangle \geq 0$ и при $\alpha \leq \frac{1}{L}$:

$$\|x_{k+1} - x^*\|_2^2 \leq (1 - \alpha\mu) \|x_k - x^*\|^2,$$

что и является линейной сходимостью с параметром $1 - \frac{\mu}{L}$.

0. Пусть $r : \mathbb{R}^n \rightarrow \mathbb{R} \cup \{+\infty\}$ выпуклая функция, для которой определен prox_r . Тогда $\forall x, y \in \mathbb{R}^n$, следующие условия эквивалентны:

- $\text{prox}_r(x) = y$,
- $x - y \in \partial r(y)$,
- $\langle x - y, z - y \rangle \leq r(z) - r(y)$ for any $z \in \mathbb{R}^n$.

Доказательство.

1 \Leftrightarrow 2. Первое условие может быть переписано в виде

$$y = \arg \min_{\tilde{x} \in \mathbb{R}^d} \left(r(\tilde{x}) + \frac{1}{2} \|x - \tilde{x}\|^2 \right).$$

Из условия оптимальности для выпуклой функции r , это эквивалентно:

$$0 \in \partial \left(r(\tilde{x}) + \frac{1}{2} \|x - \tilde{x}\|^2 \right) \Big|_{\tilde{x}=y} = \partial r(y) + y - x \Leftrightarrow x - y \in \partial r(y).$$

2 ⇒ 3. По определению субдифференциала, $\forall g \in \partial r(y)$, $\forall z \in \mathbb{R}^d$:

$$\langle g, z - y \rangle \leq r(z) - r(y).$$

В частности, это верно для $g = x - y \Rightarrow \langle x - y, z - y \rangle \leq r(z) - r(y)$

3 ⇒ 2. Пусть выполняется $\langle x - y, z - y \rangle \leq r(z) - r(y) \Rightarrow$ по определению субградиента $x - y \in \partial r(y)$.

- Пусть $f : \mathbb{R}^n \rightarrow \mathbb{R} \cup \{+\infty\}$ и $r : \mathbb{R}^n \rightarrow \mathbb{R} \cup \{+\infty\}$ выпуклые функции. Пусть f - L -гладкая, и для r определен оператор prox_r . Тогда, x^* - решение составной задачи оптимизации $\Leftrightarrow \forall \alpha > 0$, выполняется:

$$x^* = \text{prox}_{r,\alpha}(x^* - \alpha \nabla f(x^*))$$

Доказательство. Условия оптимальности:

$$\begin{aligned} 0 &\in \nabla f(x^*) + \partial r(x^*) \\ -\alpha \nabla f(x^*) &\in \alpha \partial r(x^*) \\ x^* - \alpha \nabla f(x^*) - x^* &\in \alpha \partial r(x^*) \end{aligned}$$

Воспользуемся пунктом 0:

$$\text{prox}_r(x) = y \Leftrightarrow x - y \in \partial r(y) \Rightarrow x^* = \text{prox}_{\alpha r}(x^* - \alpha \nabla f(x^*)) = \text{prox}_{r,\alpha}(x^* - \alpha \nabla f(x^*)).$$

- Оператор $\text{prox}_r(x)$ - firmly nonexpansive, т. е.

$$\|\text{prox}_r(x) - \text{prox}_r(y)\|_2^2 \leq \langle \text{prox}_r(x) - \text{prox}_r(y), x - y \rangle$$

и nonexpansive:

$$\|\text{prox}_r(x) - \text{prox}_r(y)\|_2 \leq \|x - y\|_2.$$

Доказательство. Пусть $u = \text{prox}_r(x)$, and $v = \text{prox}_r(y)$. Тогда из пункта 0:

$$\begin{aligned} \langle x - u, z_1 - u \rangle &\leq r(z_1) - r(u) \\ \langle y - v, z_2 - v \rangle &\leq r(z_2) - r(v). \end{aligned}$$

Полагая $z_1 = v$ и $z_2 = u$ и суммируя, получаем:

$$\langle x - u, v - u \rangle + \langle y - v, u - v \rangle \leq 0$$

$$\langle x - y, v - u \rangle + \|v - u\|_2^2 \leq 0$$

$$\|u - v\|_2^2 \leq \langle x - y, u - v \rangle \Rightarrow \|\text{prox}_r(x) - \text{prox}_r(y)\|_2^2 \leq \langle \text{prox}_r(x) - \text{prox}_r(y), x - y \rangle$$

Применяя неравенство Коши-Буняковского:

$$\|u - v\|_2^2 \leq \langle x - y, u - v \rangle \leq \|x - y\| \|u - v\| \Rightarrow \|u - v\| \leq \|x - y\| \Rightarrow \|\text{prox}_r(x) - \text{prox}_r(y)\|_2 \leq \|x - y\|$$

3. Пусть $f : \mathbb{R}^n \rightarrow \mathbb{R}$ - L -гладкая выпуклая функция. тогда $\forall x, y \in \mathbb{R}^n$, следующее неравенство сохраняются:

$$\begin{aligned} f(x) + \langle \nabla f(x), y - x \rangle + \frac{1}{2L} \|\nabla f(x) - \nabla f(y)\|_2^2 &\leq f(y) \Leftrightarrow \\ \Leftrightarrow \|\nabla f(y) - \nabla f(x)\|_2^2 &= \|\nabla f(x) - \nabla f(y)\|_2^2 \leq 2L(f(x) - f(y) - \langle \nabla f(y), x - y \rangle) \end{aligned}$$

Доказательство. Рассмотрим другую функцию $\varphi(y) = f(y) - \langle \nabla f(x), y \rangle$. Она, выпуклая как сумма выпуклых функций, а также L -гладкая, так как $\nabla \varphi(y) = \nabla f(y) - \nabla f(x)$ и $\|\nabla \varphi(y_1) - \nabla \varphi(y_2)\| = \|\nabla f(y_1) - \nabla f(y_2)\| \leq L\|y_1 - y_2\|$. То есть для φ выполняется $\varphi(y) \leq \varphi(x) + \langle \nabla \varphi(x), y - x \rangle + \frac{L}{2}\|y - x\|_2^2$.

$$\begin{aligned} x := y, y := y - \frac{1}{L}\nabla \varphi(y) \Rightarrow \varphi\left(y - \frac{1}{L}\nabla \varphi(y)\right) &\leq \varphi(y) + \left\langle \nabla \varphi(y), -\frac{1}{L}\nabla \varphi(y) \right\rangle + \frac{1}{2L}\|\nabla \varphi(y)\|_2^2 \\ \varphi\left(y - \frac{1}{L}\nabla \varphi(y)\right) &\leq \varphi(y) - \frac{1}{2L}\|\nabla \varphi(y)\|_2^2 \end{aligned}$$

По дифференциальному критерию первого порядка, оптимальная точка для φ определяется условием: $\nabla \varphi(y) = \nabla f(y) - \nabla f(x) = 0$. Поэтому $\forall x$, минимум функции $\varphi(y)$ находится в точке $y = x$. Тогда:

$$\varphi(x) \leq \varphi\left(y - \frac{1}{L}\nabla \varphi(y)\right) \leq \varphi(y) - \frac{1}{2L}\|\nabla \varphi(y)\|_2^2$$

Наконец, подставляем $\varphi(y) = f(y) - \langle \nabla f(x), y \rangle$:

$$\begin{aligned} f(x) - \langle \nabla f(x), x \rangle &\leq f(y) - \langle \nabla f(x), y \rangle - \frac{1}{2L}\|\nabla f(y) - \nabla f(x)\|_2^2 \\ f(x) + \langle \nabla f(x), y - x \rangle + \frac{1}{2L} \|\nabla f(x) - \nabla f(y)\|_2^2 &\leq f(y) \\ \|\nabla f(y) - \nabla f(x)\|_2^2 &\leq 2L(f(y) - f(x) - \langle \nabla f(x), y - x \rangle) \\ \text{меняем местами } x \text{ и } y \Rightarrow \|\nabla f(x) - \nabla f(y)\|_2^2 &\leq 2L(f(x) - f(y) - \langle \nabla f(y), x - y \rangle) \end{aligned}$$