Определения и формулировки

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

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

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

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

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

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

2. Метод градиентного спуска.

   Решаем задачу минимизации

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

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

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

3. Наискорейший спуск.

   Решаем задачу минимизации

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

   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.

4. *Липшицева парабола для гладкой функции.*

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

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

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 $\varphi_1(x) \leqslant f(x) \leqslant \varphi_2(x)\ \forall x$

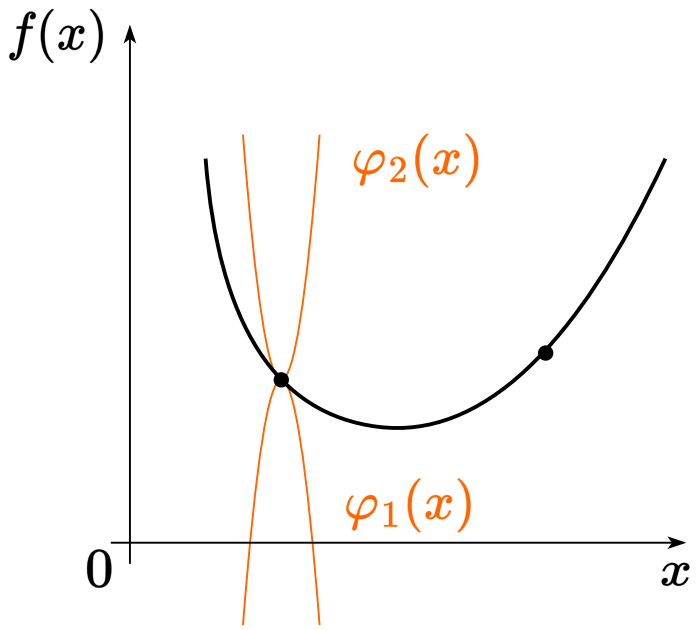


Рисунок 1: Иллюстрация Липшицевых парабол, между которыми зажата гладкая функция. Чаще нас интересует мажорирующая из них.

5. Размер шага наискорейшего спуска для квадратичной функции.

Решаем задачу минимизации методом наискорейшего спуска

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

$$f(x) = \frac{1}{2}x^T Ax - b^T x + c \to \min_{x \in \mathbb{R}^d}$$
$$x_{k+1} = x_k - \alpha(Ax_k - b)$$
$$\alpha_{opt} = \frac{2}{\mu + L}, \text{ где } \mu = \lambda_{\min}(A), L = \lambda_{\max}(A)$$
$$\kappa = \frac{L}{\mu} \geqslant 1$$
$$\rho = \frac{\kappa - 1}{\kappa + 1}$$
$$\|x_k - x^*\| \leqslant \rho^k \|x_0 - x^*\|$$

12. Связь PL-функций и сильно выпуклых функций.

Пусть f μ-сильно выпуклая и дифференицруемая $\Rightarrow f \in$ PL.
Обратное неверно - $f(x) = x^2 + 3\sin^2 x \in$ PL, но не сильно выпуклая (она вообще не выпуклая).

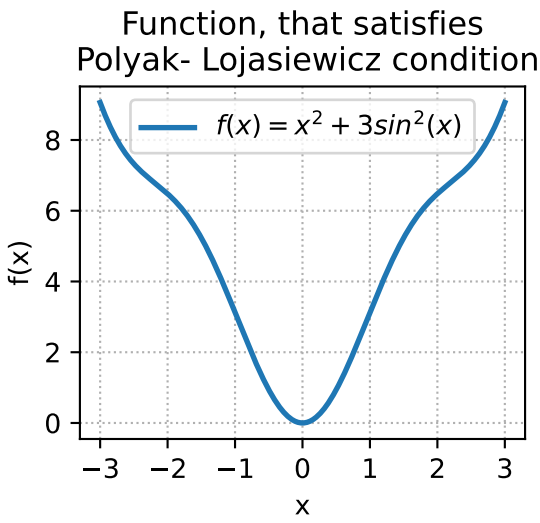


Рисунок 2: Пример невыпуклой PL функции

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

Рассмотрим задачу минимизации функции:

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

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

14. Привести пример сильно выпуклой задачи линейных наименьших квадратов (возможно, с

регуляризацией).

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.

15. 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.

16. Субградиент. Субдифференциал.

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

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

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

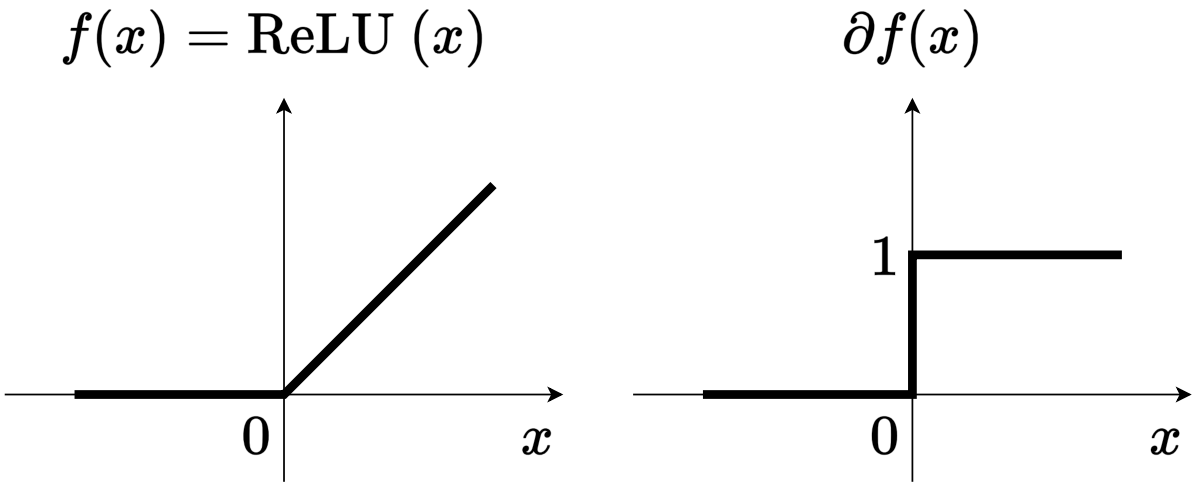


Рисунок 3: Субдифференциал функции ReLU.

17. 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.

18. 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.

19. 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	$f(x_k) - f^* \sim \mathcal{O}\left(\frac{1}{k^2}\right)$

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

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

21. Метод тяжелого шарика (Поляка).

Задача: $f(x) \to \min\limits_{x \in R^d}$, $f(x)$ - непрерывно дифференцируемая функция

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

22. Ускоренный градиентный метод Нестерова для выпуклых гладких функций.

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

$$\begin{aligned}
&\textbf{Обновление градиента:} && y_{k+1} = x_k - \frac{1}{L}\nabla f(x_k)\\
&\textbf{Экстраполяция:} && x_{k+1} = (1-\gamma_k)y_{k+1} + \gamma_k y_k\\
&\textbf{Экстраполяция веса:} && \lambda_{k+1} = \frac{1+\sqrt{1+4\lambda_k^2}}{2}\\
&\textbf{Экстраполяция веса:} && \gamma_k = \frac{1-\lambda_k}{\lambda_{k+1}}
\end{aligned}$$

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

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

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

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

$$\begin{aligned}
&\textbf{Обновление градиента:} && y_{k+1} = x_k - \frac{1}{L}\nabla f(x_k)\\
&\textbf{Экстрополяция:} && x_{k+1} = (1-\gamma_k)y_{k+1} + \gamma_k y_k\\
&\textbf{Экстрополяция весов:} && \gamma_k = \frac{\sqrt{L}-\sqrt{\mu}}{\sqrt{L}+\sqrt{\mu}}
\end{aligned}$$

Метод сходится линейно, а именно:

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

24. Проекция.

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

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

25. Достаточное условие существования проекции точки на множество.

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

26. Достаточное условие единственности проекции точки на множество.

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

27. Метод проекции градиента.

Рассматривается задача $f(x) \to \min\limits_{x \in S}$, 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.

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

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

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

29. Проекция как нерастягивающий оператор.

Проекция на выпуклое множество S является нерастягивающим оператором:

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

30. Метод Франк-Вульфа.

Рассматриваем задачу $f(x) \to \min\limits_{x \in S}$. Метод Франк-Вульфа имеет вид:

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

где γ_k - гиперпараметр.

31. 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.

32. 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 $\mathcal{O}\left(\left(1 - \frac{\mu}{L}\right)^k\right)$, где k — число итераций. То есть сходимость такая же, как и для безусловной задачи, но стоимость итерации может быть выше из-за проекции.

33. 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.

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

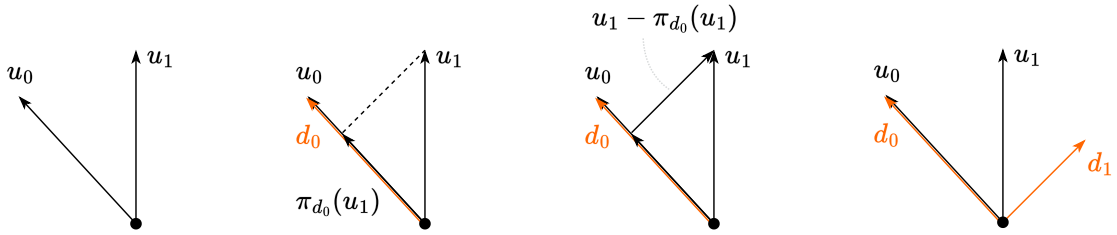
> Для гладких сильно выпуклых функций метод Франк-Вульфа имеет сходимость порядка $\mathcal{O}\left(\frac{1}{k}\right)$, где k — число итераций.

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

> A-ортогональность (сопряженность):
>
> $$x \perp_A y \iff x^T A y = 0.$$

36. 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:
>
> $$\begin{aligned} b_1 &= a_1 \\ b_2 &= a_2 - \text{proj}_{b_1} a_2 \\ b_3 &= a_3 - \text{proj}_{b_1} a_3 - \text{proj}_{b_2} a_2 \\ &\dots \\ b_n &= a_n - \sum_{i=1}^{n-1} \text{proj}_{b_i} a_n \end{aligned}$$



37. Метод сопряженных направлений.

Рассматриваем задачу

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

Идея

- В изотропном $A = I$ мире, наискорейший спуск стартующий из произвольной точки в *любом* пространстве натянутом на *линейную* оболочку из n ортогональных *Л*Н векторов будет сходится за n шагов в точной арифметике. Мы попытаемся в случае $A \neq I$ провести A-ортогональнализацию, 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. $x_0 - x^* = \sum\limits_{i=0}^{d-1} \alpha_i d_i$, 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 $k < n$
  - *Л*инейный поиск шага α: $f(x_k + \alpha_k d_k) \to \min\limits_{\alpha} \Rightarrow$

$$\alpha_k = -\frac{d_k^\top (Ax_k - b)}{d_k^\top A d_k}$$

  - 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.
  - $k := k + 1$

38. Метод сопряженных градиентов.

Рассматриваем задачу

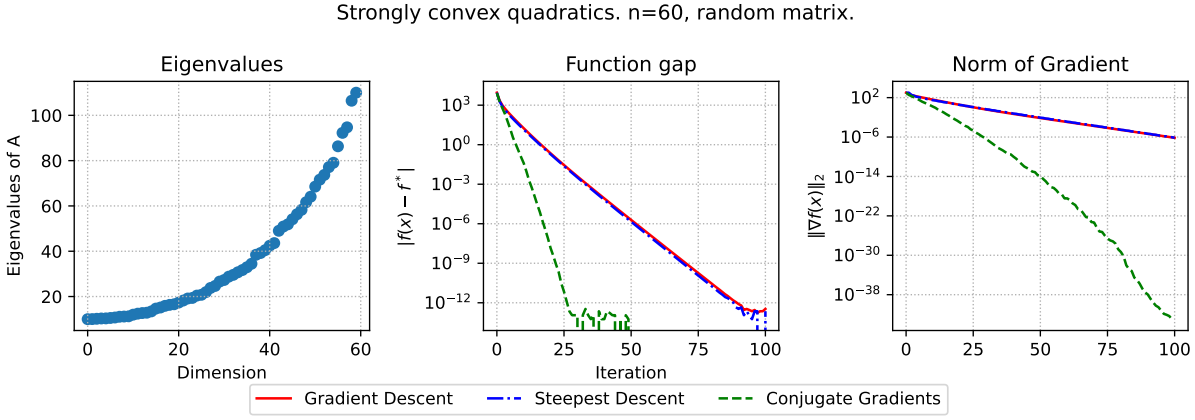
$$f(x) = \frac{1}{2}x^T Ax - b^T x + c \to \min_{x \in \mathbb{R}^d}$$

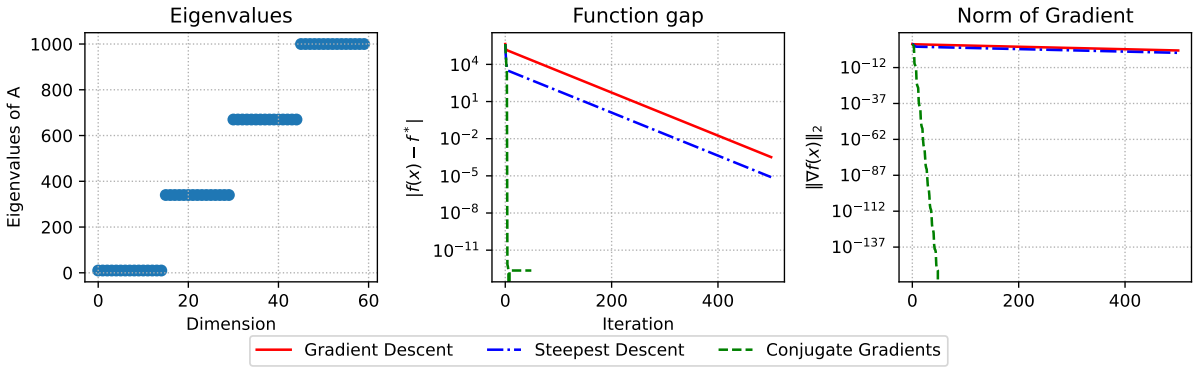
Метод сопряженных градиентов:

- $r_0 := b - Ax_0$
- 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.

39. Зависимость сходимости метода сопряженных градиентнов от спектра матрицы.

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





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

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

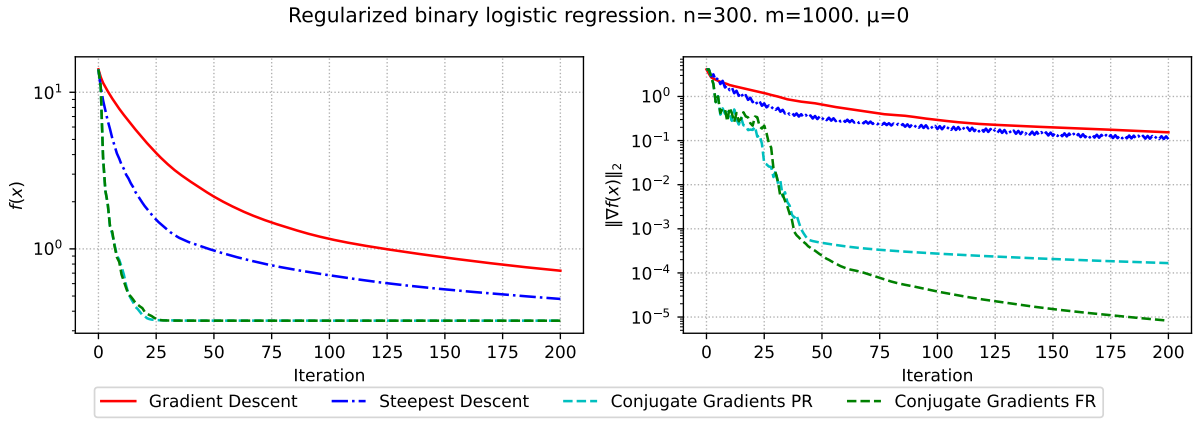
Имеет место оценка числа итераций при заданной точности ε: $\|x_k - x^*\|_A \leqslant \varepsilon \|x_0 - x^*\|_A$

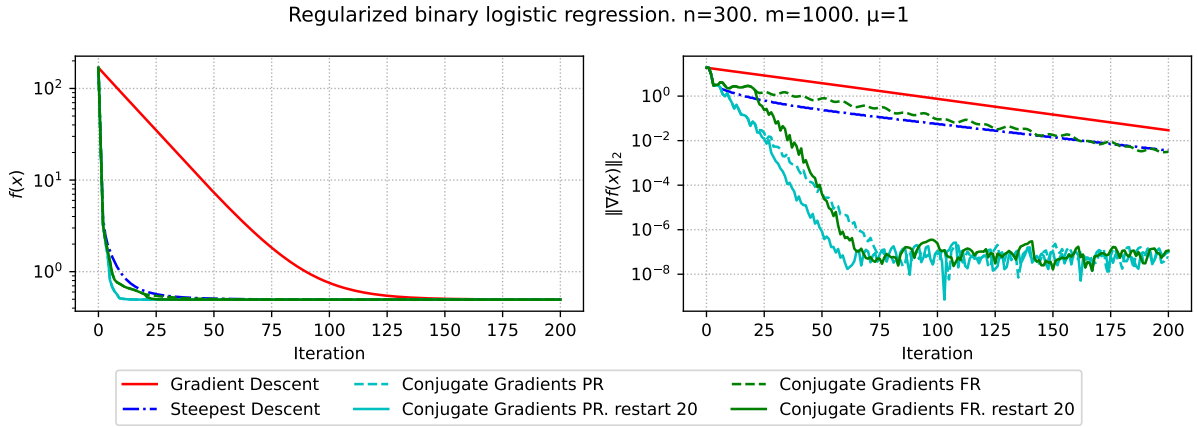
$$k \leqslant \left\lceil \frac{1}{2}\sqrt{\kappa(A)}\ln\left(\frac{2}{\varepsilon}\right)\right\rceil$$

41. 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.

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





42. Метод Ньютона.

Рассматривается задача минимизации функции с невырожденным гессианом.

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

43. Сходимость метода Ньютона для квадратичной функции.

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

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

44. 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 выпоняются 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.

45. Демпфированный метод Ньютона.

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

где α_k находят с помощью линейного поиска. Сходимость глобальная.

46. 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 $k = 1, 2, 3, \ldots$, 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.

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

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

48. Проксимальный оператор.

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

49. 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 $\text{prox}_r(x_k) = \arg\min_{x \in \mathbb{R}^n} \left[\frac{1}{2}\|x - x_k\|_2^2 + r(x)\right]$.

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

> Рассматривается задача: $\varphi(x) \to \min_{x \in \mathbb{R}^n}$, где $\varphi(x) = f(x) + r(x)$, $f(x)$ - гладкая выпуклая, $r(x)$ - негладкая выпуклая, проксимально дружественная.
>
> $$x_{k+1} = \text{prox}_{r,\alpha}\left(x_k - \alpha \nabla f(x_k)\right)$$
>
> Сходится за $\mathcal{O}\left(\frac{1}{k}\right)$.

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

> Рассматривается задача: $\varphi(x) \to \min_{x \in \mathbb{R}^n}$, 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.

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

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

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

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

54. Проксимальный оператор как нерастягивающий оператор.

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

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

и нерастягивающий:

$$\left\|\text{prox}_r(x) - \text{prox}_r(y)\right\|_2 \leqslant \|x - y\|_2$$

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

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

56. Метод стохастического градиентного спуска.

Решаемая задача: $f(x) \to \min\limits_{x \in \mathbb{R}^p}$, где $f(x) = \frac{1}{n}\sum\limits_{i=1}^{n} f_i(x)$

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

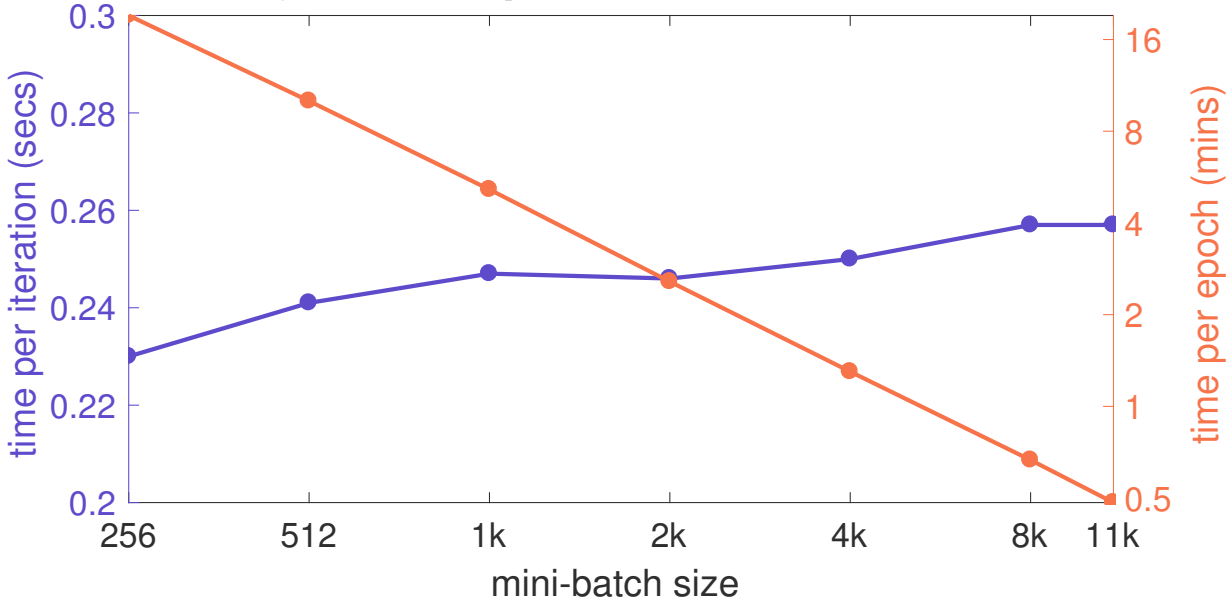
где i_k - случайно выбранный индекс. Если $\mathbb{P}(i_k = i) = \frac{1}{n}$, то $\mathbb{E}[\nabla f_{i_k}(x)] = \nabla f(x)$

57. 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).



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

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

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

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

60. 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.

61. Основная идея методов уменьшения дисперсии.

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

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

62. 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 $\tilde{x} = x_t$

63. 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: $g^{(k)} = g^{(k-1)} + \frac{1}{n}(g_{i_k}^{(k)} - g_{i_k}^{(k-1)}) = \frac{1}{n}\sum\limits_{i=1}^{n} g_i^{(k)}$
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:
$\mathbb{E}\left[f(x^{(k)}) - f^*\right] \leq \left(1 - \min\left(\frac{\mu}{16L}, \frac{1}{8n}\right)\right)^k \left(\frac{3}{2}(f(x^{(0)}) - f^*) + \frac{4L}{n}\|x^{(0)} - x^*\|^2\right) = O(\gamma^k)$

64. 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 $j = 1, \dots, p$:

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

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

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

65. Метод RMSProp.

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

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 $j = 1, \dots, p$:

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

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

66. 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:

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

67. Метод Adam.

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

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

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

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

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

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

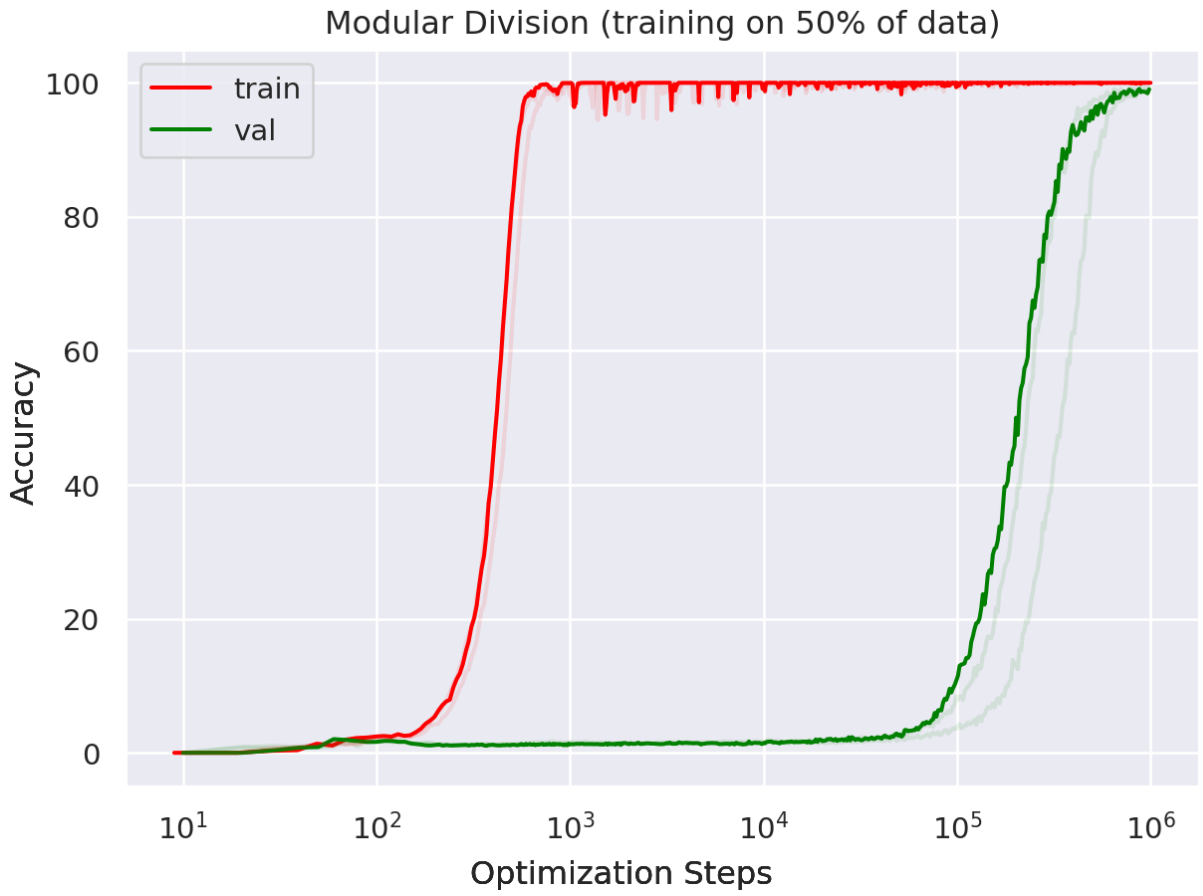
$$L(\alpha, \beta) = L(w_0 + \alpha w_1 + \beta w_2)$$

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.

69. 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.



70. 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.

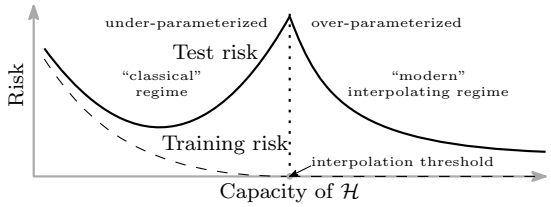


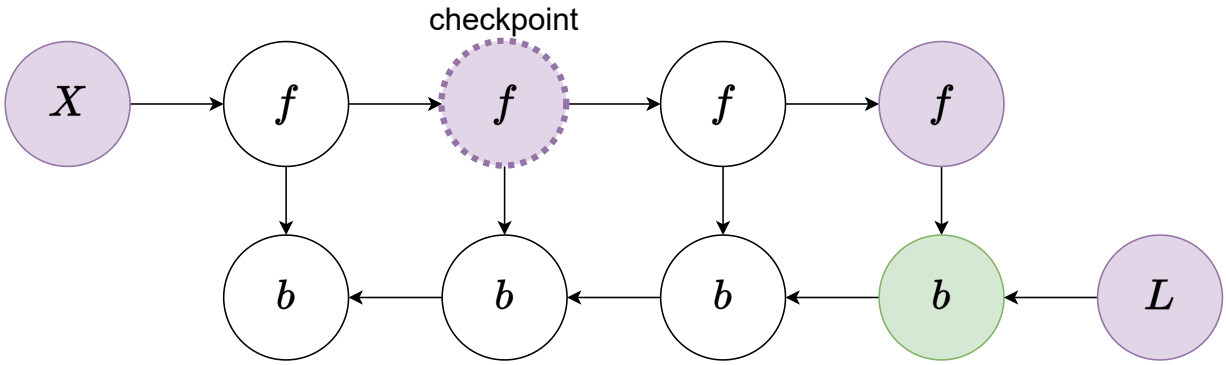
Рисунок 4: Иллюстрация зависимости обобщающей способности модели от размера.

71. 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: $\frac{\partial L}{\partial W} = \frac{\partial L}{\partial z^{(n)}} \cdot \frac{\partial z^{(n)}}{\partial z^{(n-1)}} \cdots \frac{\partial z^{(2)}}{\partial z^{(1)}} \cdot \frac{\partial z^{(1)}}{\partial W}$, 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.

72. 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.



73. 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.

74. 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.

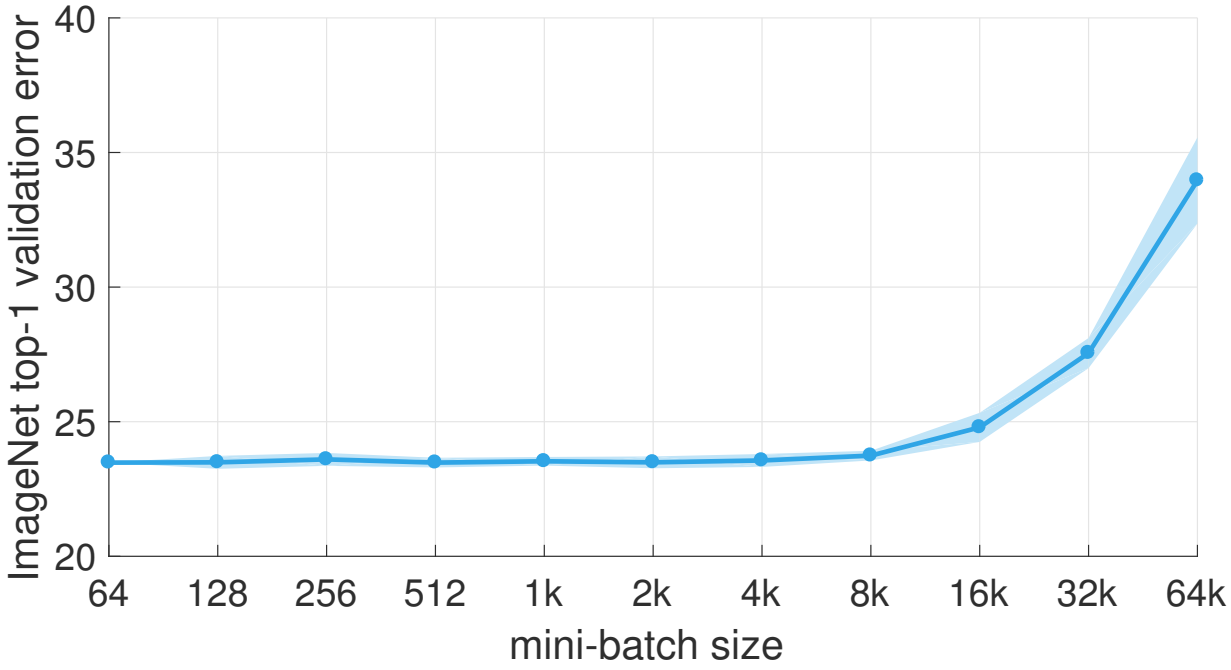


Рисунок 5: Увеличение размера батча приводит к росту ошибки из-за сложности минимизируемой функции.

75. *Дифференциальное уравнение градиентного потока.*

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

76. Характер сходимости траектории градиентного потока *для выпуклых функций* в терминах $\mathcal{O}(t)$.

$$f(x(t)) - f^* \leqslant \frac{1}{2t}\|x(0) - x^*\|^2 \quad \Rightarrow \quad \mathcal{O}\left(\frac{1}{t}\right).$$

77. Характер сходимости траектории градиентного потока *для* PL-функций в терминах $\mathcal{O}(t)$.

$$f(x(t)) - f^* \leqslant \exp\{-2\mu t\}(f(x(0)) - f^*) \quad \Rightarrow \quad \mathcal{O}\left(\exp\{-2\mu t\}\right).$$

78. Дифференциальное уравнение Нестеровского ускоренного градиентного потока.

$$\ddot{X}(t) + \frac{3}{t}\dot{X}(t) + \nabla f(X(t)) = 0.$$

79. Метод двойственного градиентного подъема.

Рассматривается задача:

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

Двойственная задача:

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

где $f^*(y) = \max_x \left[y^T x - f(x)\right]$ - 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 $x \in \arg\min_z \left[f(z) + u^T Az\right]$. Тогда определим метод двойственного градиентного подъема:

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

80. 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 $\frac{1}{\mu}$.

81. Идея dual decomposition.

Рассматриваем задачу $\sum_{i=1}^{B} f_i(x_i) \to \min_{Ax=b}$. 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

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

Тогда метод двойственного подъема запишется следующим образом:

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

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

Рассматриваем задачу $\sum\limits_{i=1}^{B} f_i(x_i) \to \min\limits_{\sum\limits_{i=1}^{B} A_i x_i \preccurlyeq b}$.

$$x_i^k \in \arg\min_{x_i} \left[f_i(x_i) + (u^{k-1})^T A_i x_i \right], \quad i = \overline{1, B}$$

$$u^k = \left(u^{k-1} + \alpha_k \left[\sum_{i=1}^{B} A_i x_i^k - b \right] \right)_+,$$

где $(u)_+$ обозначает $(u_+)_i = \max\{0, u_i\}$, $i = \overline{0, m}$.

83. Метод модифицированной функции Лагранжа.

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

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

В этом случае имеет место следующее:

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

84. Метод ADMM.

Рассматриваем задачу

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

После добавления штрафа за выход из бюджетного множества имеем $f(x) + g(z) + \|Ax + Bz - c\|^2 \to \min\limits_{Ax+Bz=c}$, 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:

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

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

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

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

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

Теоремы с доказательствами

1. Теорема сходимости градиентного спуска для гладких выпуклых функций.

Рассматриваем задачу

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

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, $0 < \alpha \leqslant \frac{1}{L}$. Тогда градиентный спуск сходится сублинейно, то есть:

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

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

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

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

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

$\left(\frac{L}{2}\alpha^2 - \alpha\right) \to \min\limits_\alpha$. Получаем оптимальный шаг: $\alpha = \frac{1}{L}$ и $f(x_k) - f(x_{k+1}) \geqslant \frac{1}{2L}\|\nabla f(x_k)\|^2$

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

$$y := x^*, x := x_k \Rightarrow f(x^*) \geqslant f(x_k) + \nabla f(x_k)^T(x^* - x_k) \Rightarrow$$

$$\Rightarrow f(x_k) \leqslant f(x^*) + \nabla f(x_k)^T(x_k - x^*) \Rightarrow f(x_k) - f(x^*) \leqslant \nabla f(x_k)^T(x_k - x^*)$$

Подставим $f(x_k)$ в (1):

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

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

$$(a-b)^T(a+b) = a^Ta - b^Ta + a^Tb - b^Tb = a^Ta - b^Tb = \|a\|^2 - \|b\|^2.$$

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

$$= \langle \nabla f(x_k), x_k - x^* \rangle - \frac{1}{2L}\langle \nabla f(x_k), \nabla f(x_k) \rangle =$$

$$= \frac{L}{2}\left(\left\langle \frac{\nabla f(x_k)}{L}, 2x_k + 2x^* - \frac{\nabla f(x_k)}{L} \right\rangle \right) =$$

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

Подставим в (2):

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

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

$$\sum_{k=0}^{T-1}(f(x_{k+1}) - f(x^*)) \leqslant \frac{L}{2}\left(\|x_0 - x^*\|^2 - \|x_T - x^*\|^2 \right) = \frac{L}{2}\left(R^2 - \|x_T - x^*\|^2 \right) \leqslant \frac{LR^2}{2}$$

$$\frac{1}{T}\sum_{k=0}^{T-1} f(x_{k+1}) - f(x^*) \leqslant \frac{LR^2}{2T}$$

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

Итого имеем:

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

То есть сходимость сублинейная.

2. Теорема сходимости градиентного спуска для гладких PL функций.

> Рассматриваем задачу
> $$f(x) \to \min_{x \in \mathbb{R}^d}$$
> и предполагаем, что f - μ-PL-функция, L-гладкая, для некоторых $L \geqslant \mu > 0$.
> Пусть $(x_T)_{T \in \mathbb{N}}$ это последовательность, созданная градиентным спуском с постоянным шагом $\alpha:\ 0 < \alpha \leqslant \frac{1}{L}$. Тогда:
> $$f(x_T) - f^* \leqslant (1 - \alpha\mu)^T(f(x^0) - f^*).$$

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

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

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

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

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

Условие PL: $\|\nabla f(x_T)\|^2 \geqslant 2\mu(f(x_T) - f^*)$

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

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

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

> **i**
>
> $$f(x) \to \min_{x \in \mathbb{R}^d}$$
>
> $$f(x) = \frac{1}{2}x^T Ax - b^T x + c,\ A \in \mathbb{S}_{++}$$
>
> Тогда градиентный спуск с шагом $\alpha = \frac{2}{\mu + L}$ сходится линейно с показателем $\frac{L-\mu}{L+\mu}$
>
> $$f(x_k) - f^* \leqslant \left(\frac{L - \mu}{L + \mu}\right)^k (f(x_0) - f^*).$$

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

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.

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

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

$$x_{k+1} - x^* = (I - \alpha Q\Lambda Q^T)(x_k - x^*) \mid \cdot Q^T$$

$$Q^T(x_{k+1} - x^*) = (Q^T - \alpha\Lambda Q^T)(x_k - x^*) = (I - \alpha\Lambda)Q^T(x_k - x^*)$$

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

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

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

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

4. 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
>
> $$f(\overline{x}) - f^* \leq \frac{G\|x_0 - x^*\|_2}{\sqrt{K}} \qquad \overline{x} = \frac{1}{K}\sum_{k=0}^{K-1} x_i$$
>
> Для фиксированных стратегий выбора шага можно получить оценки сходимости вида $\mathcal{O}\left(\frac{1}{\sqrt{k}}\right)$

Рассматривается задача

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

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

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

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

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

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

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

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

1. $\alpha_k = \alpha$.

   $$f_k^{best} - f^* \leqslant \frac{R^2 + G^2 k\alpha^2}{2k\alpha} = \frac{R^2}{2k\alpha} + G^2\alpha \underset{k\to\infty}{\to} G^2\alpha$$

   То есть при таком шаге о сходимости сказать ничего нельзя.

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

   $$f_k^{best} - f^* \leqslant \frac{R^2 + G^2\sum\limits_{i=1}^{k}\frac{R^2}{G^2 i}}{2\sum\limits_{i=1}^{k}\frac{R}{G\sqrt{i}}} = \frac{RG}{2}\frac{1 + \sum\limits_{i=1}^{k}\frac{1}{i}}{\sum\limits_{i=1}^{k}\frac{1}{\sqrt{i}}} \underset{k\gg 1}{\approx} \frac{RG}{2}\frac{1 + \int\limits_1^k \frac{dx}{x}}{\int\limits_1^k \frac{dx}{\sqrt{x}}} = \frac{RG}{2}\frac{1 + \ln k}{\sqrt{k} - 1} \underset{k\to\infty}{\to} 0.$$

   То есть при таком шаге есть сублинейная сходимость.

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

$$f_k^{best} - f^* \leqslant \frac{R^2 + G^2 \sum\limits_{i=1}^{k} \frac{1}{i^2}}{2 \sum\limits_{i=1}^{k} \frac{1}{i}} \underset{k \gg 1}{\approx} \frac{R^2 + G^2 \frac{\pi}{6}}{2 \ln k} \underset{k \to \infty}{\to} 0.$$

То есть при таком шаге есть сублинейная сходимость.

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

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

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

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

$$f_k^{best} - f^* \leqslant R\sqrt{\frac{G}{k}} \underset{k \to \infty}{\to} 0.$$

То есть при таком шаге есть сублинейная сходимость.

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

> Рассматривается задача
>
> $$f(x) \to \min_{x \in \mathbb{R}^d}$$
>
> $$f(x) = \frac{1}{2}x^T A x - b^T x + c, \quad A \in \mathbb{S}_{++} \Rightarrow \nabla f(x) = Ax - b \overset{\nabla f(x^*) = 0}{\Rightarrow} Ax^* = b.$$
>
> 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
>
> $$\|x_k - x^*\| \leqslant \left(\frac{\sqrt{\kappa} - 1}{\sqrt{\kappa} + 1}\right)^k \|x_0 - x^*\|.$$

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.

$$\tilde{x} = Q^T(x - x^*),$$

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

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:

$$
\tilde{z}_{k+1} = M\tilde{z}_k
$$

$$
M = \begin{pmatrix} I - \alpha\Lambda + \beta I & -\beta I \\ I & 0 \end{pmatrix} \in \mathbb{R}^{d \times d}
$$

Сделаем reshape:

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

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:

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

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

Собственные значения M_i имеют вид:

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

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

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

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

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

Итого получаем оценку

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

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

> Рассматриваем задачу
> $$f(x) \to \min_{x \in S},$$
> 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:
> $$f(x_T) - f^* \leqslant \frac{L\|x_0 - x^*\|_2^2}{2T} = \frac{LR^2}{2T}.$$

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

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

L-гладкость:

$$f(x_{k+1}) \leqslant f(x_k) + \langle \nabla f(x_k), x_{k+1} - x_k\rangle + \frac{L}{2}||x_{k+1} - x_k||_2^2 =$$

$$= f(x_k) - L\langle y_k - x_k, x_{k+1} - x_k\rangle + \frac{L}{2}||x_{k+1} - x_k||_2^2 =$$

$$= f(x_k) - \frac{L}{2}\left(||y_k - x_k||^2 + ||x_{k+1} - x_k||^2 - ||y_k - x_{k+1}||^2\right) + \frac{L}{2}||x_{k+1} - x_k||_2^2 =$$

$$= f(x_k) - \frac{1}{2L}\|\nabla f(x_k)\|^2 + \frac{L}{2}\|y_k - x_{k+1}\|^2.$$

$$(1) \Rightarrow \qquad \left\langle \frac{1}{L}\nabla f(x_k), x_k - x^*\right\rangle = \frac{1}{2}\left(\frac{1}{L^2}\|\nabla f(x_k)\|^2 + \|x_k - x^*\|^2 - \|x_k - x^* - \frac{1}{L}\nabla f(x_k)\|^2\right)$$

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

Воспользуемся свойством проекции:

$$\|x - \text{proj}_S(y)\|^2 + \|y - \text{proj}_S(y)\|^2 \leqslant \|x - y\|^2 \quad \forall x \in S$$

$$x := x^*, y := y_k \Rightarrow \|x^* - \text{proj}_S(y_k)\|^2 + \|y_k - \text{proj}_S(y_k)\|^2 \leqslant \|x^* - y_k\|^2$$

$$\|y_k - x^*\|^2 \geqslant \|x^* - x_{k+1}\|^2 + \|y_k - x_{k+1}\|^2$$

Выпуклость:

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

Суммируем от 0 до $T-1$:

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

$$\leqslant f(x_0) - f(x_T) + \frac{L}{2}\sum_{k=0}^{T-1}\|y_k - x_{k+1}\|^2 + \frac{L}{2}\|x_0 - x^*\|^2 - \frac{L}{2}\sum_{k=0}^{T-1}\|y_k - x_{k+1}\|^2 \leqslant$$

$$\leqslant f(x_0) - f(x_T) + \frac{L}{2}\|x_0 - x^*\|^2$$

$$\sum_{k=0}^{T-1} f(x_k) - Tf^* \leqslant f(x_0) - f(x_T) + \frac{L}{2}\|x_0 - x^*\|^2$$

$$\sum_{k=1}^{T}[f(x_k) - f^*] \leqslant \frac{L}{2}\|x_0 - x^*\|^2$$

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

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

То есть сходимость сублинейная.

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

> **i** Рассматриваем задачу
>
> $$f(x) \to \min_{x \in S},$$
>
> где f -μ-сильно выпуклая и L-гладкая. Пусть $S \subseteq \mathbb{R}^n$ замкнутое выпуклое множество. Тогда метод проекции градиента с постоянным шагом $\alpha \leqslant \frac{1}{L}$:
>
> $$x_{k+1} = \text{proj}_S(x_k - \alpha\nabla f(x_k))$$
>
> сходится со линейно и $\forall T$ выполняется неравенство:
>
> $$f(x_T) - f^* \leqslant (1 - \alpha\mu)^T(f(x_0) - f^*).$$

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

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

Рассматриваем задачу

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

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

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

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

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

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

L-гладкость:

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

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

Выпуклость:

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

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

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

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

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

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

9. Доказательство сходимости метода сопряженных градиентов и вывод формулы.

Рассматриваем задачу

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

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 $d_1, d_2, ...d_m$ попарно $A-$ ортогональные вектора. Тогда они линейно независимы. $A \in S^n_{++}$

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

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

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. $e_0 = x_0 - x^* = \sum\limits_{i=0}^{n-1} \delta_i d_i$

Доказательство:

Пусть есть n A-ортогональных векторов: $d_0, ..., d_{n-1}$. $x_{k+1} = x_0 + \sum\limits_{i=0}^{k} \alpha_i d_i$. α_i подбирается LineSearch. $\alpha_i = \frac{d_i^T r_i}{d_i^T A d_i}$ и необходимо показать, что $\delta_i = -\alpha_i$, $x_0 + \sum\limits_{i=0}^{n-1} \alpha_i d_i = x^*$.

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

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

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

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

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

Лемма:

1. В методе сопряженных градиентов мы расматриваем ортогонолизацию Грамма-Шмидта для невязок, те $u_i = r_i$. Формула Грамма-Шмидта: $d_i = u_i + \sum\limits_{i=0}^{i-1} \beta_{ij} d_j$, $\beta_{ij} = -\frac{u_i^T A d_j}{d_j^T A d_j}$.

2. 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: $e_k = -\sum\limits_{j=k}^{n-1} \alpha_j d_j$ и для некоторого l домножим 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.

3. 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 $r_k^T d_k = r_k^T u_k = r_k^T r_k$

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

   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:

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

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

   Если $i = j$: $\alpha_i \langle r_i, A d_i \rangle = \langle r_i, r_i \rangle - \langle r_i, r_{i+1} \rangle = \langle r_i, r_i \rangle$ 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:

$$\beta_{ij} = -\frac{r_i^T A d_j}{d_J^T A d_j} = \frac{1}{\alpha_j} \frac{r_i^T r_i}{d_J^T A d_j} = \left\{ \alpha_j = \frac{d_j^T r_j}{d_j^T A d_j} \right\} =$$

$$= \frac{d_j^T A d_j}{d_j^T r_j} \cdot \frac{r_i^T r_j}{d_j^T A d_j} = \frac{r_i^T r_i}{d_j^T r_j} = \frac{r_i^T r_j}{d_j^T r_i} = \frac{r_i^T r_j}{d_j^T r_j} = \frac{r_i^T r_j}{r_j^T r_j} = [i = j - 1] = \frac{r_i^T r_j}{r_{i-1}^T r_{i-1}}$$

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

> Рассматривам задачу
>
> $$f(x) \to \min_{x \in \mathbb{R}^d},$$
>
> 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:

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

$$= x_k - x^* - \left[\nabla^2 f(x_k)\right]^{-1} \int_0^1 \nabla^2 f(x^* + \tau(x_k - x^*))(x_k - x^*) d\tau =$$

$$= \left(I - \left[\nabla^2 f(x_k)\right]^{-1} \int_0^1 \nabla^2 f(x^* + \tau(x_k - x^*)) d\tau \right) (x_k - x^*) =$$

$$= \left[\nabla^2 f(x_k)\right]^{-1} \left(\nabla^2 f(x_k) - \int_0^1 \nabla^2 f(x^* + \tau(x_k - x^*)) d\tau \right) (x_k - x^*) =$$

$$= \left[\nabla^2 f(x_k)\right]^{-1} \left(\int_0^1 (\nabla^2 f(x_k) - \nabla^2 f(x^* + \tau(x_k - x^*)) d\tau \right) (x_k - x^*) =$$

$$= \left[\nabla^2 f(x_k)\right]^{-1} G_k (x_k - x^*)$$

3. Замена:

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

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

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

где $r_k = \|x_k - x^*\|$.

4. Итак:

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

5. Из-за непрерывности и симметрии Липшица Гессиана:

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

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

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

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

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

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

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^* $\left(\|x_0 - x^*\| < \frac{2\mu}{3M}\right)$ сходится квадратично к решению.

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

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

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

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

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:

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

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 $ □ 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 (DVP)**

   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:

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

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

   Умножая на Δy_k и ипользуя 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 Бройдена-Флетчера-Гольдфарба-Шанно (BFGS).

12. Теорема о сходимости проксимального градиентного для выпуклой гладкой функции f.

Рассматриваем задачу

$$\varphi(x) \to \min_{x \in \mathbb{R}^d}.$$

Причем $\varphi(x) = f(x) + r(x)$, и

- f-выпуклая и L-гладкая, $\text{dom} f = \mathbb{R}^n$
- r - выпуклая и $\text{prox}_{\alpha r}(x_k) = \arg\min_{x \in \mathbb{R}^n} \left[\alpha r(x) + \frac{1}{2}\|x - x_k\|^2\right]$ может быть вычислен

Тогда для проксимального метода с фиксированным шагом $\alpha = \frac{1}{L}$

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

выполнаяется

$$\varphi(x_k) - \varphi^* \leqslant \frac{L\|x_0 - x^*\|^2}{2k},$$

То есть имеет место сублинейная сходимость.

1. Представим отображение градиента, обозначаемое как $G_\alpha(x)$, действует как "градиентоподобный объект":

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

где $G_\alpha(x)$ имеет вид:

$$G_\alpha(x) = \frac{1}{\alpha}\left(x - \text{prox}_{\alpha r}\left(x - \alpha \nabla f(x)\right)\right)$$

$G_\alpha(x) = 0 \Leftrightarrow x = x^* \Rightarrow G_\alpha$ аналогичен ∇f.

2. L-гладкость:

$$f(x_{k+1}) \leqslant f(x_k) + \langle \nabla f(x_k), x_{k+1} - x_k\rangle + \frac{L}{2}\|x_{k+1} - x_k\|_2^2$$

Выпуклость:

$$f(x) \geqslant f(x_k) + \langle \nabla f(x_k), x - x_k\rangle$$

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

$$\leqslant f(x) + \langle \nabla f(x_k), x_{k+1} - x\rangle + \frac{\alpha^2 L}{2}\|G_\alpha(x_k)\|_2^2 \qquad (1)$$

3. Воспользуемся свойством проксимального оператора:

$$x_{k+1} = \text{prox}_{\alpha r}\left(x_k - \alpha \nabla f(x_k)\right) \quad \Leftrightarrow \quad x_k - \alpha \nabla f(x_k) - x_{k+1} \in \partial \alpha r(x_{k+1})$$

$$x_k - x_{k+1} = \alpha G_\alpha(x_k) \qquad \Rightarrow \qquad \alpha G_\alpha(x_k) - \alpha \nabla f(x_k) \in \partial \alpha r(x_{k+1})$$
$$G_\alpha(x_k) - \nabla f(x_k) \in \partial r(x_{k+1})$$

4. По определению субградиента:

$$r(x) \geqslant r(x_{k+1}) + \langle g, x - x_{k+1}\rangle, \quad g \in \partial r(x_{k+1})$$
$$r(x) \geqslant r(x_{k+1}) + \langle G_\alpha(x_k) - \nabla f(x), x - x_{k+1}\rangle$$

$$r(x) \geqslant r(x_{k+1}) + \langle G_\alpha(x_k), x - x_{k+1}\rangle - \langle \nabla f(x), x - x_{k+1}\rangle$$
$$\langle \nabla f(x), x_{k+1} - x\rangle \leqslant r(x) - r(x_{k+1}) - \langle G_\alpha(x_k), x - x_{k+1}\rangle$$

5. Подставляем полученные результаты в (1):

$$f(x_{k+1}) \leqslant f(x) + \langle \nabla f(x_k), x_{k+1} - x\rangle + \frac{\alpha^2 L}{2}\|G_\alpha(x_k)\|_2^2$$
$$f(x_{k+1}) \leqslant f(x) + r(x) - r(x_{k+1}) - \langle G_\alpha(x_k), x - x_{k+1}\rangle + \frac{\alpha^2 L}{2}\|G_\alpha(x_k)\|_2^2$$
$$f(x_{k+1}) + r(x_{k+1}) \leqslant f(x) + r(x) - \langle G_\alpha(x_k), x - x_k + \alpha G_\alpha(x_k)\rangle + \frac{\alpha^2 L}{2}\|G_\alpha(x_k)\|_2^2$$

6. Используя $□ (x) = f(x) + r(x) $ доказываем монотонное уменьшение итерации:

$$\varphi(x_{k+1}) \leqslant \varphi(x) - \langle G_\alpha(x_k), x - x_k\rangle - \langle G_\alpha(x_k), \alpha G_\alpha(x_k)\rangle + \frac{\alpha^2 L}{2}\|G_\alpha(x_k)\|_2^2$$
$$\varphi(x_{k+1}) \leqslant \varphi(x) + \langle G_\alpha(x_k), x_k - x\rangle + \frac{\alpha}{2}(\alpha L - 2)\|G_\alpha(x_k)\|_2^2$$
$$\left(\alpha \leqslant \frac{1}{L} \Rightarrow \frac{\alpha}{2}(\alpha L - 2) \leqslant -\frac{\alpha}{2}\right) \Rightarrow \quad \varphi(x_{k+1}) \leqslant \varphi(x) + \langle G_\alpha(x_k), x_k - x\rangle - \frac{\alpha}{2}\|G_\alpha(x_k)\|_2^2$$
$$x := x_k \Rightarrow \qquad \varphi(x_{k+1}) \leqslant \varphi(x_k) - \frac{\alpha}{2}\|G_\alpha(x_k)\|_2^2$$

7. Рассмотрим теперь $x = x^*$:

$$\varphi(x_{k+1}) \leqslant \varphi(x^*) + \langle G_\alpha(x_k), x_k - x^*\rangle - \frac{\alpha}{2}\|G_\alpha(x_k)\|_2^2$$
$$\varphi(x_{k+1}) - \varphi(x^*) \leqslant \langle G_\alpha(x_k), x_k - x^*\rangle - \frac{\alpha}{2}\|G_\alpha(x_k)\|_2^2 \leqslant \frac{1}{2\alpha}\left[2\langle \alpha G_\alpha(x_k), x_k - x^*\rangle - \|\alpha G_\alpha(x_k)\|_2^2\right] \leqslant$$
$$\leqslant \frac{1}{2\alpha}\left[2\langle \alpha G_\alpha(x_k), x_k - x^*\rangle - \|\alpha G_\alpha(x_k)\|_2^2 - \|x_k - x^*\|_2^2 + \|x_k - x^*\|_2^2\right] \leqslant$$
$$\leqslant \frac{1}{2\alpha}\left[-\|x_k - x^* - \alpha G_\alpha(x_k)\|_2^2 + \|x_k - x^*\|_2^2\right] \leqslant \frac{1}{2\alpha}\left[\|x_k - x^*\|_2^2 - \|x_{k+1} - x^*\|_2^2\right]$$

8. Суммируем $i = \overline{0, k-1}$ и суммируем их:

$$\sum_{i=0}^{k-1}\left[\varphi(x_{i+1}) - \varphi(x^*)\right] \leqslant \frac{1}{2\alpha}\left[\|x_0 - x^*\|_2^2 - \|x_k - x^*\|_2^2\right] \leqslant \frac{1}{2\alpha}\|x_0 - x^*\|_2^2$$

9. Поскольку $ □ (x_{k}) $ является убывающей последовательностью, из этого следует, что:

$$k\varphi(x_k) \leqslant \sum_{i=0}^{k-1}\varphi(x_{i+1}) \Rightarrow \varphi(x_k) \leqslant \frac{1}{k}\sum_{i=0}^{k-1}\varphi(x_{i+1})$$
$$\varphi(x_k) - \varphi(x^*) \leqslant \frac{1}{k}\sum_{i=0}^{k-1}\left[\varphi(x_{i+1}) - \varphi(x^*)\right] \leqslant \frac{\|x_0 - x^*\|_2^2}{2\alpha k} = \frac{L\|x_0 - x^*\|_2^2}{2k}$$

То есть имеет место сублинейная сходимость.

13. Теорема о сходимости проксимального градиентного для сильно выпуклой гладкой функции f.

Рассматриваем задачу

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

Пусть $\varphi(x) = f(x) + r(x)$, причем

- f - μ-сильно выпуклая, L-гладкая, $\mathrm{dom} f = \mathbb{R}^n$
- r - выпуклая и $\mathrm{prox}_{\alpha r}(x_k) = \arg\min_{x \in \mathbb{R}^n} \left[\alpha r(x) + \frac{1}{2}\|x - x_k\|^2\right]$ может быть вычислен

Тогда проксимальный градиентный спуск с фиксированным шагом $\alpha \leqslant \frac{1}{L}$

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

сходится линейно, то есть имеет место

$$\|x_{k+1} - x^*\|_2^2 \leqslant (1 - \alpha\mu)^k \|x_0 - x^*\|_2^2$$

$\|x_{k+1} - x^*\|_2^2 = \|\mathrm{prox}_{\alpha f}(x_k - \alpha\nabla f(x_k)) - x^*\|_2^2 \overset{1}{=} \|\mathrm{prox}_{\alpha f}(x_k - \alpha\nabla f(x_k)) - \mathrm{prox}_{\alpha f}(x^* - \alpha\nabla f(x^*))\|_2^2 \overset{2}{\leqslant}$

$\overset{2}{\leqslant} \|x_k - \alpha\nabla f(x_k) - x^* + \alpha\nabla f(x^*)\|_2^2 = \|x_k - x^*\|^2 - 2\alpha\langle \nabla f(x_k) - \nabla f(x^*), x_k - x^*\rangle + \alpha^2 \|\nabla f(x_k) - \nabla f(x^*)\|_2^2$

Воспользуемся L-гладкостью и сильной выпуклостью

$$\|\nabla f(x_k) - \nabla f(x^*)\|_2^2 \leqslant 2L\left(f(x_k) - f(x^*) - \langle \nabla f(x^*), x_k - x^*\rangle\right) \quad {}^{3}$$

$$-\langle \nabla f(x_k) - \nabla f(x^*), x_k - x^*\rangle \leqslant -\left(f(x_k) - f(x^*) + \frac{\mu}{2}\|x_k - x^*\|_2^2\right) - \langle \nabla f(x^*), x_k - x^*\rangle$$

Подставляем

$$\begin{aligned}\|x_{k+1} - x^*\|_2^2 &\leqslant \|x_k - x^*\|^2 - 2\alpha\left(f(x_k) - f(x^*) + \frac{\mu}{2}\|x_k - x^*\|_2^2\right) - 2\alpha\langle \nabla f(x^*), x_k - x^*\rangle + \\ &+ \alpha^2 2L\left(f(x_k) - f(x^*) - \langle \nabla f(x^*), x_k - x^*\rangle\right) \\ &\leqslant (1 - \alpha\mu)\|x_k - x^*\|^2 + 2\alpha(\alpha L - 1)\left(f(x_k) - f(x^*) - \langle \nabla f(x^*), x_k - x^*\rangle\right)\end{aligned}$$

Так как f выпуклая: $f(x_k) - f(x^*) - \langle \nabla f(x^*), x_k - x^*\rangle \geqslant 0$ и при $\alpha \leqslant \frac{1}{L}$:

$$|x_{k+1} - x^*\|_2^2 \leqslant (1 - \alpha\mu)\|x_k - x^*\|^2,$$

что и является линейной сходимостью с параметром $1 - \frac{\mu}{L}$.

0. Пусть $r : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ выпуклая функция, для которой определен prox_r. Тогда $\forall x, y \in \mathbb{R}^n$, следующие условия эквивалентны:

   - $\mathrm{prox}_r(x) = y$,
   - $x - y \in \partial r(y)$,
   - $\langle x - y, z - y\rangle \leqslant r(z) - r(y)$ 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:

$$0 \in \partial \left(r(\tilde{x}) + \frac{1}{2}\|x - \tilde{x}\|^2 \right)\Big|_{\tilde{x}=y} = \partial r(y) + y - x \Leftrightarrow x - y \in \partial r(y).$$

$2 \Rightarrow 3$. По определению субдифференциала, $\forall g \in \partial r(y),\ \forall z \in \mathbb{R}^d$:

$$\langle g, z - y \rangle \leqslant r(z) - r(y).$$

В частности, это верно для $g = x - y \Rightarrow \langle x - y, z - y \rangle \leqslant r(z) - r(y)$

$3 \Rightarrow 2$. Пусть выполняется $\langle x - y, z - y \rangle \leqslant r(z) - r(y) \Rightarrow$ по определению субградиента $x - y \in \partial r(y)$.

1. Пусть $f : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ и $r : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ 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:
$$0 \in \nabla f(x^*) + \partial r(x^*)$$
$$-\alpha \nabla f(x^*) \in \alpha \partial r(x^*)$$
$$x^* - \alpha \nabla f(x^*) - x^* \in \alpha \partial r(x^*)$$
Воспользуемся пунктом 0:
$$\text{prox}_r(x) = y \Leftrightarrow x - y \in \partial r(y) \Rightarrow x^* = \text{prox}_{\alpha r}(x^* - \alpha \nabla f(x^*)) = \text{prox}_{r,\alpha}(x^* - \alpha \nabla f(x^*)).$$

2. Оператор $\text{prox}_r(x)$ - firmly nonexpansive, т. е.
$$\|\text{prox}_r(x) - \text{prox}_r(y)\|_2^2 \leqslant \langle \text{prox}_r(x) - \text{prox}_r(y), x - y \rangle$$
и nonexpansive:
$$\|\text{prox}_r(x) - \text{prox}_r(y)\|_2 \leqslant \|x - y\|_2.$$
Доказательство. Пусть $u = \text{prox}_r(x)$, and $v = \text{prox}_r(y)$. Тогда из пункта 0:
$$\langle x - u, z_1 - u \rangle \leqslant r(z_1) - r(u)$$
$$\langle y - v, z_2 - v \rangle \leqslant r(z_2) - r(v).$$
Полагая $z_1 = v$ и $z_2 = u$ и суммируя, получаем:
$$\langle x - u, v - u \rangle + \langle y - v, u - v \rangle \leqslant 0$$
$$\langle x - y, v - u \rangle + \|v - u\|_2^2 \leqslant 0$$
$$\|u - v\|_2^2 \leqslant \langle x - y, u - v \rangle \Rightarrow \|\text{prox}_r(x) - \text{prox}_r(y)\|_2^2 \leqslant \langle \text{prox}_r(x) - \text{prox}_r(y), x - y \rangle$$
Применяя неравенство Коши-Буняковского:
$$\|u - v\|_2^2 \leqslant \langle x - y, u - v \rangle \leqslant \|x - y\|\|u - v\| \Rightarrow \|u - v\| \leqslant \|x - y\| \Rightarrow \|\text{prox}_r(x) - \text{prox}_r(y)\|_2 \leqslant \|x - y\|$$

3. Пусть $f : \mathbb{R}^n \to \mathbb{R}$ - L-гладкая выпуклая функция. тогда $\forall x, y \in \mathbb{R}^n$, следующее неравенство сохраняются:

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

$$\Leftrightarrow \|\nabla f(y) - \nabla f(x)\|_2^2 = \|\nabla f(x) - \nabla f(y)\|_2^2 \leqslant 2L\left(f(x) - f(y) - \langle \nabla f(y), x - y \rangle\right)$$

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 $\|\nabla\varphi(y_1) - \nabla\varphi(y_2)\| = \|\nabla f(y_1) - \nabla f(y_2)\| \leqslant L\|y_1 - y_2\|$. То есть для φ выполняется $\varphi(y) \leqslant \varphi(x) + \langle \nabla\varphi(x), y - x \rangle + \frac{L}{2}\|y - x\|_2^2$.

$$x := y, y := y - \frac{1}{L}\nabla\varphi(y) \quad \Rightarrow \quad \varphi\left(y - \frac{1}{L}\nabla\varphi(y)\right) \leqslant \varphi(y) + \left\langle \nabla\varphi(y), -\frac{1}{L}\nabla\varphi(y) \right\rangle + \frac{1}{2L}\|\nabla\varphi(y)\|_2^2$$

$$\varphi\left(y - \frac{1}{L}\nabla\varphi(y)\right) \leqslant \varphi(y) - \frac{1}{2L}\|\nabla\varphi(y)\|_2^2$$

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:

$$\varphi(x) \leqslant \varphi\left(y - \frac{1}{L}\nabla\varphi(y)\right) \leqslant \varphi(y) - \frac{1}{2L}\|\nabla\varphi(y)\|_2^2$$

Наконец, подставляем $\varphi(y) = f(y) - \langle \nabla f(x), y \rangle$:

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

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

$$\|\nabla f(y) - \nabla f(x)\|_2^2 \leqslant 2L\left(f(y) - f(x) - \langle \nabla f(x), y - x \rangle\right)$$

меняем местами x и y $\quad \Rightarrow \quad \|\nabla f(x) - \nabla f(y)\|_2^2 \leqslant 2L\left(f(x) - f(y) - \langle \nabla f(y), x - y \rangle\right)$

14. Теорема о сходимости стохастического градиентного спуска в гладком PL-случае.

> Рассмотрим задачу оптимизации $f(x(t)) \to \min\limits_{\substack{x(t) \in \mathbb{R}^d \\ t \in T}}$, функции $f(x)$ и $x(t)$ гладкие и f - выпуклая, тогда градиентный поток сходится сублинейно со скоростью $\mathcal{O}\left(\frac{1}{t}\right)$, то есть выполняется $f(x(t)) - f^* \leqslant \frac{1}{2t}\|x(0) - x^*\|^2$.

Докажем монотонность сходимости.

$$\frac{d}{dx}f(x(t)) = \nabla f(x(t))^T \underbrace{\dot{x}(t)}_{= \frac{dx(t)}{dt} = -\nabla f(x)} = -\|\nabla f(x(t))\|_2^2 \leqslant 0.$$

Если $f(x)$ - ограничена снизу, то $f(x(t))$ сходится как невозрастающая ограниченная снизу функция. Очевидно, что GF сходится к стационарной точке, в которой $\nabla f = 0$ (потенциально минимум, максимум или седловая точка).

Воспользуемся выпуклостью. По дифференциальному критерию 1-го порядка:

$$f(x) \geqslant f(y) + \nabla f(y)^T(x - y) \quad \Rightarrow \quad \nabla f(y)^T(x - y) \leqslant f(x) - f(y)$$

$$\begin{aligned}\frac{d}{dt}\left[\|x(t)-x^*\|^2\right] &= -2(x(t)-x^*)^T\nabla f(x(t)) \\ &= 2(x^*-x(t))^T\nabla f(x(y)) \\ &\leqslant 2\left[f^*-f(x(t))\right] \\ &= -2\left[f(x(t))-f^*\right] \\ f(x(t))-f^* &\leqslant -\frac{1}{2}\frac{d}{dt}\left[\|x(t)-x^*\|^2\right]\end{aligned}$$

$$f(x(t))-f^* \leqslant \frac{1}{t}\int\limits_0^t \left[f(x(s))-f^*\right]ds \leqslant -\frac{1}{2t}\int\limits_0^t \frac{d}{ds}\left[\|x(s)-x^*\|^2\right]ds = -\frac{1}{2t}\|x(t)-x^*\|^2+\frac{1}{2t}\|x(0)-x^*\|^2 \leqslant$$

$$\leqslant \frac{1}{2t}\|x(0)-x^*\|^2.$$

15. Теорема сходимости траектории градиентного потока для выпуклых и PL-функций.

> **i** Рассмотрим задачу оптимизации $f(x(t)) \to \min\limits_{\substack{x(t)\in\mathbb{R}^d \\ t\in T}}$, функции $f(x)$ и $x(t)$ гладкие, f - выпуклая и удовлетворяет условию PL, т.е. $\forall x \mapsto \|\nabla f(x)|^2 \geqslant 2\mu(f(x)-f^*)$, тогда градиентный поток сходится линейно, а именно, выполняется неравенство $f(x(t))-f^* \leqslant \exp\{-2\mu t\}\left(f(x(0))-f^*\right)$.

Предположим, у нас выполняется условие Поляка-Лоясевича $\|\nabla f(x)\|^2 \geqslant 2\mu(f(x)-f^*)$, тогда

$$\frac{d}{dt}\left(f(x(t))-f(x^*)\right) = \nabla f(x(t))^T \underbrace{\dot{x}(t)}_{=\frac{dx}{dt}=-\nabla f(x)} = -\|\nabla f(x(t))|_2^2 \leqslant -2\mu\left(f(x(t))-f^*\right)$$

Интегрируя, получаем:

$$f(x(t))-f^* \leqslant \exp\{-2\mu t\}\left(f(x(0))-f^*\right)$$