1 БАЗА

2. Теор.

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

learning rate

Gradient Descent

Daniil Merkulov

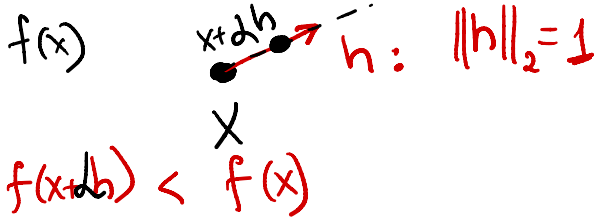
Optimization methods. MIPT

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

x_{k+1}: $n \times 1$; x_k: $n \times 1$; α_k: 1×1; $\nabla f(x_k)$: $n \times 1$

Direction of local steepest descent

Let's consider a linear approximation of the differentiable function f along some direction $h, \|h\|_2 = 1$:



Direction of local steepest descent

Let's consider a linear approximation of the differentiable function f along some direction $h, \|h\|_2 = 1$:

$$f(x + \alpha h) = f(x) + \alpha \langle f'(x), h \rangle + o(\alpha)$$

Direction of local steepest descent

Let's consider a linear approximation of the differentiable function f along some direction $h, \|h\|_2 = 1$:

$$f(x + \alpha h) = f(x) + \alpha \langle f'(x), h \rangle + o(\alpha)$$

We want h to be a decreasing direction:

$$\boxed{f(x + \alpha h) < f(x)}$$

трёб. убыв.

$$f(x) + \alpha \langle f'(x), h \rangle + o(\alpha) < f(x)$$

Direction of local steepest descent

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

Let's consider a linear approximation of the differentiable function f along some direction $h, \|h\|_2 = 1$:

$$f(x + \alpha h) = f(x) + \alpha \langle f'(x), h \rangle + o(\alpha)$$

We want h to be a decreasing direction:

$$f(x + \alpha h) < f(x)$$

$$f(x) + \alpha \langle f'(x), h \rangle + o(\alpha) < f(x)$$

and going to the limit at $\alpha \to 0$:

$$\langle f'(x), h \rangle \leq 0$$

Direction of local steepest descent

Let's consider a linear approximation of the differentiable function f along some direction $h, \|h\|_2 = 1$:

$$f(x + \alpha h) = f(x) + \alpha \langle f'(x), h \rangle + o(\alpha)$$

We want h to be a decreasing direction:

$$f(x + \alpha h) < f(x)$$

$$f(x) + \alpha \langle f'(x), h \rangle + o(\alpha) < f(x)$$

and going to the limit at $\alpha \to 0$:

$$\langle f'(x), h \rangle \leq 0$$

Also from Cauchy–Bunyakovsky–Schwarz inequality:

$$|\langle f'(x), h \rangle| \leq \|f'(x)\|_2 \|h\|_2$$

$$\langle f'(x), h \rangle \geq -\|f'(x)\|_2 \|h\|_2 = -\|f'(x)\|_2$$

$$\max -\langle \nabla f, h \rangle \Leftrightarrow \min \langle \nabla f, h \rangle, \quad \|h\|^2 = 1$$

$$L = \nabla f^T h + \lambda (h^T h - 1)$$

$$\frac{\partial L}{\partial h} = \nabla f + 2\lambda h \Rightarrow h = -\frac{1}{2\lambda} \nabla f$$

$$\frac{\partial L}{\partial \lambda} \Rightarrow \|h\|_2^2 = 1 \qquad \frac{1}{4\lambda^2} \|\nabla f\|^2 = 1$$

$$\Rightarrow 2\lambda = \|\nabla f\|$$

Direction of local steepest descent

Let's consider a linear approximation of the differentiable function f along some direction $h, \|h\|_2 = 1$:

$$\boxed{f(x + \alpha h) = f(x) + \alpha \langle f'(x), h \rangle + o(\alpha)}$$

We want h to be a decreasing direction:

$$f(x + \alpha h) < f(x)$$

$$f(x) + \alpha \langle f'(x), h \rangle + o(\alpha) < f(x)$$

and going to the limit at $\alpha \to 0$:

$$\langle f'(x), h \rangle \leq 0$$

Also from Cauchy–Bunyakovsky–Schwarz inequality:

$$|\langle f'(x), h \rangle| \leq \|f'(x)\|_2 \|h\|_2$$
$$\langle f'(x), h \rangle \geq -\|f'(x)\|_2 \|h\|_2 = -\|f'(x)\|_2$$

Thus, the direction of the antigradient

$$\boxed{h = -\frac{f'(x)}{\|f'(x)\|_2}}$$

gives the direction of the **steepest local** decreasing of the function f.

$\max \langle \nabla f(x), h \rangle \Leftrightarrow \min \langle \nabla f, h \rangle$

$\|h\|_2^2 = 1$

$L = -\nabla f(x)^T h + \lambda(\|h\|_2^2 - 1)$

$\frac{\partial L}{\partial h} = +\nabla f + 2\lambda h$

$h = -\nabla f \frac{1}{2\lambda}$

Direction of local steepest descent

Let's consider a linear approximation of the differentiable function f along some direction $h, \|h\|_2 = 1$:

$$f(x + \alpha h) = f(x) + \alpha \langle f'(x), h \rangle + o(\alpha)$$

We want h to be a decreasing direction:

$$f(x + \alpha h) < f(x)$$

$$f(x) + \alpha \langle f'(x), h \rangle + o(\alpha) < f(x)$$

and going to the limit at $\alpha \to 0$:

$$\langle f'(x), h \rangle \leq 0$$

Also from Cauchy–Bunyakovsky–Schwarz inequality:

$$|\langle f'(x), h \rangle| \leq \|f'(x)\|_2 \|h\|_2$$
$$\langle f'(x), h \rangle \geq -\|f'(x)\|_2 \|h\|_2 = -\|f'(x)\|_2$$

Thus, the direction of the antigradient

$$h = -\frac{f'(x)}{\|f'(x)\|_2}$$

gives the direction of the **steepest local** decreasing of the function f. The result of this method is

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

Gradient flow ODE

Let's consider the following ODE, which is referred as Gradient Flow equation.

$$\frac{dx}{dt} = -f'(x(t)) \tag{GF}$$

Gradient flow ODE

Let's consider the following ODE, which is referred as Gradient Flow equation.

$$\frac{dx}{dt} = -f'(x(t)) \tag{GF}$$

and discretize it on a uniform grid with α step:

$$\frac{x_{k+1} - x_k}{\alpha} = -f'(x_k),$$

дискретизация Эйлера

Gradient flow ODE

Let's consider the following ODE, which is referred as Gradient Flow equation.

$$\frac{dx}{dt} = -f'(x(t)) \qquad \text{(GF)}$$

and discretize it on a uniform grid with α step:

$$\frac{x_{k+1} - x_k}{\alpha} = -f'(x_k),$$

where $x_k \equiv x(t_k)$ and $\alpha = t_{k+1} - t_k$ - is the grid step.
From here we get the expression for x_{k+1}

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

which is exactly gradient descent.
Open In Colab ♣

дискретизация Эйлера ODE GF

Gradient flow ODE

Let's consider the following ODE, which is referred as Gradient Flow equation.

$$\frac{dx}{dt} = -f'(x(t)) \qquad \text{(GF)}$$

and discretize it on a uniform grid with α step:

$$\frac{x_{k+1} - x_k}{\alpha} = -f'(x_k),$$

where $x_k \equiv x(t_k)$ and $\alpha = t_{k+1} - t_k$ - is the grid step.
From here we get the expression for x_{k+1}

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

which is exactly gradient descent.
Open In Colab ♣

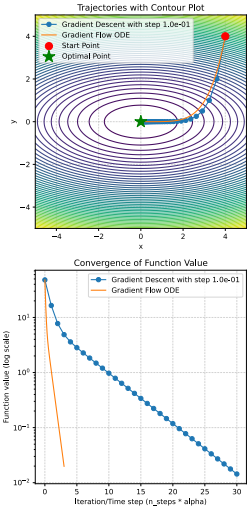


Figure 1: Gradient flow trajectory

Necessary local minimum condition

$$f'(x) = 0$$

$$-\eta f'(x) = 0 \quad +x$$

$$x - \eta f'(x) = x$$

$$x_k - \eta f'(x_k) = x_{k+1}$$

k $\quad$ $k+1$

Minimizer of Lipschitz parabola

If a function $f : \mathbb{R}^n \to \mathbb{R}$ is continuously differentiable and its gradient satisfies Lipschitz conditions with constant L, then $\forall x, y \in \mathbb{R}^n$:

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

∇f - L липшицев

$$\|\nabla f(x) - \nabla f(y)\| \leq L \cdot \|x - y\|$$

Minimizer of Lipschitz parabola

If a function $f : \mathbb{R}^n \to \mathbb{R}$ is continuously differentiable and its gradient satisfies Lipschitz conditions with constant L, then $\forall x, y \in \mathbb{R}^n$:

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

which geometrically means, that if we'll fix some point $x_0 \in \mathbb{R}^n$ and define two parabolas:

$$\phi_1(x) = f(x_0) + \langle \nabla f(x_0), x - x_0 \rangle - \frac{L}{2} \|x - x_0\|^2,$$
$$\phi_2(x) = f(x_0) + \langle \nabla f(x_0), x - x_0 \rangle + \frac{L}{2} \|x - x_0\|^2.$$

Minimizer of Lipschitz parabola

If a function $f : \mathbb{R}^n \to \mathbb{R}$ is continuously differentiable and its gradient satisfies Lipschitz conditions with constant L, then $\forall x, y \in \mathbb{R}^n$:

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

which geometrically means, that if we'll fix some point $x_0 \in \mathbb{R}^n$ and define two parabolas:

$$\phi_1(x) = f(x_0) + \langle \nabla f(x_0), x - x_0 \rangle - \frac{L}{2} \|x - x_0\|^2,$$
$$\phi_2(x) = f(x_0) + \langle \nabla f(x_0), x - x_0 \rangle + \frac{L}{2} \|x - x_0\|^2.$$

Then

$$\phi_1(x) \leq f(x) \leq \phi_2(x) \quad \forall x \in \mathbb{R}^n.$$

Minimizer of Lipschitz parabola

If a function $f : \mathbb{R}^n \to \mathbb{R}$ is continuously differentiable and its gradient satisfies Lipschitz conditions with constant L, then $\forall x, y \in \mathbb{R}^n$:

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

which geometrically means, that if we'll fix some point $x_0 \in \mathbb{R}^n$ and define two parabolas:

$$\phi_1(x) = f(x_0) + \langle \nabla f(x_0), x - x_0 \rangle - \frac{L}{2} \|x - x_0\|^2,$$
$$\phi_2(x) = f(x_0) + \langle \nabla f(x_0), x - x_0 \rangle + \frac{L}{2} \|x - x_0\|^2.$$

Then

$$\phi_1(x) \leq f(x) \leq \phi_2(x) \quad \forall x \in \mathbb{R}^n.$$

Now, if we have global upper bound on the function, in a form of parabola, we can try to go directly to its minimum.

Minimizer of Lipschitz parabola

If a function $f : \mathbb{R}^n \to \mathbb{R}$ is continuously differentiable and its gradient satisfies Lipschitz conditions with constant L, then $\forall x, y \in \mathbb{R}^n$:

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

which geometrically means, that if we'll fix some point $x_0 \in \mathbb{R}^n$ and define two parabolas:

$$\phi_1(x) = f(x_0) + \langle \nabla f(x_0), x - x_0 \rangle - \frac{L}{2}\|x - x_0\|^2,$$
$$\phi_2(x) = f(x_0) + \langle \nabla f(x_0), x - x_0 \rangle + \frac{L}{2}\|x - x_0\|^2.$$

Then

$$\phi_1(x) \leq f(x) \leq \phi_2(x) \quad \forall x \in \mathbb{R}^n.$$

Now, if we have global upper bound on the function, in a form of parabola, we can try to go directly to its minimum.

L-smooth function

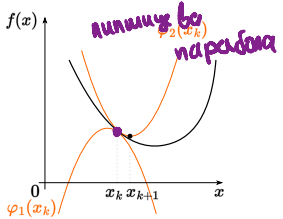


Figure 2: Illustration

$\nabla \varphi_2(x_{k+1}) = 0$

Minimizer of Lipschitz parabola

If a function $f : \mathbb{R}^n \to \mathbb{R}$ is continuously differentiable and its gradient satisfies Lipschitz conditions with constant L, then $\forall x, y \in \mathbb{R}^n$:

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

which geometrically means, that if we'll fix some point $x_0 \in \mathbb{R}^n$ and define two parabolas:

$$\phi_1(x) = f(x_0) + \langle \nabla f(x_0), x - x_0 \rangle - \frac{L}{2}\|x - x_0\|^2,$$

$$\phi_2(x) = f(x_0) + \langle \nabla f(x_0), x - x_0 \rangle + \frac{L}{2}\|x - x_0\|^2.$$

Then

$$\phi_1(x) \leq f(x) \leq \phi_2(x) \quad \forall x \in \mathbb{R}^n.$$

Now, if we have global upper bound on the function, in a form of parabola, we can try to go directly to its minimum.

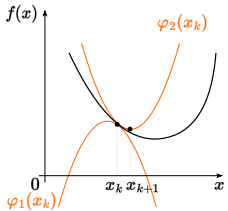


Figure 2: Illustration

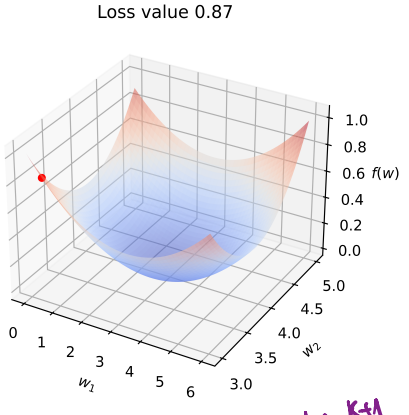
$$\nabla \phi_2(x) = 0$$
$$\nabla f(x_0) + L(x^{k+1} - x_0) = 0$$
$$x^* = x_0 - \frac{1}{L}\nabla f(x_0)$$
$$x_{k+1} = x_k - \frac{1}{L}\nabla f(x_k)$$

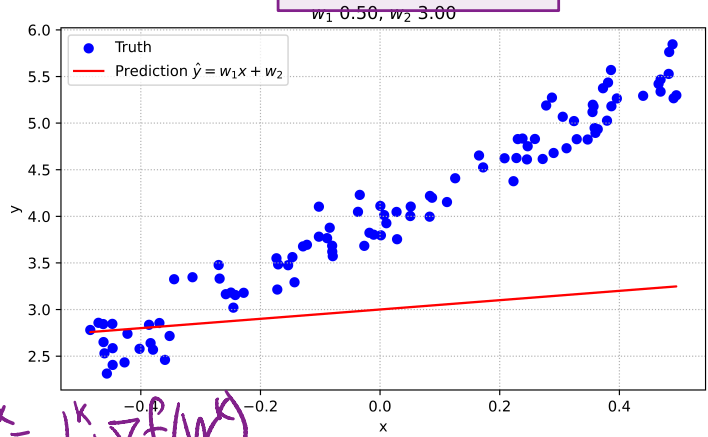
This way leads to the $\frac{1}{L}$ stepsize choosing. However, often the L constant is not known.

Convergence of Gradient Descent algorithm

$y = w_1 x + w_2$

Heavily depends on the choice of the learning rate α:





$w^{k+1} = w^k - \alpha^k \cdot \nabla f(w^k)$

Exact line search aka steepest descent

$$\alpha_k = \arg\min_{\alpha \in \mathbb{R}^+} f(x_{k+1}) = \arg\min_{\alpha \in \mathbb{R}^+} f(x_k - \alpha \nabla f(x_k))$$

More theoretical than practical approach. It also allows you to analyze the convergence, but often exact line search can be difficult if the function calculation takes too long or costs a lot. Interesting theoretical property of this method is that each following iteration is orthogonal to the previous one:

Метод наискорейшего спуска

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

Optimality conditions:

$$\nabla f(x_{k+1})^\top \nabla f(x_k) = 0$$

Пример задача найти. f(x)

$\frac{1}{2}\|Ax-b\|_2^2 \to \min_{x \in \mathbb{R}^n}$ к квадрат

$\nabla f = A^T(Ax-b)$

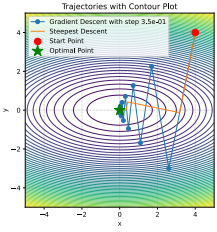


Figure 3: Steepest Descent

Open In Colab ♣

Exact line search aka steepest descent

$$\alpha_k = \arg\min_{\alpha \in \mathbb{R}^+} f(x_{k+1}) = \arg\min_{\alpha \in \mathbb{R}^+} f(x_k - \alpha \nabla f(x_k))$$

More theoretical than practical approach. It also allows you to analyze the convergence, but often exact line search can be difficult if the function calculation takes too long or costs a lot. Interesting theoretical property of this method is that each following iteration is orthogonal to the previous one:

$$\frac{\partial f}{\partial \alpha} = \frac{\partial f}{\partial x_{k+1}}^\top \cdot \frac{\partial x_{k+1}}{\partial \alpha} = 0$$

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

$$\left[A^\top (A x_{k+1} - b)\right]^\top A A x_k - b = 0$$

$$\left[A^\top (A(x_k - \alpha g_k) - b)\right]^\top g_k = 0$$

Optimality conditions:

$$\nabla f(x_{k+1})^\top \nabla f(x_k) = 0$$

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

$$\alpha_k = \arg\min_{\alpha} f(x_{k+1}) = \arg\min_{\alpha} f\left(x_k - \alpha \cdot A^\top (A x_k - b)\right)$$

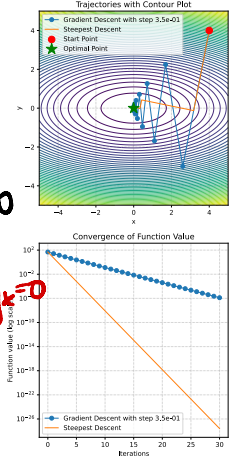


Figure 3: Steepest Descent

Open In Colab ♣

$$\left[A^T(Ax_{k+1}-b)\right]^T \underbrace{(Ax_k-b)}_{g_k}=0$$

$$\left[A^T(A(x_k-\alpha g_k)-b)\right]^T g_k=0$$

$$g_k^T\cdot A^T(Ax_k-\alpha Ag_k-b)=0$$

$$g_k^T A^T(Ax_k-b-\alpha Ag_k)=0$$

$$g_k^T g_k-\alpha g_k^T A^T Ag_k=0$$

$$\boxed{\alpha=\frac{g_k^T\cdot g_k}{g_k^T A^T Ag_k}}=\frac{g_k^T g_k}{g_k^T\nabla^2 f(x_k)g_k}$$

Exact line search aka steepest descent

$$\alpha_k = \arg\min_{\alpha \in \mathbb{R}^+} f(x_{k+1}) = \arg\min_{\alpha \in \mathbb{R}^+} f(x_k - \alpha \nabla f(x_k))$$

More theoretical than practical approach. It also allows you to analyze the convergence, but often exact line search can be difficult if the function calculation takes too long or costs a lot. Interesting theoretical property of this method is that each following iteration is orthogonal to the previous one:

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

$\nabla f = Ax_k$

Optimality conditions:

$$\nabla f(x_{k+1})^\top \nabla f(x_k) = 0$$

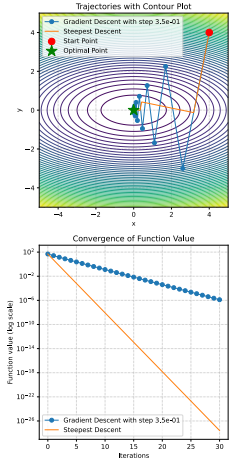


Figure 3: Steepest Descent

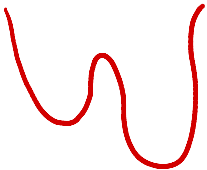
Open In Colab ♣

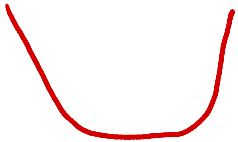
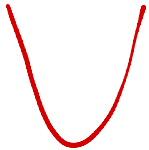
Convergence rates

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

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

smooth	convex	smooth & convex	smooth & strongly convex (or PL)
$\|\nabla f(x_k)\|^2 \approx \mathcal{O}\left(\frac{1}{k}\right)$	$f(x_k) - f^* \approx \mathcal{O}\left(\frac{1}{\sqrt{k}}\right)$	$f(x_k) - f^* \approx \mathcal{O}\left(\frac{1}{k}\right)$	$\|x_k - x^*\|^2 \approx \mathcal{O}\left(\left(1 - \frac{\mu}{L}\right)^k\right)$



Gradient Descent convergence. Smooth convex case

Факты [1] Пусть $f(x): \mathbb{R}^n \to \mathbb{R}^m$

Тогда f – L – Липшицево iff

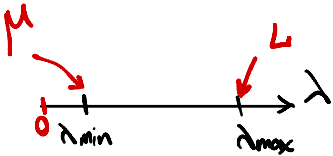
$\forall x \in \mathbb{R}^n : \|Df(x)\| \le L$

частный случай

если ∇f – L – Липшицев

$\|\nabla^2 f(x)\| \le L$

$\lambda_{max}(\nabla^2 f(x)) \le L$.



[2] Если $f: \mathbb{R}^n \to \mathbb{R}$ – L-гладкая (Липшицев ∇f с кон. Лип. L)

то $\forall x, y \in \mathbb{R}^n$: $f(y) \le f(x) + \langle \nabla f(x), y - x \rangle + \frac{L}{2}\|y - x\|^2$

[3] Сходимость GD в гладком выпуклом случае

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

т.к. f – L-гладкая

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

$f(x^{k+1}) \le f(x^k) - \alpha \cdot \|\nabla f(x^k)\|^2 + \frac{L}{2}\alpha^2 \|\nabla f(x^k)\|^2$

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

Gradient Descent convergence. Smooth convex case

2) $f(x)$ – выпуклая

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

$$\frac{1}{2}\alpha^2 L - \alpha \to \min_{\alpha} \to \alpha^{opt} = \frac{1}{L}$$

$$\begin{cases} y = x^* \\ x = x^k \end{cases} \to f(x^*) \geq f(x^k) + \nabla f(x^k)^T (x^* - x^k)$$

$$f(x^k) \leq f(x^*) + \nabla f(x^k)^T (x^k - x^*)$$

$$f(x^k) - f(x^*) \leq \nabla f(x^k)^T (x^k - x^*) \quad \text{conv}$$

$$\to f(x^{k+1}) \leq f(x^k) - \frac{1}{2L}\|\nabla f(x^k)\|^2 \quad \text{декремент}$$

$$f(x^k) - f(x^{k+1}) \geqslant \frac{1}{2L}\|\nabla f(x^k)\|^2$$

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

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

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

$$a^T a - b^T b = (a-b)^T (a+b)$$

$$\nabla f(x^k)^T \left(x^k - x^* - \frac{1}{2L}\nabla f(x^k)\right)$$

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

Gradient Descent convergence. Smooth convex case

$$\leq \nabla f(x^k)^T \left[x^k - x^* - \frac{1}{L} \nabla f(x^k) \right] =$$

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

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

$$\sum_{k=0}^{T-1} f(x^{k+1}) - f^* \leq \frac{L}{2} \left(\|x^0 - x^*\|^2 - \|x^T - x^*\|^2 \right) \leq$$

$\|x^0 - x^*\|^2 \leq R^2$

$$\sum_{k=0}^{T-1} \left(f(x^{k+1}) - f^* \right) \leq \frac{LR^2}{2} \quad | : T$$

$$\boxed{\sum_{k=0}^{T-1} \left(\frac{1}{T} f(x^{k+1}) - \frac{1}{T} f^* \right) \leq \frac{LR^2}{2T}}$$

$$f\left(\sum_{k=0}^{T-1} \frac{1}{T} \cdot x^{k+1} \right) \leq \frac{1}{T} \cdot \sum_{k=0}^{T-1} f(x^{k+1})$$

нер-во Йенсена: $f\left(\sum_i \theta_i x_i\right) \leq \sum_i \theta_i f(x_i)$

$$\sum_{k=0}^{T-1} \frac{1}{T} f(x^{k+1})$$

$f(\theta x + (1-\theta) y) \leq \theta f(x) + (1-\theta) f(y)$

$x_1, \ldots, x_T$

$\theta_i x_1 + \ldots + \theta_T x_T$

$\theta_i \geqslant 0$

$\sum \theta_i = 1$

$\theta_i = \frac{1}{T}$

$$f\left(\frac{1}{T} x_1 + \ldots + \frac{1}{T} x_T \right) \leq \frac{1}{T} \sum_{i=1}^{T} f(x_i)$$

$$\leq \boxed{\sum_{k=0}^{T-1} \left(\frac{1}{T} f(x^{k+1}) - \frac{1}{T} f^* \right) \leq \frac{LR^2}{2T}}$$

$$T \cdot f(\bar{x}^T) \leq f(x^1) + f(x^2) + \ldots + f(x^T)$$

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

субдифференциал ск-ть

Докажем, что GD сходится субдифференциально для гладких выпуклых функций

$\alpha = \frac{1}{L}$

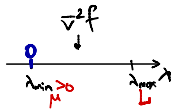
$$f(x) = \frac{1}{2} \|Ax - b\|^2$$

$$\nabla f = A^T (Ax - b) \rightarrow \nabla^2 f = A^T A$$

$$\|\nabla^2 f(x)\| \leq L$$

$$\boxed{\lambda_{max}(A^T A)}$$

$\nabla^2 f$



Gradient Descent convergence. Smooth μ-strongly convex case

Gradient Descent convergence. Polyak-Lojasiewicz case

④ Сильно выпуклый мягкий спуск:

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

$$\|x^{k+1} - x^*\|^2 = \|x^k - \alpha \nabla f(x^k) - x^*\|^2 =$$

$$= \|x^k - x^*\|^2 + \alpha^2 \|\nabla f(x^k)\|^2 - 2\alpha \nabla f(x^k)^T (x^k - x^*) \leq$$

$$\leq (1 - 2\mu)\|x^k - x^*\|^2 + \alpha^2 \|\nabla f(x^k)\|^2 - 2\alpha (f(x^k) - f^*) \quad (\leq)$$

сильная выпуклость: $f(x^*) \geq f(x^k) + \nabla f(x^k)^T \cdot (x^* - x^k) + \frac{\mu}{2}\|x^k - x^*\|^2$

$$\longrightarrow -2\alpha \cdot \nabla f(x^k)^T (x^k - x^*) \leq 2\alpha (f(x^*) - f(x^k)) - \alpha\mu \|x^k - x^*\|^2$$

PL: $f: \mathbb{R}^n \to \mathbb{R}$ – дифференцируема, $\mu > 0$: $\forall x \in \mathbb{R}^n$

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

$$f(x) - f^* \leq \frac{1}{2\mu} \|\nabla f(x)\|^2$$

Лемма: Если $f(x)$ – μ сильновыпукла, то она удовл. PL:

Пусть $x^* = \operatorname{argmin} f(x)$ $\quad f^* = f(x^*)$

Сильная выпуклость:

$\forall x, y \in \mathbb{R}^n \quad f(y) \geq f(x) + \nabla f(x)^T (y - x) + \frac{\mu}{2}\|y - x\|^2$

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

$$f(x) - f(x^*) \leq \nabla f(x)^T (x - x^*) - \frac{\mu}{2}\|x - x^*\|^2 =$$

$$= \left[\nabla f(x) - \frac{\mu}{2}(x - x^*)\right]^T (x - x^*) =$$

$$= \underbrace{\left[\frac{1}{\sqrt{\mu}} \nabla f(x) - \frac{\sqrt{\mu}}{2}(x - x^*)\right]^T}_{(a-b)^T} \underbrace{\sqrt{\mu}(x - x^*)}_{a+b} \quad (=)$$

$b = \sqrt{\mu}(x - x^*) - \frac{1}{\sqrt{\mu}}\nabla f(x)$ $\qquad a + b = \sqrt{\mu}(x - x^*)$

$a = \frac{1}{\sqrt{\mu}}\nabla f(x)$ $\qquad a - b = -\sqrt{\mu}(x - x^*) + \frac{2}{\sqrt{\mu}}\nabla f(x)$

$$\equiv \frac{1}{2}\left[\frac{1}{\mu}\|\nabla f(x)\|^2 - \left\|\sqrt{\mu}(x - x^*) - \frac{1}{\sqrt{\mu}}\nabla f(x)\right\|^2\right] \leq$$

$$\leq \frac{1}{2\mu}\|\nabla f(x)\|^2 \qquad \text{q.t.g.}$$

$$\|\nabla f(x^k)\|^2 \geq 2\mu \cdot (f(x^k) - f^*)$$

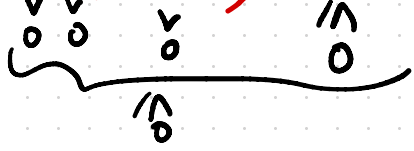
$$\leq (1 - \alpha\mu)\|x^k - x^*\|^2 + \alpha^2\|\nabla f(x^k)\|^2 - 2\alpha(f(x^k) - f^*) \leq$$

$$\leq (1 - \alpha\mu)\|x^k - x^*\|^2 + \alpha^2 \cdot 2L(f(x^k) - f^*) - 2\alpha(f(x^k) - f^*) =$$

$$= (1 - \alpha\mu)\|x^k - x^*\|^2 + 2\alpha(f(x^k) - f^*)[\alpha L - 1] =$$

$$= (1 - \alpha\mu)\|x^k - x^*\|^2 + 2\alpha L(f(x^k) - f^*)\left[\alpha - \frac{1}{L}\right] \leq$$

$\alpha \leq \frac{1}{L}$



$R^2 = \|x^0 - x^*\|^2$

$$\leq (1 - \alpha\mu)\|x^k - x^*\|^2$$

$1 - \alpha\mu < 1$

$$\|x^{k+1} - x^*\|^2 \leq (1 - \alpha\mu)\|x^k - x^*\|^2$$

$$\|x^k - x^*\|^2 \leq (1 - \alpha\mu)^k \cdot R^2$$

$\alpha\mu - 1 < 0$ $\qquad$ $1 - \alpha\mu > 0$

$-1 < 1 - \alpha\mu < 1$ $\qquad$ $-1 < 1 - \alpha\mu < 1$

$\alpha \le \frac{1}{L}$

$0 < 1 - \alpha\mu < 1$

$\alpha = \frac{1}{L}$

$1 - \alpha\mu > 0 \Rightarrow \alpha\mu < 1$

$\frac{\mu}{L} < 1$

$1 - \alpha\mu < 1$

$-\alpha\mu < 0$

$\alpha > 0 \quad \mu > 0$

$\alpha \le \frac{1}{L}$

(5) Сходимость GD для PL:

L - гладкость

$$f(x^{k+1}) \le f(x^k) + \nabla f(x^k)^T (x^{k+1} - x^k) + \frac{L}{2}\|x^{k+1} - x^k\|^2 =$$

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

$$= f(x^k) + \nabla f(x^k)^T (-\alpha \nabla f(x^k)) + \frac{L}{2}\alpha^2 \|\nabla f(x^k)\|^2 =$$

$$= f(x^k) - \alpha \|\nabla f(x^k)\|^2 + \frac{\alpha^2 L}{2}\|\nabla f(x^k)\|^2 =$$

$$= f(x^k) + \|\nabla f(x^k)\|^2 \left(\frac{\alpha^2 L}{2} - \alpha\right) =$$

$$= f(x^k) + \frac{\alpha}{2}\|\nabla f(x^k)\|^2 (\alpha L - 2) =$$

$\alpha \le \frac{2}{L}$

$\alpha L \le 2$

$$= f(x^k) - \frac{\alpha}{2}\|\nabla f(x^k)\|^2 (2 - \alpha L) \le$$

$$\leq f(x^k) - \frac{\alpha}{2} \cdot 2\mu\left(f(x^k) - f^*\right)(2 - \alpha L)$$

$$f(x^{k+1}) - f^* \leq f(x^k) - f^* - \alpha\mu\left(f(x^k) - f^*\right)(2 - \alpha L)$$

$$f(x^{k+1}) - f(x^*) \leq \left[1 - \alpha\mu(2 - \alpha L)\right] \cdot \left(f(x^k) - f^*\right)$$

$$f(x^k) - f^* \leq \left[1 - \alpha\mu(2 - \alpha L)\right]^k \cdot \left(f(x^0) - f^*\right)$$

$$-1 < 1 - \alpha\mu(2 - \alpha L) < 1$$

$$-2 < -\alpha\mu(2 - \alpha L) < 0$$

$$0 < \alpha\mu(2 - \alpha L) < 2$$

$$\alpha\mu(2 - \alpha L) < 2$$

$$\frac{1}{L}\mu(2 - 1) < 2$$

$$\frac{\mu}{L} \cdot 1 < 2$$

$$\frac{2}{L}\mu\underbrace{(2 - 2)}_{0} < 2$$

$$2 - \alpha L > 0$$

$$\alpha < \frac{2}{L}$$

$$\alpha = \frac{1}{L}$$

$$\alpha = \frac{2}{L}$$

$$\alpha = \frac{1}{L}$$