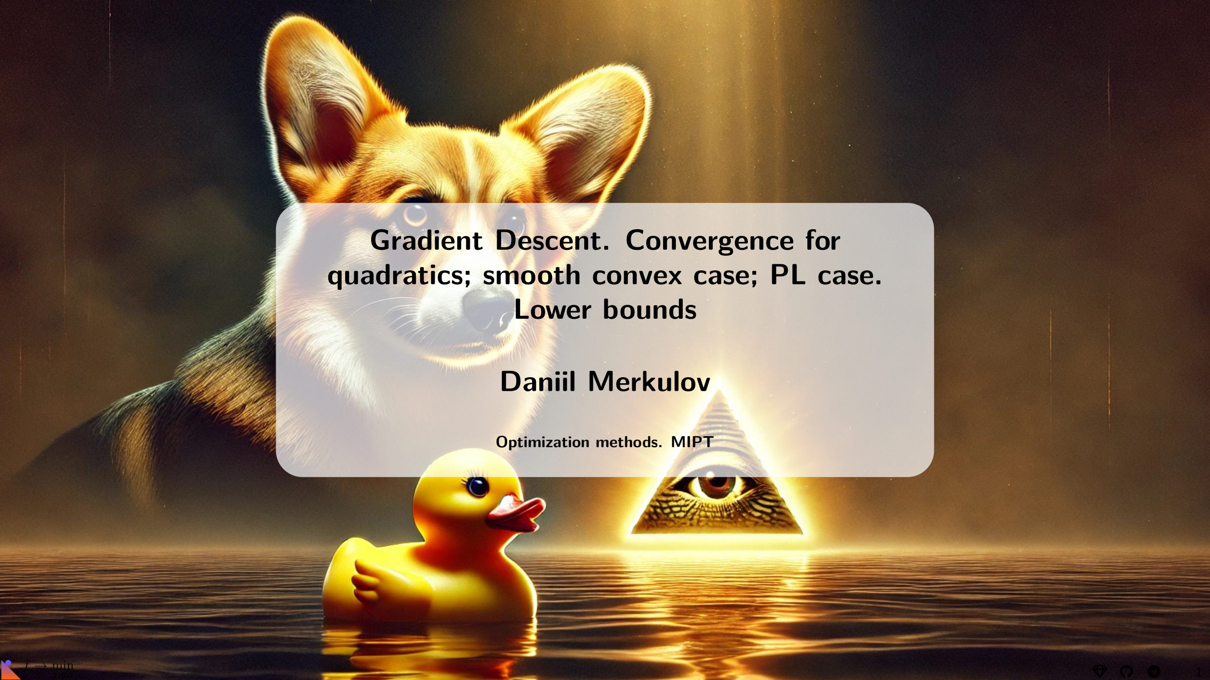


Gradient Descent. Convergence for quadratics; smooth convex case; PL case. Lower bounds

Daniil Merkulov

Optimization methods. MIPT

Gradient Descent

Direction of local steepest descent

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

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

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

$n \times 1$ $\quad$ $n \times 1$ $\quad$ 1×1 $\quad$ $n \times 1$

шаг

learning rate

Direction of local steepest descent

$f(x): R^n \to R$

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

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

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: выбираем h:

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

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

Direction of local steepest descent

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

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

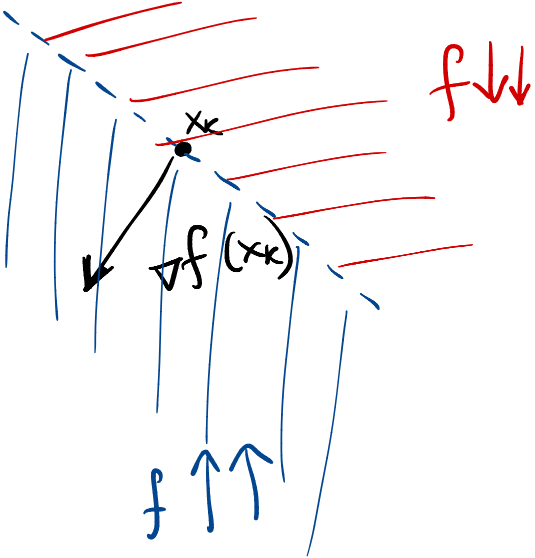
We want h to be a decreasing direction:

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

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

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

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



Direction of local steepest descent

$\|h\|_2 = 1$

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:

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

$|x| \leq 3$

$-3 \leq x \leq 3$

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

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

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

$$\langle \nabla f(x), h\rangle = \langle \nabla f(x), -\frac{\nabla f(x)}{\|\nabla f(x)\|}\rangle =$$

$$= -\frac{\|\nabla f(x)\|^2}{\|\nabla f(x)\|} = -\|\nabla f(x)\|$$

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:

$$\begin{aligned} |\langle f'(x), h \rangle| &\leq \|f'(x)\|_2 \|h\|_2 \\ \langle f'(x), h \rangle &\geq -\|f'(x)\|_2 \|h\|_2 = -\|f'(x)\|_2 \end{aligned}$$

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 to as the Gradient Flow equation.

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

$$\frac{dx}{dt}$$

$$dx = x_{k+1} - x_k$$

$$dt = t - 0$$

Gradient flow ODE

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

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

and discretize it on a uniform grid with α step:

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

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

явная схема дискретизации Эйлера

Gradient flow ODE

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

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

and discretize it on a uniform grid with α step:

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

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 ♣

Gradient flow ODE

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

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

and discretize it on a uniform grid with α step:

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

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 ♣

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

$$f = \frac{1}{2} x^T A x$$

$$f' = Ax$$

$$\frac{dx}{dt} = -A \cdot x(t)$$

пусть - скаляр   A - скаляр

$$\frac{dx}{dt} = -ax(t)$$

$$x(t) = \tilde{x} \cdot \exp(-at)$$

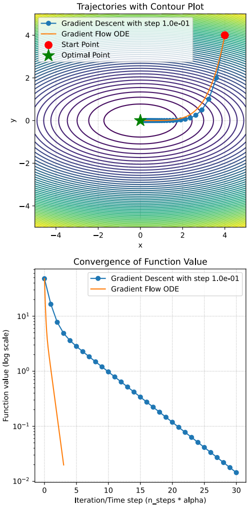
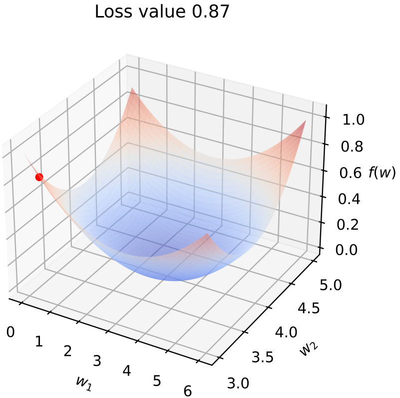
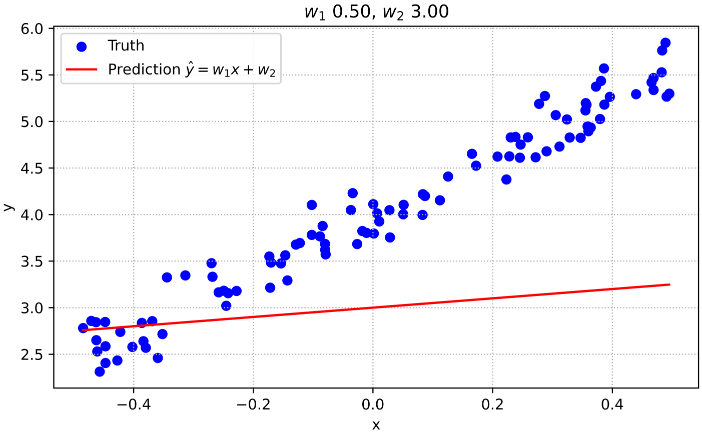


Figure 1: Gradient flow trajectory

Convergence of Gradient Descent algorithm

Heavily depends on the choice of the learning rate α:





Exact line search aka steepest descent

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

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. An interesting theoretical property of this method is that each following iteration is orthogonal to the previous one:

НАИСКОРЕЙШИЙ СПУСК:

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

$\text{argmin}_{\alpha} f(x_{k+1})$

задача: находясь в x_k выбрать лучшую длину шага α_k

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

Exact line search aka steepest descent

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

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.
An interesting theoretical property of this method is that each following iteration is orthogonal to the previous one:

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

$$\frac{\partial f(x_{k+1})}{\partial \alpha} = 0$$

Optimality conditions:

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

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

При наиск. спуске

$$\nabla f(x_{k+1}) \cdot \left(-\nabla f(x_k)\right) = 0$$

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

Exact line search aka steepest descent

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

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. An interesting theoretical property of this method is that each following iteration is orthogonal to the previous one:

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

Optimality conditions: $A \in \mathbb{S}^n_{++}$

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

Пусть $f(x) = \frac{1}{2} x^T A x$

$$x_{k+1} = x_k - \alpha_k \cdot A x_k$$

$$\alpha_k = \text{arg}\min_{\alpha} f(x_{k+1}) = \text{arg}\min \frac{1}{2} \overbrace{(x_k - \alpha A x_k)^T A (x_k - \alpha A x_k)}^{x_{k+1}(\alpha)}$$

$$\alpha = ?$$

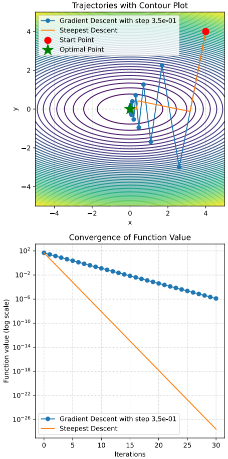


Figure 2: Steepest Descent

Open In Colab ♣

Exact line search aka steepest descent

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

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. An interesting theoretical property of this method is that each following iteration is orthogonal to the previous one:

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

Optimality conditions:

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

$$g_k = A x_k$$

$$(A(x_k - \alpha A x_k))^T A x_k = 0$$

$$(x_k - \alpha g_k)^T \cdot A^T g_k = 0$$

$$g_k^T$$

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

$$\Rightarrow \boxed{\alpha = \frac{g_k^T \cdot g_k}{g_k^T A g_k}}$$

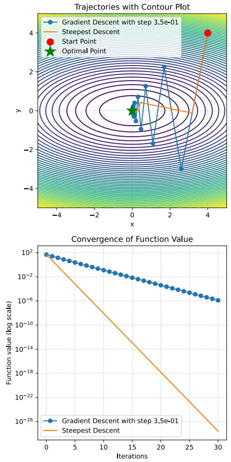


Figure 2: Steepest Descent

Open In Colab ♣

Strongly convex quadratics

Coordinate shift

Consider the following quadratic optimization problem:

$$\min_{x \in \mathbb{R}^d} f(x) = \min_{x \in \mathbb{R}^d} \frac{1}{2} x^\top A x - b^\top x + c, \text{ where } A \in \mathbb{S}^d_{++}.$$

Наша задача:
проанализир.
ex-ть GD
для

Coordinate shift

Consider the following quadratic optimization problem:

$$\min_{x \in \mathbb{R}^d} f(x) = \min_{x \in \mathbb{R}^d} \frac{1}{2} x^\top A x - b^\top x + c, \text{ where } A \in \mathbb{S}^d_{++}.$$

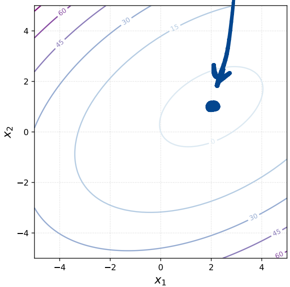
- Firstly, without loss of generality we can set $c = 0$, which will or affect optimization process.

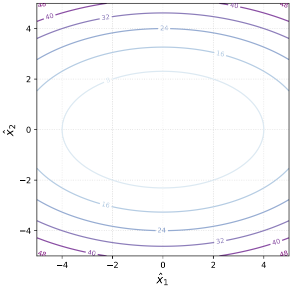
$$\nabla f(x^*) = 0$$

$$Ax^* - b = 0$$

$$\Rightarrow \boxed{x^* = A^{-1} b}$$

$x^* = ?$





Coordinate shift

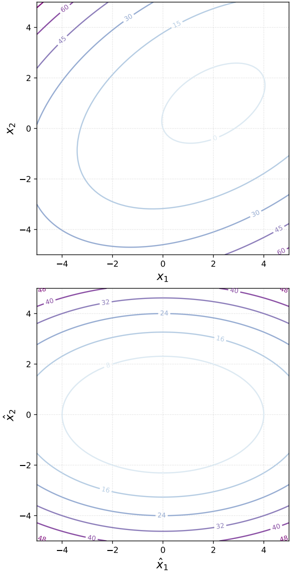
Consider the following quadratic optimization problem:

$$\min_{x \in \mathbb{R}^d} f(x) = \min_{x \in \mathbb{R}^d} \frac{1}{2} x^\top A x - b^\top x + c, \text{ where } A \in \mathbb{S}^d_{++}.$$

- Firstly, without loss of generality we can set $c = 0$, which will or affect optimization process.
- Secondly, we have a spectral decomposition of the matrix A:

$$\boxed{A = Q \Lambda Q^T}$$

собств. разл. матрицы

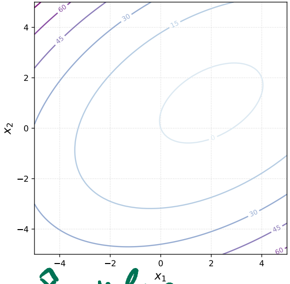


Coordinate shift

Consider the following quadratic optimization problem:

$$
\min_{x \in \mathbb{R}^d} f(x) = \min_{x \in \mathbb{R}^d} \frac{1}{2} x^\top A x - b^\top x + c, \text{ where } A \in \mathbb{S}^d_{++}.
$$

X – старые коорд



- Firstly, without loss of generality we can set $c = 0$, which will or affect optimization process.
- Secondly, we have a spectral decomposition of the matrix A:

$$
A = Q \Lambda Q^T
$$

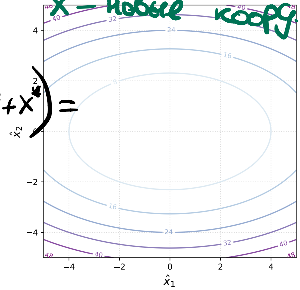
$QQ^T = I$

поворот

сдвиг в 0

- Let's show, that we can switch coordinates to make an analysis a little bit easier. Let $\boxed{\hat{x} = Q^T(x - x^*)}$, where x^* is the minimum point of initial function, defined by $Ax^* = b$. At the same time $\boxed{x = Q\hat{x} + x^*}$.

$\hat{x}$ – новые коорд



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

$=$

Coordinate shift

Consider the following quadratic optimization problem:

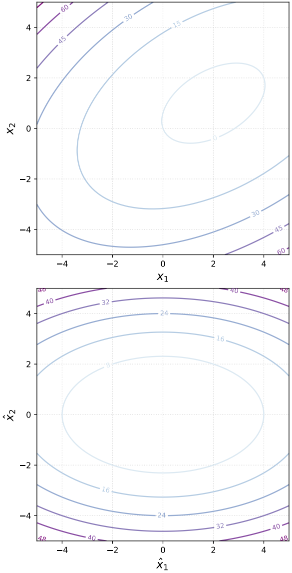
$$\min_{x \in \mathbb{R}^d} f(x) = \min_{x \in \mathbb{R}^d} \frac{1}{2} x^\top A x - b^\top x + c, \text{ where } A \in \mathbb{S}^d_{++}.$$

- Firstly, without loss of generality we can set $c = 0$, which will or affect optimization process.
- Secondly, we have a spectral decomposition of the matrix A:

$$A = Q \Lambda Q^T$$

- Let's show, that we can switch coordinates to make an analysis a little bit easier. Let $\hat{x} = Q^T(x - x^*)$, where x^* is the minimum point of initial function, defined by $Ax^* = b$. At the same time $x = Q\hat{x} + x^*$.

$$f(\hat{x}) = \frac{1}{2}(Q\hat{x} + x^*)^\top A (Q\hat{x} + x^*) - b^\top (Q\hat{x} + x^*)$$



Coordinate shift

Consider the following quadratic optimization problem:

$$
\min_{x \in \mathbb{R}^d} f(x) = \min_{x \in \mathbb{R}^d} \frac{1}{2} x^\top A x - b^\top x + c, \text{ where } A \in \mathbb{S}^d_{++}.
$$

- Firstly, without loss of generality we can set $c = 0$, which will or affect optimization process.
- Secondly, we have a spectral decomposition of the matrix A:

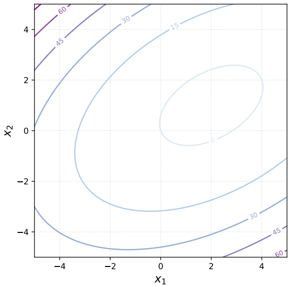
$$
A = Q \Lambda Q^T
$$

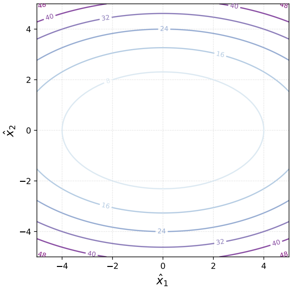
- Let's show, that we can switch coordinates to make an analysis a little bit easier. Let $\hat{x} = Q^T(x - x^*)$, where x^* is the minimum point of initial function, defined by $Ax^* = b$. At the same time $x = Q\hat{x} + x^*$.

$$
-\frac{1}{2}\hat{x}^T Q^T x^*
$$

$$
\begin{split}
f(\hat{x}) &= \frac{1}{2}(Q\hat{x} + x^*)^\top A (Q\hat{x} + x^*) - b^\top (Q\hat{x} + x^*) \\
&= \frac{1}{2}\hat{x}^T Q^T A Q \hat{x} + (x^*)^T A Q \hat{x} + \frac{1}{2}(x^*)^T A (x^*)^* - b^T Q \hat{x} - b^T x^*
\end{split}
$$

$$
\hat{x}^T Q^T A x^* = \hat{x}^T \underbrace{Q^T Q}_{I} \Lambda Q^T x^* = \hat{x}^T \Lambda Q^T x^*
$$





Coordinate shift

Consider the following quadratic optimization problem:

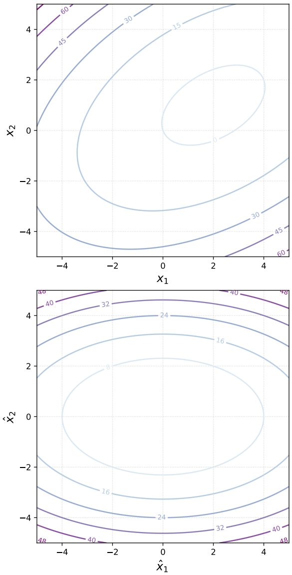
$$\min_{x\in\mathbb{R}^d} f(x) = \min_{x\in\mathbb{R}^d} \frac{1}{2}x^\top A x - b^\top x + c, \text{ where } A \in \mathbb{S}^d_{++}.$$

- Firstly, without loss of generality we can set $c = 0$, which will or affect optimization process.
- Secondly, we have a spectral decomposition of the matrix A:

$$A = Q\Lambda Q^T$$

- Let's show, that we can switch coordinates to make an analysis a little bit easier. Let $\hat{x} = Q^T(x - x^*)$, where x^* is the minimum point of initial function, defined by $Ax^* = b$. At the same time $x = Q\hat{x} + x^*$.

$$
\begin{aligned}
f(\hat{x}) &= \frac{1}{2}(Q\hat{x} + x^*)^\top A(Q\hat{x} + x^*) - b^\top(Q\hat{x} + x^*) \\
&= \frac{1}{2}\hat{x}^T \underbrace{Q^T A Q}_{=\Lambda}\hat{x} + (x^*)^T AQ\hat{x} + \frac{1}{2}(x^*)^T A(x^*)^T - b^T Q\hat{x} - b^T x^* \\
&= \frac{1}{2}\hat{x}^T\Lambda\hat{x} \quad (+\text{ const})
\end{aligned}
$$



Convergence analysis

Now we can work with the function $\boxed{f(x) = \frac{1}{2}x^T\Lambda x}$ with $x^* = 0$ without loss of generality (drop the hat from the $\hat{x}$)

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

Convergence analysis

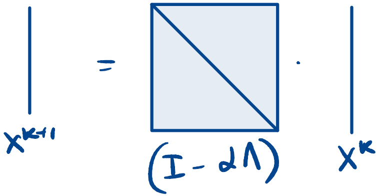
Now we can work with the function $f(x) = \frac{1}{2}x^T\Lambda x$ with $x^* = 0$ without loss of generality (drop the hat from the $\hat{x}$)

$$x^{k+1} = x^k - \alpha^k \nabla f(x^k) = x^k - \alpha^k \underbrace{\Lambda x^k}_{\nabla f(x_k)} = (I - \alpha\Lambda)\, x^k$$

Convergence analysis

Now we can work with the function $f(x) = \frac{1}{2}x^T\Lambda x$ with $x^* = 0$ without loss of generality (drop the hat from the $\hat{x}$)

$$
\begin{aligned}
\boxed{x^{k+1}} &= x^k - \alpha^k \nabla f(x^k) = x^k - \alpha^k \Lambda x^k \\
&= \boxed{(I - \alpha^k \Lambda)x^k}
\end{aligned}
$$



Convergence analysis

Now we can work with the function $f(x) = \frac{1}{2}x^T \Lambda x$ with $x^* = 0$ without loss of generality (drop the hat from the $\hat{x}$)

$$
\begin{aligned}
x^{k+1} &= x^k - \alpha^k \nabla f(x^k) = x^k - \alpha^k \Lambda x^k \\
&= (I - \alpha^k \Lambda) x^k
\end{aligned}
$$

$$\boxed{x^{k+1}_{(i)} = (1 - \alpha^k \lambda_{(i)}) x^k_{(i)}} \text{ For } i\text{-th coordinate}$$

Convergence analysis

Now we can work with the function $f(x) = \frac{1}{2}x^T\Lambda x$ with $x^* = 0$ without loss of generality (drop the hat from the $\hat{x}$)

$$
\begin{aligned}
x^{k+1} &= x^k - \alpha^k \nabla f(x^k) = x^k - \alpha^k \Lambda x^k \\
&= (I - \alpha^k \Lambda) x^k \\
x^{k+1}_{(i)} &= (1 - \alpha^k \lambda_{(i)}) x^k_{(i)} \text{ For } i\text{-th coordinate} \\
x^{k+1}_{(i)} &= (1 - \alpha^k \lambda_{(i)})^k x^0_{(i)}
\end{aligned}
$$

$\alpha_k = \text{const} = \alpha$

$$X_i^{k+1} = \left(1 - \alpha \lambda_{(i)}\right)^k \cdot X^0_{(i)}$$

Convergence analysis

Now we can work with the function $f(x) = \frac{1}{2}x^T\Lambda x$ with $x^* = 0$ without loss of generality (drop the hat from the $\hat{x}$)

$$
\begin{aligned}
x^{k+1} &= x^k - \alpha^k \nabla f(x^k) = x^k - \alpha^k \Lambda x^k \\
&= (I - \alpha^k \Lambda)x^k \\
x^{k+1}_{(i)} &= (1 - \alpha^k \lambda_{(i)})x^k_{(i)} \text{ For } i\text{-th coordinate} \\
x^{k+1}_{(i)} &= (1 - \alpha^k \lambda_{(i)})^k x^0_{(i)}
\end{aligned}
$$

СКАЛЯРЫ:

$$x^{k+1} = (1 - \alpha\lambda)^k \cdot x^0$$

$$|1 - \alpha\lambda| < 1$$

Let's use constant stepsize $\alpha^k = \alpha$. Convergence condition:

$$\boxed{\rho(\alpha) = \max_i |1 - \alpha\lambda_{(i)}| < 1}$$

Remember, that $\lambda_{\text{min}} = \mu > 0, \lambda_{\text{max}} = L \geq \mu$.

$$|1 - \alpha\mu| < 1$$
$$-1 < 1 - \alpha\mu < 1$$
$$-2 < -\alpha\mu < 0$$
$$0 < \alpha\mu < 2$$

$$|1 - \alpha L| < 1$$
$$-1 < 1 - \alpha L < 1$$
$$-2 < -\alpha L < 0$$
$$0 < \alpha L < 2$$

КАК ВЫБРАТЬ

α, ЧТО БЫ

$$\rho(\alpha) \to \min_{\alpha \in \mathbb{R}}$$

Convergence analysis

Now we can work with the function $f(x) = \frac{1}{2}x^T\Lambda x$ with $x^* = 0$ without loss of generality (drop the hat from the $\hat{x}$)

$$
\begin{aligned}
x^{k+1} &= x^k - \alpha^k \nabla f(x^k) = x^k - \alpha^k \Lambda x^k \\
&= (I - \alpha^k \Lambda) x^k \\
x^{k+1}_{(i)} &= (1 - \alpha^k \lambda_{(i)}) x^k_{(i)} \text{ For } i\text{-th coordinate} \\
x^{k+1}_{(i)} &= (1 - \alpha^k \lambda_{(i)})^k x^0_{(i)}
\end{aligned}
$$

Let's use constant stepsize $\alpha^k = \alpha$. Convergence condition:

$$
\rho(\alpha) = \max_i |1 - \alpha \lambda_{(i)}| < 1
$$

Remember, that $\lambda_{\text{min}} = \mu > 0, \lambda_{\text{max}} = L \geq \mu$.

$$
|1 - \alpha\mu| < 1
$$

Convergence analysis

Now we can work with the function $f(x) = \frac{1}{2}x^T\Lambda x$ with $x^* = 0$ without loss of generality (drop the hat from the $\hat{x}$)

$$
\begin{aligned}
x^{k+1} &= x^k - \alpha^k \nabla f(x^k) = x^k - \alpha^k \Lambda x^k \\
&= (I - \alpha^k \Lambda) x^k \\
x^{k+1}_{(i)} &= (1 - \alpha^k \lambda_{(i)}) x^k_{(i)} \text{ For } i\text{-th coordinate} \\
x^{k+1}_{(i)} &= (1 - \alpha^k \lambda_{(i)})^k x^0_{(i)}
\end{aligned}
$$

Let's use constant stepsize $\alpha^k = \alpha$. Convergence condition:

$$
\rho(\alpha) = \max_i |1 - \alpha\lambda_{(i)}| < 1
$$

Remember, that $\lambda_{\text{min}} = \mu > 0, \lambda_{\text{max}} = L \geq \mu$.

$$
\begin{aligned}
&|1 - \alpha\mu| < 1 \\
&-1 < 1 - \alpha\mu < 1
\end{aligned}
$$

Convergence analysis

Now we can work with the function $f(x) = \frac{1}{2}x^T\Lambda x$ with $x^* = 0$ without loss of generality (drop the hat from the $\hat{x}$)

$$
\begin{aligned}
x^{k+1} &= x^k - \alpha^k \nabla f(x^k) = x^k - \alpha^k \Lambda x^k \\
&= (I - \alpha^k \Lambda) x^k \\
x^{k+1}_{(i)} &= (1 - \alpha^k \lambda_{(i)}) x^k_{(i)} \text{ For } i\text{-th coordinate} \\
x^{k+1}_{(i)} &= (1 - \alpha^k \lambda_{(i)})^k x^0_{(i)}
\end{aligned}
$$

Let's use constant stepsize $\alpha^k = \alpha$. Convergence condition:

$$
\rho(\alpha) = \max_i |1 - \alpha \lambda_{(i)}| < 1
$$

Remember, that $\lambda_{\text{min}} = \mu > 0, \lambda_{\text{max}} = L \geq \mu$.

$$
\begin{aligned}
|1 - \alpha\mu| &< 1 \\
-1 < 1 - \alpha\mu &< 1 \\
\alpha < \frac{2}{\mu} \qquad & \alpha\mu > 0
\end{aligned}
$$

Convergence analysis

Now we can work with the function $f(x) = \frac{1}{2}x^T\Lambda x$ with $x^* = 0$ without loss of generality (drop the hat from the $\hat{x}$)

$$
\begin{aligned}
x^{k+1} &= x^k - \alpha^k \nabla f(x^k) = x^k - \alpha^k \Lambda x^k \\
&= (I - \alpha^k \Lambda) x^k \\
x^{k+1}_{(i)} &= (1 - \alpha^k \lambda_{(i)}) x^k_{(i)} \text{ For } i\text{-th coordinate} \\
x^{k+1}_{(i)} &= (1 - \alpha^k \lambda_{(i)})^k x^0_{(i)}
\end{aligned}
$$

Let's use constant stepsize $\alpha^k = \alpha$. Convergence condition:

$$
\rho(\alpha) = \max_i |1 - \alpha\lambda_{(i)}| < 1
$$

Remember, that $\lambda_{\text{min}} = \mu > 0, \lambda_{\text{max}} = L \geq \mu$.

$$
\begin{aligned}
|1 - \alpha\mu| &< 1 \\
-1 < 1 - \alpha\mu &< 1 \\
\alpha < \frac{2}{\mu} \quad & \quad \alpha\mu > 0
\end{aligned}
$$

$$
|1 - \alpha L| < 1
$$

Convergence analysis

Now we can work with the function $f(x) = \frac{1}{2}x^T\Lambda x$ with $x^* = 0$ without loss of generality (drop the hat from the $\hat{x}$)

$$
\begin{aligned}
x^{k+1} &= x^k - \alpha^k \nabla f(x^k) = x^k - \alpha^k \Lambda x^k \\
&= (I - \alpha^k \Lambda) x^k \\
x^{k+1}_{(i)} &= (1 - \alpha^k \lambda_{(i)}) x^k_{(i)} \text{ For } i\text{-th coordinate} \\
x^{k+1}_{(i)} &= (1 - \alpha^k \lambda_{(i)})^k x^0_{(i)}
\end{aligned}
$$

Let's use constant stepsize $\alpha^k = \alpha$. Convergence condition:

$$
\rho(\alpha) = \max_i |1 - \alpha \lambda_{(i)}| < 1
$$

Remember, that $\lambda_{\text{min}} = \mu > 0, \lambda_{\text{max}} = L \geq \mu$.

$$
\begin{aligned}
|1 - \alpha \mu| &< 1 \\
-1 < 1 - \alpha \mu &< 1 \\
\alpha < \frac{2}{\mu} \quad & \quad \alpha \mu > 0
\end{aligned}
\qquad
\begin{aligned}
|1 - \alpha L| &< 1 \\
-1 < 1 - \alpha L &< 1
\end{aligned}
$$

Convergence analysis

Now we can work with the function $f(x) = \frac{1}{2}x^T\Lambda x$ with $x^* = 0$ without loss of generality (drop the hat from the $\hat{x}$)

$$
\begin{aligned}
x^{k+1} &= x^k - \alpha^k \nabla f(x^k) = x^k - \alpha^k \Lambda x^k \\
&= (I - \alpha^k \Lambda) x^k \\
x^{k+1}_{(i)} &= (1 - \alpha^k \lambda_{(i)}) x^k_{(i)} \text{ For } i\text{-th coordinate} \\
x^{k+1}_{(i)} &= (1 - \alpha^k \lambda_{(i)})^k x^0_{(i)}
\end{aligned}
$$

ЗАКОН Б

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

Let's use constant stepsize $\alpha^k = \alpha$. Convergence condition:

$$\rho(\alpha) = \max_i |1 - \alpha \lambda_{(i)}| < 1$$

Remember, that $\lambda_{\text{min}} = \mu > 0, \lambda_{\text{max}} = L \geq \mu$.

$$
\begin{aligned}
|1 - \alpha\mu| &< 1 \\
-1 < 1 &- \alpha\mu < 1 \\
\boxed{\alpha < \frac{2}{\mu}} \quad & \quad \alpha\mu > 0
\end{aligned}
\qquad
\begin{aligned}
|1 - \alpha L| &< 1 \\
-1 < 1 &- \alpha L < 1 \\
\boxed{\alpha < \frac{2}{L}} \quad & \quad \alpha L > 0
\end{aligned}
$$

Convergence analysis

Now we can work with the function $f(x) = \frac{1}{2}x^T\Lambda x$ with $x^* = 0$ without loss of generality (drop the hat from the $\hat{x}$)

$$
\begin{aligned}
x^{k+1} &= x^k - \alpha^k \nabla f(x^k) = x^k - \alpha^k \Lambda x^k \\
&= (I - \alpha^k \Lambda) x^k \\
x^{k+1}_{(i)} &= (1 - \alpha^k \lambda_{(i)}) x^k_{(i)} \text{ For } i\text{-th coordinate} \\
x^{k+1}_{(i)} &= (1 - \alpha^k \lambda_{(i)})^k x^0_{(i)}
\end{aligned}
$$

Let's use constant stepsize $\alpha^k = \alpha$. Convergence condition:

$$
\rho(\alpha) = \max_i |1 - \alpha \lambda_{(i)}| < 1
$$

Remember, that $\lambda_{\text{min}} = \mu > 0, \lambda_{\text{max}} = L \geq \mu$.

$$
\begin{aligned}
&|1 - \alpha\mu| < 1 \\
&-1 < 1 - \alpha\mu < 1 \\
&\alpha < \frac{2}{\mu} \quad\quad \alpha\mu > 0
\end{aligned}
\qquad\qquad
\begin{aligned}
&|1 - \alpha L| < 1 \\
&-1 < 1 - \alpha L < 1 \\
&\alpha < \frac{2}{L} \quad\quad \alpha L > 0
\end{aligned}
$$

Convergence analysis

Now we can work with the function $f(x) = \frac{1}{2}x^T\Lambda x$ with $x^* = 0$ without loss of generality (drop the hat from the $\hat{x}$)

$$
\begin{aligned}
x^{k+1} &= x^k - \alpha^k \nabla f(x^k) = x^k - \alpha^k \Lambda x^k \\
&= (I - \alpha^k \Lambda) x^k \\
x^{k+1}_{(i)} &= (1 - \alpha^k \lambda_{(i)}) x^k_{(i)} \text{ For } i\text{-th coordinate} \\
x^{k+1}_{(i)} &= (1 - \alpha^k \lambda_{(i)})^k x^0_{(i)}
\end{aligned}
$$

Let's use constant stepsize $\alpha^k = \alpha$. Convergence condition:

$$
\rho(\alpha) = \max_i |1 - \alpha \lambda_{(i)}| < 1
$$

Remember, that $\lambda_{\text{min}} = \mu > 0, \lambda_{\text{max}} = L \geq \mu$.

$$
\begin{aligned}
&|1 - \alpha\mu| < 1 && |1 - \alpha L| < 1 \\
&-1 < 1 - \alpha\mu < 1 && -1 < 1 - \alpha L < 1 \\
&\alpha < \frac{2}{\mu} \quad \alpha\mu > 0 && \alpha < \frac{2}{L} \quad \alpha L > 0
\end{aligned}
$$

$\alpha < \frac{2}{L}$ is needed for convergence.

Convergence analysis

Now we can work with the function $f(x) = \frac{1}{2}x^T\Lambda x$ with $x^* = 0$ without loss of generality (drop the hat from the $\hat{x}$)

$$
\begin{aligned}
x^{k+1} &= x^k - \alpha^k \nabla f(x^k) = x^k - \alpha^k \Lambda x^k \\
&= (I - \alpha^k \Lambda) x^k \\
x^{k+1}_{(i)} &= (1 - \alpha^k \lambda_{(i)}) x^k_{(i)} \text{ For } i\text{-th coordinate} \\
x^{k+1}_{(i)} &= (1 - \alpha^k \lambda_{(i)})^k x^0_{(i)}
\end{aligned}
$$

Let's use constant stepsize $\alpha^k = \alpha$. Convergence condition:

$$
\rho(\alpha) = \max_i |1 - \alpha\lambda_{(i)}| < 1
$$

Remember, that $\lambda_{\text{min}} = \mu > 0, \lambda_{\text{max}} = L \geq \mu$.

$$
\begin{aligned}
|1 - \alpha\mu| &< 1 \\
-1 < 1 - \alpha\mu &< 1 \\
\alpha < \frac{2}{\mu} \quad & \quad \alpha\mu > 0
\end{aligned}
\qquad\qquad
\begin{aligned}
|1 - \alpha L| &< 1 \\
-1 < 1 - \alpha L &< 1 \\
\alpha < \frac{2}{L} \quad & \quad \alpha L > 0
\end{aligned}
$$

$\alpha < \frac{2}{L}$ is needed for convergence.

Now we would like to tune α to choose the best (lowest) convergence rate

$$
\rho^* = \min_\alpha \rho(\alpha)
$$

Convergence analysis

Now we can work with the function $f(x) = \frac{1}{2}x^T \Lambda x$ with $x^* = 0$ without loss of generality (drop the hat from the $\hat{x}$)

$$
\begin{aligned}
x^{k+1} &= x^k - \alpha^k \nabla f(x^k) = x^k - \alpha^k \Lambda x^k \\
&= (I - \alpha^k \Lambda) x^k \\
x^{k+1}_{(i)} &= (1 - \alpha^k \lambda_{(i)}) x^k_{(i)} \text{ For } i\text{-th coordinate} \\
x^{k+1}_{(i)} &= (1 - \alpha^k \lambda_{(i)})^k x^0_{(i)}
\end{aligned}
$$

Let's use constant stepsize $\alpha^k = \alpha$. Convergence condition:

$$
\rho(\alpha) = \max_i |1 - \alpha \lambda_{(i)}| < 1
$$

Remember, that $\lambda_{\text{min}} = \mu > 0, \lambda_{\text{max}} = L \geq \mu$.

$$
\begin{aligned}
|1 - \alpha \mu| &< 1 & |1 - \alpha L| &< 1 \\
-1 < 1 - \alpha \mu &< 1 & -1 < 1 - \alpha L &< 1 \\
\alpha < \frac{2}{\mu} \quad & \alpha \mu > 0 & \alpha < \frac{2}{L} \quad & \alpha L > 0
\end{aligned}
$$

$\alpha < \frac{2}{L}$ is needed for convergence.

Now we would like to tune α to choose the best (lowest) convergence rate

$$
\rho^* = \min_{\alpha} \underline{\rho(\alpha)} = \underline{\min_{\alpha} \max_i |1 - \alpha \lambda_{(i)}|}
$$

Convergence analysis

Now we can work with the function $f(x) = \frac{1}{2}x^T\Lambda x$ with $x^* = 0$ without loss of generality (drop the hat from the $\hat{x}$)

$$
\begin{aligned}
x^{k+1} &= x^k - \alpha^k \nabla f(x^k) = x^k - \alpha^k \Lambda x^k \\
&= (I - \alpha^k \Lambda) x^k \\
x^{k+1}_{(i)} &= (1 - \alpha^k \lambda_{(i)}) x^k_{(i)} \text{ For } i\text{-th coordinate} \\
x^{k+1}_{(i)} &= (1 - \alpha^k \lambda_{(i)})^k x^0_{(i)}
\end{aligned}
$$

Let's use constant stepsize $\alpha^k = \alpha$. Convergence condition:

$$
\rho(\alpha) = \max_i |1 - \alpha\lambda_{(i)}| < 1
$$

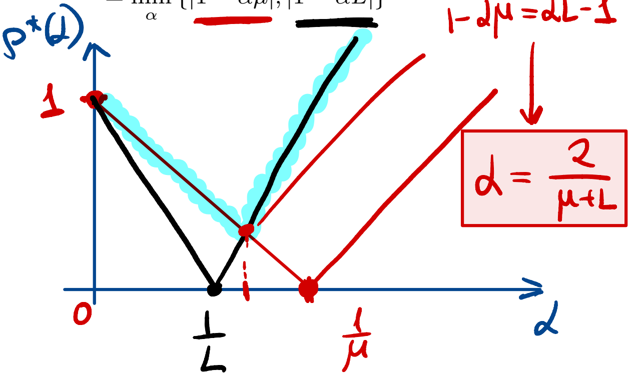
Remember, that $\lambda_{\text{min}} = \mu > 0, \lambda_{\text{max}} = L \geq \mu$.

$$
\begin{aligned}
&|1 - \alpha\mu| < 1 && |1 - \alpha L| < 1 \\
&-1 < 1 - \alpha\mu < 1 && -1 < 1 - \alpha L < 1 \\
&\alpha < \frac{2}{\mu} \quad \alpha\mu > 0 && \alpha < \frac{2}{L} \quad \alpha L > 0
\end{aligned}
$$

$\alpha < \frac{2}{L}$ is needed for convergence.

Now we would like to tune α to choose the best (lowest) convergence rate

$$
\begin{aligned}
\rho^* &= \min_\alpha \rho(\alpha) = \min_\alpha \max_i |1 - \alpha\lambda_{(i)}| \\
&= \min_\alpha \overset{\text{max}}{\left\{|1 - \alpha\mu|, |1 - \alpha L|\right\}}
\end{aligned}
$$



Convergence analysis

Now we can work with the function $f(x) = \frac{1}{2}x^T \Lambda x$ with $x^* = 0$ without loss of generality (drop the hat from the $\hat{x}$)

$$
\begin{aligned}
x^{k+1} &= x^k - \alpha^k \nabla f(x^k) = x^k - \alpha^k \Lambda x^k \\
&= (I - \alpha^k \Lambda) x^k \\
x^{k+1}_{(i)} &= (1 - \alpha^k \lambda_{(i)}) x^k_{(i)} \text{ For } i\text{-th coordinate} \\
x^{k+1}_{(i)} &= (1 - \alpha^k \lambda_{(i)})^k x^0_{(i)}
\end{aligned}
$$

Let's use constant stepsize $\alpha^k = \alpha$. Convergence condition:

$$
\rho(\alpha) = \max_i |1 - \alpha \lambda_{(i)}| < 1
$$

Remember, that $\lambda_{\text{min}} = \mu > 0, \lambda_{\text{max}} = L \geq \mu$.

$$
\begin{aligned}
&|1 - \alpha \mu| < 1 && |1 - \alpha L| < 1 \\
&-1 < 1 - \alpha \mu < 1 && -1 < 1 - \alpha L < 1 \\
&\alpha < \frac{2}{\mu} \qquad \alpha \mu > 0 && \alpha < \frac{2}{L} \qquad \alpha L > 0
\end{aligned}
$$

$\alpha < \frac{2}{L}$ is needed for convergence.

Now we would like to tune α to choose the best (lowest) convergence rate

$$
\begin{aligned}
\rho^* &= \min_\alpha \rho(\alpha) = \min_\alpha \max_i |1 - \alpha \lambda_{(i)}| \\
&= \min_\alpha \left\{ |1 - \alpha \mu|, |1 - \alpha L| \right\} \\
\alpha^* &: \quad 1 - \alpha^* \mu = \alpha^* L - 1
\end{aligned}
$$

Convergence analysis

Now we can work with the function $f(x) = \frac{1}{2}x^T\Lambda x$ with $x^* = 0$ without loss of generality (drop the hat from the $\hat{x}$)

$$
\begin{aligned}
x^{k+1} &= x^k - \alpha^k \nabla f(x^k) = x^k - \alpha^k \Lambda x^k \\
&= (I - \alpha^k \Lambda) x^k \\
x^{k+1}_{(i)} &= (1 - \alpha^k \lambda_{(i)}) x^k_{(i)} \text{ For } i\text{-th coordinate} \\
x^{k+1}_{(i)} &= (1 - \alpha^k \lambda_{(i)})^k x^0_{(i)}
\end{aligned}
$$

Let's use constant stepsize $\alpha^k = \alpha$. Convergence condition:

$$
\rho(\alpha) = \max_i |1 - \alpha \lambda_{(i)}| < 1
$$

Remember, that $\lambda_{\text{min}} = \mu > 0, \lambda_{\text{max}} = L \geq \mu$.

$$
\begin{aligned}
&|1 - \alpha\mu| < 1 && |1 - \alpha L| < 1 \\
&-1 < 1 - \alpha\mu < 1 && -1 < 1 - \alpha L < 1 \\
&\alpha < \frac{2}{\mu} \qquad \alpha\mu > 0 && \alpha < \frac{2}{L} \qquad \alpha L > 0
\end{aligned}
$$

$\alpha < \frac{2}{L}$ is needed for convergence.

Now we would like to tune α to choose the best (lowest) convergence rate

$$
\begin{aligned}
\rho^* &= \min_\alpha \rho(\alpha) = \min_\alpha \max_i |1 - \alpha \lambda_{(i)}| \\
&= \min_\alpha \left\{ |1 - \alpha\mu|, |1 - \alpha L| \right\} \\
\alpha^* &: \quad 1 - \alpha^*\mu = \alpha^* L - 1 \\
\alpha^* &= \frac{2}{\mu + L}
\end{aligned}
$$

Convergence analysis

Now we can work with the function $f(x) = \frac{1}{2}x^T \Lambda x$ with $x^* = 0$ without loss of generality (drop the hat from the $\hat{x}$)

$$
\begin{aligned}
x^{k+1} &= x^k - \alpha^k \nabla f(x^k) = x^k - \alpha^k \Lambda x^k \\
&= (I - \alpha^k \Lambda) x^k \\
x^{k+1}_{(i)} &= (1 - \alpha^k \lambda_{(i)}) x^k_{(i)} \text{ For } i\text{-th coordinate} \\
x^{k+1}_{(i)} &= (1 - \alpha^k \lambda_{(i)})^k x^0_{(i)}
\end{aligned}
$$

Let's use constant stepsize $\alpha^k = \alpha$. Convergence condition:

$$
\rho(\alpha) = \max_i |1 - \alpha \lambda_{(i)}| < 1
$$

Remember, that $\lambda_{\text{min}} = \mu > 0, \lambda_{\text{max}} = L \geq \mu$.

$$
\begin{aligned}
&|1 - \alpha\mu| < 1 && |1 - \alpha L| < 1 \\
&-1 < 1 - \alpha\mu < 1 && -1 < 1 - \alpha L < 1 \\
&\alpha < \frac{2}{\mu} \quad \alpha\mu > 0 && \alpha < \frac{2}{L} \quad \alpha L > 0
\end{aligned}
$$

$\alpha < \frac{2}{L}$ is needed for convergence.

Now we would like to tune α to choose the best (lowest) convergence rate

$$
\begin{aligned}
\rho^* &= \min_\alpha \rho(\alpha) = \min_\alpha \max_i |1 - \alpha \lambda_{(i)}| \\
&= \min_\alpha \left\{ |1 - \alpha\mu|, |1 - \alpha L| \right\}
\end{aligned}
$$

$$
\alpha^* : \quad 1 - \alpha^* \mu = \alpha^* L - 1
$$

$$
\boxed{\alpha^* = \frac{2}{\mu + L}} \quad \boxed{\rho^* = \frac{L - \mu}{L + \mu}} = \frac{\frac{L}{\mu} - 1}{\frac{L}{\mu} + 1} = \frac{\varkappa - 1}{\varkappa + 1}
$$

Convergence analysis

Now we can work with the function $f(x) = \frac{1}{2}x^T\Lambda x$ with $x^* = 0$ without loss of generality (drop the hat from the $\hat{x}$)

$$
\begin{aligned}
x^{k+1} &= x^k - \alpha^k \nabla f(x^k) = x^k - \alpha^k \Lambda x^k \\
&= (I - \alpha^k \Lambda) x^k \\
x^{k+1}_{(i)} &= (1 - \alpha^k \lambda_{(i)}) x^k_{(i)} \text{ For } i\text{-th coordinate} \\
x^{k+1}_{(i)} &= (1 - \alpha^k \lambda_{(i)})^k x^0_{(i)}
\end{aligned}
$$

Let's use constant stepsize $\alpha^k = \alpha$. Convergence condition:

$$
\rho(\alpha) = \max_i |1 - \alpha\lambda_{(i)}| < 1
$$

Remember, that $\lambda_{\text{min}} = \mu > 0, \lambda_{\text{max}} = L \geq \mu$.

$$
\begin{aligned}
&|1 - \alpha\mu| < 1 && |1 - \alpha L| < 1 \\
&-1 < 1 - \alpha\mu < 1 && -1 < 1 - \alpha L < 1 \\
&\alpha < \frac{2}{\mu} \quad\quad \alpha\mu > 0 && \alpha < \frac{2}{L} \quad\quad \alpha L > 0
\end{aligned}
$$

$\alpha < \frac{2}{L}$ is needed for convergence.

Now we would like to tune α to choose the best (lowest) convergence rate

$$
\begin{aligned}
\rho^* &= \min_\alpha \rho(\alpha) = \min_\alpha \max_i |1 - \alpha\lambda_{(i)}| \\
&= \min_\alpha \left\{ |1 - \alpha\mu|, |1 - \alpha L| \right\} \\
\alpha^* &: \quad 1 - \alpha^*\mu = \alpha^* L - 1 \\
\alpha^* &= \frac{2}{\mu + L} \quad\quad \rho^* = \frac{L - \mu}{L + \mu} \\
x^{k+1} &= \left(\frac{L - \mu}{L + \mu} \right)^k x^0
\end{aligned}
$$

Convergence analysis

Now we can work with the function $f(x) = \frac{1}{2}x^T\Lambda x$ with $x^* = 0$ without loss of generality (drop the hat from the $\hat{x}$)

$$
\begin{aligned}
x^{k+1} &= x^k - \alpha^k \nabla f(x^k) = x^k - \alpha^k \Lambda x^k \\
&= (I - \alpha^k \Lambda) x^k \\
x^{k+1}_{(i)} &= (1 - \alpha^k \lambda_{(i)}) x^k_{(i)} \text{ For } i\text{-th coordinate} \\
x^{k+1}_{(i)} &= (1 - \alpha^k \lambda_{(i)})^k x^0_{(i)}
\end{aligned}
$$

Let's use constant stepsize $\alpha^k = \alpha$. Convergence condition:

$$
\rho(\alpha) = \max_i |1 - \alpha\lambda_{(i)}| < 1
$$

Remember, that $\lambda_{\text{min}} = \mu > 0, \lambda_{\text{max}} = L \geq \mu$.

$$
\begin{aligned}
|1 - \alpha\mu| &< 1 & \qquad |1 - \alpha L| &< 1 \\
-1 < 1 - \alpha\mu &< 1 & \qquad -1 < 1 - \alpha L &< 1 \\
\alpha < \frac{2}{\mu} \qquad \alpha\mu &> 0 & \qquad \alpha < \frac{2}{L} \qquad \alpha L &> 0
\end{aligned}
$$

$\alpha < \frac{2}{L}$ is needed for convergence.

Now we would like to tune α to choose the best (lowest) convergence rate

$$
\begin{aligned}
\rho^* &= \min_\alpha \rho(\alpha) = \min_\alpha \max_i |1 - \alpha\lambda_{(i)}| \\
&= \min_\alpha \{|1 - \alpha\mu|, |1 - \alpha L|\} \\
\alpha^* &: \quad 1 - \alpha^*\mu = \alpha^* L - 1 \\
\alpha^* &= \frac{2}{\mu + L} \qquad \rho^* = \frac{L - \mu}{L + \mu}
\end{aligned}
$$

$$
\|x^{k+1}\| = \left(\frac{L - \mu}{L + \mu}\right)^k \|x^0\| \qquad f(x^{k+1}) = \left(\frac{L - \mu}{L + \mu}\right)^{2k} f(x^0)
$$

Convergence analysis

So, we have a linear convergence in the domain with rate $\frac{\kappa-1}{\kappa+1} = 1 - \frac{2}{\kappa+1}$, where $\kappa = \frac{L}{\mu}$ is sometimes called *condition number* of the quadratic problem.

κ	ρ	Iterations to decrease domain gap 10 times	Iterations to decrease function gap 10 times
1.1	0.05	1	1
2	0.33	3	2
5	0.67	6	3
10	0.82	12	6
50	0.96	58	29
100	0.98	116	58
500	0.996	576	288
1000	0.998	1152	576

Polyak-Lojasiewicz smooth case

Polyak-Lojasiewicz condition. Linear convergence of gradient descent without convexity

условие градиентного доминирования

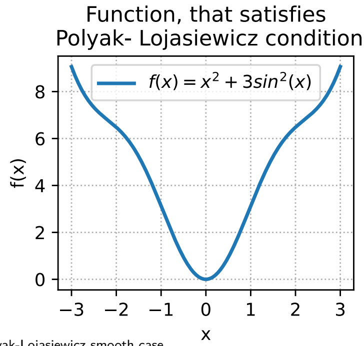
PL inequality holds if the following condition is satisfied for some $\mu > 0$,

$$\boxed{\|\nabla f(x)\|^2 \geq 2\mu(f(x) - f^*)} \quad \forall x$$

It is interesting, that the Gradient Descent algorithm might converge linearly even without convexity.

The following functions satisfy the PL condition but are not convex. Link to the code

$$f(x) = x^2 + 3\sin^2(x)$$



Polyak-Lojasiewicz condition. Linear convergence of gradient descent without convexity

PL inequality holds if the following condition is satisfied for some $\mu > 0$,

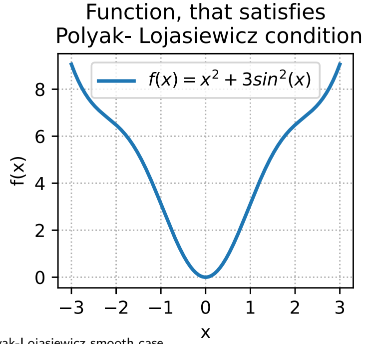
$$\|\nabla f(x)\|^2 \geq 2\mu(f(x) - f^*) \quad \forall x$$

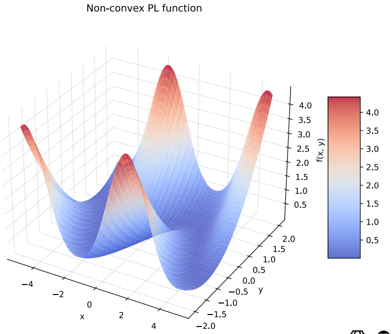
It is interesting, that the Gradient Descent algorithm might converge linearly even without convexity.

The following functions satisfy the PL condition but are not convex. Link to the code

$$f(x) = x^2 + 3\sin^2(x)$$

$$f(x,y) = \frac{(y - \sin x)^2}{2}$$





Convergence analysis

Theorem

Consider the Problem

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

and assume that f is μ-Polyak-Lojasiewicz and L-smooth, for some $L \geq \mu > 0$.
Consider $(x^k)_{k \in \mathbb{N}}$ a sequence generated by the gradient descent constant stepsize algorithm, with a stepsize satisfying $0 < \alpha \leq \frac{1}{L}$. Then:

$$\boxed{f(x^k) - f^* \leq (1 - \alpha\mu)^k (f(x^0) - f^*).}$$

Convergence analysis

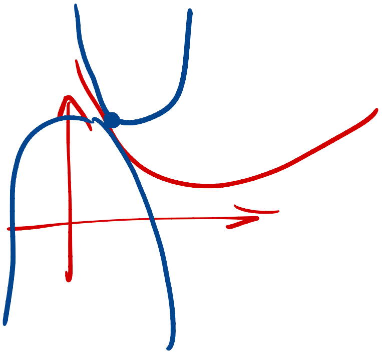
We can use L-smoothness, together with the update rule of the algorithm, to write

гладкость

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

липшицева парабола

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



Convergence analysis

We can use L-smoothness, together with the update rule of the algorithm, to write

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

подставляем GD

Convergence analysis

We can use L-smoothness, together with the update rule of the algorithm, to write

$$
\begin{aligned}
f(x^{k+1}) &\leq f(x^k) + \langle \nabla f(x^k), x^{k+1} - x^k \rangle + \frac{L}{2}\|x^{k+1} - x^k\|^2 \\
&= f(x^k) - \alpha\|\nabla f(x^k)\|^2 + \frac{L\alpha^2}{2}\|\nabla f(x^k)\|^2 \\
&= f(x^k) - \frac{\alpha}{2}\left(2 - L\alpha\right)\|\nabla f(x^k)\|^2
\end{aligned}
$$

выносим $\|\nabla f(x_k)\|^2$ за скобки

Convergence analysis

We can use L-smoothness, together with the update rule of the algorithm, to write

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

пусть $\alpha \leq \frac{1}{L}$

$\alpha L \leq 1$

$\frac{\alpha}{2}(\alpha L - 2) \leq$

$\leq \frac{\alpha}{2}(1-2) \leq -\frac{\alpha}{2}$

Convergence analysis

We can use L-smoothness, together with the update rule of the algorithm, to write

$$
\begin{aligned}
f(x^{k+1}) &\leq f(x^k) + \langle \nabla f(x^k), x^{k+1} - x^k \rangle + \frac{L}{2}\|x^{k+1} - x^k\|^2 \\
&= f(x^k) - \alpha\|\nabla f(x^k)\|^2 + \frac{L\alpha^2}{2}\|\nabla f(x^k)\|^2 \\
&= f(x^k) - \frac{\alpha}{2}\left(2 - L\alpha\right)\|\nabla f(x^k)\|^2 \\
&\leq f(x^k) - \frac{\alpha}{2}\|\nabla f(x^k)\|^2,
\end{aligned}
$$

Convergence analysis

We can use L-smoothness, together with the update rule of the algorithm, to write

$$
\begin{aligned}
f(x^{k+1}) &\leq f(x^k) + \langle \nabla f(x^k), x^{k+1} - x^k \rangle + \frac{L}{2}\|x^{k+1} - x^k\|^2 \\
&= f(x^k) - \alpha\|\nabla f(x^k)\|^2 + \frac{L\alpha^2}{2}\|\nabla f(x^k)\|^2 \\
&= f(x^k) - \frac{\alpha}{2}\left(2 - L\alpha\right)\|\nabla f(x^k)\|^2 \\
&\leq f(x^k) - \frac{\alpha}{2}\|\nabla f(x^k)\|^2,
\end{aligned}
$$

where in the last inequality we used our hypothesis on the stepsize that $\alpha L \leq 1$.

Convergence analysis

We can use L-smoothness, together with the update rule of the algorithm, to write

$$
\begin{aligned}
f(x^{k+1}) &\leq f(x^k) + \langle \nabla f(x^k), x^{k+1} - x^k \rangle + \frac{L}{2}\|x^{k+1} - x^k\|^2 \\
&= f(x^k) - \alpha\|\nabla f(x^k)\|^2 + \frac{L\alpha^2}{2}\|\nabla f(x^k)\|^2 \\
&= f(x^k) - \frac{\alpha}{2}\left(2 - L\alpha\right)\|\nabla f(x^k)\|^2 \\
&\leq f(x^k) - \frac{\alpha}{2}\|\nabla f(x^k)\|^2,
\end{aligned}
$$

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

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

where in the last inequality we used our hypothesis on the stepsize that $\alpha L \leq 1$.

We can now use the Polyak-Lojasiewicz property to write:

$-f^*$ $\quad$ $-f^*$

$$
\boxed{f(x^{k+1}) \leq f(x^k) - \alpha\mu(f(x^k) - f^*).}
$$

$f(x^{k+1}) - f^* \leq$

$\leq f(x_k) - f^* - \alpha\mu(f(x_k) - f^*)$

The conclusion follows after subtracting f^* on both sides of this inequality and using recursion.

$= (f(x_k) - f^*)(1 - \alpha\mu)$

Any μ-strongly convex differentiable function is a PL-function

PL

μ

Theorem

If a function $f(x)$ is differentiable and μ-strongly convex, then it is a PL function.

Proof

$\forall$ μ-сильно вып $\Rightarrow$ PL функция

By first order strong convexity criterion:

$$
f(y) \geq f(x) + \nabla f(x)^T (y - x) + \frac{\mu}{2} \|y - x\|_2^2
$$

$f(x^*) = f^*$

Putting $y = x^*$:

$$
f(x^*) \geq f(x) + \nabla f(x)^T (x^* - x) + \frac{\mu}{2} \|x^* - x\|_2^2
$$

Any μ-strongly convex differentiable function is a PL-function

Theorem

If a function $f(x)$ is differentiable and μ-strongly convex, then it is a PL function.

Proof

By first order strong convexity criterion:

$$f(y) \geq f(x) + \nabla f(x)^T (y - x) + \frac{\mu}{2}\|y - x\|_2^2$$

Putting $y = x^*$:

$$f(x^*) \geq f(x) + \nabla f(x)^T (x^* - x) + \frac{\mu}{2}\|x^* - x\|_2^2$$

$$f(x) - f(x^*) \leq \nabla f(x)^T (x - x^*) - \frac{\mu}{2}\|x^* - x\|_2^2 =$$

Any μ-strongly convex differentiable function is a PL-function

Theorem

If a function $f(x)$ is differentiable and μ-strongly convex, then it is a PL function.

Proof

By first order strong convexity criterion:

$$
f(y) \geq f(x) + \nabla f(x)^T(y-x) + \frac{\mu}{2}\|y-x\|_2^2
$$

Putting $y = x^*$:

$$
\begin{split}
f(x^*) &\geq f(x) + \nabla f(x)^T(x^*-x) + \frac{\mu}{2}\|x^*-x\|_2^2 \\
f(x) - f(x^*) &\leq \nabla f(x)^T(x-x^*) - \frac{\mu}{2}\|x^*-x\|_2^2 = \\
&= \left(\nabla f(x) - \frac{\mu}{2}(x^*-x)\right)^T(x-x^*) =
\end{split}
$$

Any μ-strongly convex differentiable function is a PL-function

Theorem

If a function $f(x)$ is differentiable and μ-strongly convex, then it is a PL function.

Proof

By first order strong convexity criterion:

$$
f(y) \geq f(x) + \nabla f(x)^T(y-x) + \frac{\mu}{2}\|y-x\|_2^2
$$

Putting $y = x^*$:

$$
\begin{aligned}
f(x^*) &\geq f(x) + \nabla f(x)^T(x^*-x) + \frac{\mu}{2}\|x^*-x\|_2^2 \\
f(x) - f(x^*) &\leq \nabla f(x)^T(x-x^*) - \frac{\mu}{2}\|x^*-x\|_2^2 = \\
&= \left(\nabla f(x)^T - \frac{\mu}{2}(x^*-x)\right)^T(x-x^*) = \\
&= \frac{1}{2}\left(\frac{2}{\sqrt{\mu}}\nabla f(x)^T - \sqrt{\mu}(x^*-x)\right)^T\sqrt{\mu}(x-x^*) =
\end{aligned}
$$

Any μ-strongly convex differentiable function is a PL-function

Theorem

If a function $f(x)$ is differentiable and μ-strongly convex, then it is a PL function.

Proof

By first order strong convexity criterion:

$$f(y) \geq f(x) + \nabla f(x)^T (y - x) + \frac{\mu}{2} \|y - x\|_2^2$$

Putting $y = x^*$:

$$
\begin{aligned}
f(x^*) &\geq f(x) + \nabla f(x)^T (x^* - x) + \frac{\mu}{2} \|x^* - x\|_2^2 \\
f(x) - f(x^*) &\leq \nabla f(x)^T (x - x^*) - \frac{\mu}{2} \|x^* - x\|_2^2 = \\
&= \left(\nabla f(x)^T - \frac{\mu}{2}(x^* - x) \right)^T (x - x^*) = \\
&= \frac{1}{2} \left(\frac{2}{\sqrt{\mu}} \nabla f(x)^T - \sqrt{\mu}(x^* - x) \right)^T \sqrt{\mu}(x - x^*) =
\end{aligned}
$$

Any μ-strongly convex differentiable function is a PL-function

Theorem

If a function $f(x)$ is differentiable and μ-strongly convex, then it is a PL function.

Proof

By first order strong convexity criterion:

$$f(y) \geq f(x) + \nabla f(x)^T (y - x) + \frac{\mu}{2}\|y - x\|_2^2$$

Putting $y = x^*$:

$$f(x^*) \geq f(x) + \nabla f(x)^T (x^* - x) + \frac{\mu}{2}\|x^* - x\|_2^2$$

$$\begin{aligned}
f(x) - f(x^*) &\leq \nabla f(x)^T (x - x^*) - \frac{\mu}{2}\|x^* - x\|_2^2 = \\
&= \left(\nabla f(x)^T - \frac{\mu}{2}(x^* - x)\right)^T (x - x^*) = \\
&= \frac{1}{2}\underbrace{\left(\frac{2}{\sqrt{\mu}}\nabla f(x)^T - \sqrt{\mu}(x^* - x)\right)^T}_{a - b} \underbrace{\sqrt{\mu}(x - x^*)}_{b + a} =
\end{aligned}$$

Let $\boxed{a = \frac{1}{\sqrt{\mu}}\nabla f(x)}$ and

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

$= a^2 - b^2$

Any μ-strongly convex differentiable function is a PL-function

Theorem

If a function $f(x)$ is differentiable and μ-strongly convex, then it is a PL function.

Proof

By first order strong convexity criterion:

$$f(y) \geq f(x) + \nabla f(x)^T (y - x) + \frac{\mu}{2}\|y - x\|_2^2$$

Putting $y = x^*$:

$$f(x^*) \geq f(x) + \nabla f(x)^T (x^* - x) + \frac{\mu}{2}\|x^* - x\|_2^2$$

$$f(x) - f(x^*) \leq \nabla f(x)^T (x - x^*) - \frac{\mu}{2}\|x^* - x\|_2^2 =$$

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

$$= \frac{1}{2}\left(\frac{2}{\sqrt{\mu}}\nabla f(x)^T - \sqrt{\mu}(x^* - x)\right)^T \sqrt{\mu}(x - x^*) =$$

$\leq a^2 - b^2$

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

Any μ-strongly convex differentiable function is a PL-function

$$
f(x) - f(x^*) \leq \frac{1}{2}\left(\frac{1}{\mu}\|\nabla f(x)\|_2^2 - \left\|\sqrt{\mu}(x - x^*) - \frac{1}{\sqrt{\mu}}\nabla f(x)\right\|_2^2\right)
$$

Any μ-strongly convex differentiable function is a PL-function

≥ 0

$X \leq 100 -$ что угодно $\Rightarrow X \leq 100$

$$
f(x) - f(x^*) \leq \frac{1}{2}\left(\frac{1}{\mu}\|\nabla f(x)\|_2^2 - \left\|\sqrt{\mu}(x - x^*) - \frac{1}{\sqrt{\mu}}\nabla f(x)\right\|_2^2\right)
$$

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

$\Rightarrow$ PL

$\|\nabla f(x)\|_2^2 \geq 2\mu (f - f^*)$

Any μ-strongly convex differentiable function is a PL-function

$$
f(x) - f(x^*) \leq \frac{1}{2}\left(\frac{1}{\mu}\|\nabla f(x)\|_2^2 - \left\| \sqrt{\mu}(x - x^*) - \frac{1}{\sqrt{\mu}}\nabla f(x) \right\|_2^2 \right)
$$

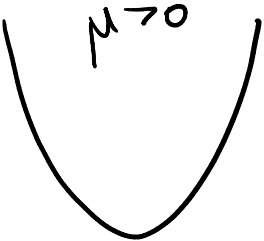
$$
f(x) - f(x^*) \leq \frac{1}{2\mu}\|\nabla f(x)\|_2^2,
$$

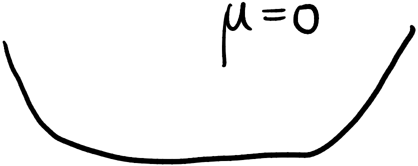
Any μ-strongly convex differentiable function is a PL-function

$$
\begin{aligned}
f(x) - f(x^*) &\leq \frac{1}{2}\left(\frac{1}{\mu}\|\nabla f(x)\|_2^2 - \left\|\sqrt{\mu}(x - x^*) - \frac{1}{\sqrt{\mu}}\nabla f(x)\right\|_2^2\right) \\
f(x) - f(x^*) &\leq \frac{1}{2\mu}\|\nabla f(x)\|_2^2,
\end{aligned}
$$

which is exactly the PL condition. It means, that we already have linear convergence proof for any strongly convex function.

Smooth convex case





линейно

Smooth convex case

!

сх-ть есть только по $f(x)$

Theorem

Consider the Problem

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

and assume that f is convex and L-smooth, for some $L > 0$.

Let $(x^k)_{k \in \mathbb{N}}$ be the sequence of iterates generated by the gradient descent constant stepsize algorithm, with a stepsize satisfying $\boxed{0 < \alpha \leq \frac{1}{L}}$. Then, for all $x^* \in \operatorname{argmin} f$, for all $k \in \mathbb{N}$ we have that

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

$$\boxed{f(x^k) - f^* \leq \frac{\|x^0 - x^*\|^2}{2\alpha k}.}$$

$O\left(\frac{1}{k}\right)$

сублинейная сх-ть

сх-ти по аргументу НЕТ !

Convergence analysis

- As it was before, we first use smoothness:

МОНОТОННОСТЬ GD при правильном α:

$$
\begin{aligned}
f(x^{k+1}) &\leq f(x^k) + \langle \nabla f(x^k), x^{k+1} - x^k \rangle + \frac{L}{2}\|x^{k+1} - x^k\|^2 \\
&= f(x^k) - \alpha\|\nabla f(x^k)\|^2 + \frac{L\alpha^2}{2}\|\nabla f(x^k)\|^2 \\
&= f(x^k) - \frac{\alpha}{2}\left(2 - L\alpha\right)\|\nabla f(x^k)\|^2 \\
&\leq f(x^k) - \frac{\alpha}{2}\|\nabla f(x^k)\|^2,
\end{aligned}
\tag{1}
$$

гладкость

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

$\alpha L \leq 1$ ← ограничение длины шага

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

$$
f(x^k) - f(x^{k+1}) \geq \frac{1}{2L}\|\nabla f(x^k)\|^2 \text{ if } \alpha \leq \frac{1}{L}
$$

Typically, for the convergent gradient descent algorithm the higher the learning rate the faster the convergence. That is why we often will use $\alpha = \frac{1}{L}$.

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

$f(x^k) - f(x^{k+1}) \geq \frac{\alpha}{2}(2 - L\alpha)\|\nabla f(x_k)\|^2$ $\Rightarrow \max\limits_{\alpha}\ \alpha - \frac{L\alpha^2}{2}$

$\alpha_{opt} = \frac{1}{L}$

$\geq \frac{1}{2L}(2 - 1)\|\nabla f(x)\|_2^2$

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

Convergence analysis

- As it was before, we first use smoothness:

$$
\begin{aligned}
f(x^{k+1}) &\leq f(x^k) + \langle \nabla f(x^k), x^{k+1} - x^k \rangle + \frac{L}{2}\|x^{k+1} - x^k\|^2 \\
&= f(x^k) - \alpha \|\nabla f(x^k)\|^2 + \frac{L\alpha^2}{2}\|\nabla f(x^k)\|^2 \\
&= f(x^k) - \frac{\alpha}{2}\left(2 - L\alpha\right)\|\nabla f(x^k)\|^2 \\
&\leq f(x^k) - \frac{\alpha}{2}\|\nabla f(x^k)\|^2, \\
f(x^k) - f(x^{k+1}) &\geq \frac{1}{2L}\|\nabla f(x^k)\|^2 \text{ if } \alpha \leq \frac{1}{L}
\end{aligned}
\tag{1}
$$

  Typically, for the convergent gradient descent algorithm the higher the learning rate the faster the convergence. That is why we often will use $\alpha = \frac{1}{L}$.
- After that we add convexity:

$$
\tag{2}
$$

Convergence analysis

- As it was before, we first use smoothness:

$$
\begin{aligned}
f(x^{k+1}) &\leq f(x^k) + \langle \nabla f(x^k), x^{k+1} - x^k \rangle + \frac{L}{2}\|x^{k+1} - x^k\|^2 \\
&= f(x^k) - \alpha\|\nabla f(x^k)\|^2 + \frac{L\alpha^2}{2}\|\nabla f(x^k)\|^2 \\
&= f(x^k) - \frac{\alpha}{2}\left(2 - L\alpha\right)\|\nabla f(x^k)\|^2 \\
&\leq f(x^k) - \frac{\alpha}{2}\|\nabla f(x^k)\|^2, \\
f(x^k) - f(x^{k+1}) &\geq \frac{1}{2L}\|\nabla f(x^k)\|^2 \text{ if } \alpha \leq \frac{1}{L}
\end{aligned}
\tag{1}
$$

Typically, for the convergent gradient descent algorithm the higher the learning rate the faster the convergence. That is why we often will use $\alpha = \frac{1}{L}$.

- After that we add convexity:

✓ дифф. критерий вып. 1 порядка

$$
f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle
\tag{2}
$$

Convergence analysis

- As it was before, we first use smoothness:

$$
\begin{aligned}
f(x^{k+1}) &\leq f(x^k) + \langle \nabla f(x^k), x^{k+1} - x^k \rangle + \frac{L}{2}\|x^{k+1} - x^k\|^2 \\
&= f(x^k) - \alpha \|\nabla f(x^k)\|^2 + \frac{L\alpha^2}{2}\|\nabla f(x^k)\|^2 \\
&= f(x^k) - \frac{\alpha}{2}\left(2 - L\alpha\right)\|\nabla f(x^k)\|^2 \\
&\leq f(x^k) - \frac{\alpha}{2}\|\nabla f(x^k)\|^2, \\
f(x^k) - f(x^{k+1}) &\geq \frac{1}{2L}\|\nabla f(x^k)\|^2 \text{ if } \alpha \leq \frac{1}{L}
\end{aligned}
\tag{1}
$$

  Typically, for the convergent gradient descent algorithm the higher the learning rate the faster the convergence. That is why we often will use $\alpha = \frac{1}{L}$.
- After that we add convexity:

$$
f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle \text{ with } \underline{y = x^*}, \underline{x = x^k} \tag{2}
$$

$$
f(x^*) \geq f(x^k) + \langle \nabla f(x^k), x^* - x^k \rangle
$$

Convergence analysis

- As it was before, we first use smoothness:

$$
\begin{aligned}
f(x^{k+1}) &\leq f(x^k) + \langle \nabla f(x^k), x^{k+1} - x^k \rangle + \frac{L}{2}\|x^{k+1} - x^k\|^2 \\
&= f(x^k) - \alpha\|\nabla f(x^k)\|^2 + \frac{L\alpha^2}{2}\|\nabla f(x^k)\|^2 \\
&= f(x^k) - \frac{\alpha}{2}\left(2 - L\alpha\right)\|\nabla f(x^k)\|^2 \\
&\leq f(x^k) - \frac{\alpha}{2}\|\nabla f(x^k)\|^2,
\end{aligned} \tag{1}
$$

$$
f(x^k) - f(x^{k+1}) \geq \frac{1}{2L}\|\nabla f(x^k)\|^2 \text{ if } \alpha \leq \frac{1}{L}
$$

Typically, for the convergent gradient descent algorithm the higher the learning rate the faster the convergence. That is why we often will use $\alpha = \frac{1}{L}$.

- After that we add convexity:

$$
\begin{aligned}
&f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle \text{ with } y = x^*, x = x^k \\
&f(x^k) - f^* \leq \langle \nabla f(x^k), x^k - x^* \rangle
\end{aligned} \tag{2}
$$

$f(x^k) \leq f^* + \langle \nabla f(x^k), x^k - x^* \rangle$

Convergence analysis

- Now we put Equation 2 to Equation 1:

Convergence analysis

- Now we put Equation 2 to Equation 1:

(2)

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

(1)

Convergence analysis

- Now we put Equation 2 to Equation 1:

$$f(x^{k+1}) \leq f(x^k) - \frac{\alpha}{2}\|\nabla f(x^k)\|^2 \leq \underline{f^*} + \langle \nabla f(x^k), x^k - x^* \rangle - \frac{\alpha}{2}\|\nabla f(x^k)\|^2$$
$$= \underline{f^*} + \langle \nabla f(x^k), x^k - x^* - \frac{\alpha}{2}\nabla f(x^k) \rangle$$

выносим $\nabla f(x_k)$

Convergence analysis

- Now we put Equation 2 to Equation 1:

$$
\begin{aligned}
f(x^{k+1}) \leq f(x^k) - \frac{\alpha}{2}\|\nabla f(x^k)\|^2 &\leq f^* + \langle \nabla f(x^k), x^k - x^* \rangle - \frac{\alpha}{2}\|\nabla f(x^k)\|^2 \\
&= f^* + \langle \nabla f(x^k), x^k - x^* - \frac{\alpha}{2}\nabla f(x^k) \rangle \\
&= f^* + \frac{1}{2\alpha}\left\langle \alpha \nabla f(x^k), 2\left(x^k - x^* - \frac{\alpha}{2}\nabla f(x^k)\right)\right\rangle
\end{aligned}
$$

Convergence analysis

- Now we put Equation 2 to Equation 1:

$$
\begin{aligned}
f(x^{k+1}) \leq f(x^k) - \frac{\alpha}{2}\|\nabla f(x^k)\|^2 &\leq f^* + \langle \nabla f(x^k), x^k - x^* \rangle - \frac{\alpha}{2}\|\nabla f(x^k)\|^2 \\
&= f^* + \langle \nabla f(x^k), x^k - x^* - \frac{\alpha}{2}\nabla f(x^k) \rangle \\
&= f^* + \frac{1}{2\alpha}\left\langle \alpha \nabla f(x^k), 2\left(x^k - x^* - \frac{\alpha}{2}\nabla f(x^k)\right)\right\rangle
\end{aligned}
$$

Let $\underline{a = x^k - x^*}$ and $\underline{b = x^k - x^* - \alpha\nabla f(x^k)}$.

Convergence analysis

- Now we put Equation 2 to Equation 1:

$$
\begin{aligned}
f(x^{k+1}) \leq f(x^k) - \frac{\alpha}{2}\|\nabla f(x^k)\|^2 &\leq f^* + \langle \nabla f(x^k), x^k - x^* \rangle - \frac{\alpha}{2}\|\nabla f(x^k)\|^2 \\
&= f^* + \langle \nabla f(x^k), x^k - x^* - \frac{\alpha}{2}\nabla f(x^k) \rangle \\
&= f^* + \frac{1}{2\alpha}\left\langle \underbrace{\alpha \nabla f(x^k)}_{a-b}, \underbrace{2\left(x^k - x^* - \frac{\alpha}{2}\nabla f(x^k)\right)}_{a+b} \right\rangle
\end{aligned}
$$

Let $a = x^k - x^*$ and $b = x^k - x^* - \alpha \nabla f(x^k)$. Then $a - b = \alpha \nabla f(x^k)$ and $a + b = 2\left(x^k - x^* - \frac{\alpha}{2}\nabla f(x^k)\right)$.

$$2(x^k - x^*) - \alpha \nabla f(x_k) = a+b$$

Convergence analysis

- Now we put Equation 2 to Equation 1:

$$
\begin{aligned}
f(x^{k+1}) &\leq f(x^k) - \frac{\alpha}{2}\|\nabla f(x^k)\|^2 \leq f^* + \langle \nabla f(x^k), x^k - x^* \rangle - \frac{\alpha}{2}\|\nabla f(x^k)\|^2 \\
&= f^* + \langle \nabla f(x^k), x^k - x^* - \frac{\alpha}{2}\nabla f(x^k) \rangle \\
&= f^* + \frac{1}{2\alpha}\left\langle \alpha \nabla f(x^k), 2\left(x^k - x^* - \frac{\alpha}{2}\nabla f(x^k)\right)\right\rangle
\end{aligned}
$$

Let $a = x^k - x^*$ and $b = x^k - x^* - \alpha\nabla f(x^k)$. Then $a + b = \alpha\nabla f(x^k)$ and $a - b = 2\left(x^k - x^* - \frac{\alpha}{2}\nabla f(x^k)\right)$.

$$
f(x^{k+1}) \leq f^* + \frac{1}{2\alpha}\left[\|x^k - x^*\|_2^2 - \|x^k - x^* - \alpha\nabla f(x^k)\|_2^2\right]
$$

$a^2 \quad - \quad b^2$

Convergence analysis

- Now we put Equation 2 to Equation 1:

$$
\begin{aligned}
f(x^{k+1}) &\leq f(x^k) - \frac{\alpha}{2}\|\nabla f(x^k)\|^2 \leq f^* + \langle \nabla f(x^k), x^k - x^* \rangle - \frac{\alpha}{2}\|\nabla f(x^k)\|^2 \\
&= f^* + \langle \nabla f(x^k), x^k - x^* - \frac{\alpha}{2}\nabla f(x^k) \rangle \\
&= f^* + \frac{1}{2\alpha}\left\langle \alpha \nabla f(x^k), 2\left(x^k - x^* - \frac{\alpha}{2}\nabla f(x^k)\right)\right\rangle
\end{aligned}
$$

Let $a = x^k - x^*$ and $b = x^k - x^* - \alpha \nabla f(x^k)$. Then $a + b = \alpha \nabla f(x^k)$ and $a - b = 2\left(x^k - x^* - \frac{\alpha}{2}\nabla f(x^k)\right)$.

$$
\begin{aligned}
f(x^{k+1}) &\leq f^* + \frac{1}{2\alpha}\left[\|x^k - x^*\|_2^2 - \|x^k - x^* - \alpha \nabla f(x^k)\|_2^2\right] \\
&\leq f^* + \frac{1}{2\alpha}\left[\|x^k - x^*\|_2^2 - \|x^{k+1} - x^*\|_2^2\right]
\end{aligned}
$$

x^{k+1}

Convergence analysis

- Now we put Equation 2 to Equation 1:

$$
\begin{aligned}
f(x^{k+1}) \leq f(x^k) - \frac{\alpha}{2}\|\nabla f(x^k)\|^2 &\leq f^* + \langle \nabla f(x^k), x^k - x^* \rangle - \frac{\alpha}{2}\|\nabla f(x^k)\|^2 \\
&= f^* + \langle \nabla f(x^k), x^k - x^* - \frac{\alpha}{2}\nabla f(x^k) \rangle \\
&= f^* + \frac{1}{2\alpha}\left\langle \alpha \nabla f(x^k), 2\left(x^k - x^* - \frac{\alpha}{2}\nabla f(x^k)\right)\right\rangle
\end{aligned}
$$

Let $a = x^k - x^*$ and $b = x^k - x^* - \alpha\nabla f(x^k)$. Then $a + b = \alpha \nabla f(x^k)$ and $a - b = 2\left(x^k - x^* - \frac{\alpha}{2}\nabla f(x^k)\right)$.

$$
\begin{aligned}
f(x^{k+1}) &\leq f^* + \frac{1}{2\alpha}\left[\|x^k - x^*\|_2^2 - \|x^k - x^* - \alpha\nabla f(x^k)\|_2^2\right] \\
&\leq f^* + \frac{1}{2\alpha}\left[\|x^k - x^*\|_2^2 - \|x^{k+1} - x^*\|_2^2\right]
\end{aligned}
$$

$$
\boxed{2\alpha\left(f(x^{k+1}) - f^*\right) \leq \|x^k - x^*\|_2^2 - \|x^{k+1} - x^*\|_2^2}
$$

Convergence analysis

- Now we put Equation 2 to Equation 1:

$$
\begin{aligned}
f(x^{k+1}) \leq f(x^k) - \frac{\alpha}{2}\|\nabla f(x^k)\|^2 \leq f^* + \langle \nabla f(x^k), x^k - x^* \rangle - \frac{\alpha}{2}\|\nabla f(x^k)\|^2 \\
= f^* + \langle \nabla f(x^k), x^k - x^* - \frac{\alpha}{2}\nabla f(x^k) \rangle \\
= f^* + \frac{1}{2\alpha}\left\langle \alpha\nabla f(x^k), 2\left(x^k - x^* - \frac{\alpha}{2}\nabla f(x^k)\right)\right\rangle
\end{aligned}
$$

Let $a = x^k - x^*$ and $b = x^k - x^* - \alpha\nabla f(x^k)$. Then $a + b = \alpha\nabla f(x^k)$ and $a - b = 2\left(x^k - x^* - \frac{\alpha}{2}\nabla f(x^k)\right)$.

$$
\begin{aligned}
f(x^{k+1}) &\leq f^* + \frac{1}{2\alpha}\left[\|x^k - x^*\|_2^2 - \|x^k - x^* - \alpha\nabla f(x^k)\|_2^2\right] \\
&\leq f^* + \frac{1}{2\alpha}\left[\|x^k - x^*\|_2^2 - \|x^{k+1} - x^*\|_2^2\right] \\
2\alpha\left(f(x^{k+1}) - f^*\right) &\leq \|x^k - x^*\|_2^2 - \|x^{k+1} - x^*\|_2^2
\end{aligned}
$$

- Now suppose, that the last line is defined for some index i and we sum over $i \in [0, k-1]$. Almost all summands will vanish due to the telescopic nature of the sum:

$$
\tag{3}
$$

Convergence analysis

- Now we put Equation 2 to Equation 1:

$$
\begin{aligned}
f(x^{k+1}) \leq f(x^k) - \frac{\alpha}{2}\|\nabla f(x^k)\|^2 &\leq f^* + \langle \nabla f(x^k), x^k - x^* \rangle - \frac{\alpha}{2}\|\nabla f(x^k)\|^2 \\
&= f^* + \langle \nabla f(x^k), x^k - x^* - \frac{\alpha}{2}\nabla f(x^k) \rangle \\
&= f^* + \frac{1}{2\alpha}\left\langle \alpha \nabla f(x^k), 2\left(x^k - x^* - \frac{\alpha}{2}\nabla f(x^k)\right) \right\rangle
\end{aligned}
$$

Let $a = x^k - x^*$ and $b = x^k - x^* - \alpha \nabla f(x^k)$. Then $a + b = \alpha \nabla f(x^k)$ and $a - b = 2\left(x^k - x^* - \frac{\alpha}{2}\nabla f(x^k)\right)$.

$$
\begin{aligned}
f(x^{k+1}) &\leq f^* + \frac{1}{2\alpha}\left[\|x^k - x^*\|_2^2 - \|x^k - x^* - \alpha \nabla f(x^k)\|_2^2\right] \\
&\leq f^* + \frac{1}{2\alpha}\left[\|x^k - x^*\|_2^2 - \|x^{k+1} - x^*\|_2^2\right] \\
2\alpha\left(f(x^{k+1}) - f^*\right) &\leq \|x^k - x^*\|_2^2 - \|x^{k+1} - x^*\|_2^2
\end{aligned}
$$

- Now suppose, that the last line is defined for some index i and we sum over $i \in [0, k-1]$. Almost all summands will vanish due to the telescopic nature of the sum:

$$
2\alpha \sum_{i=0}^{k-1} \left(f(x^{i+1}) - f^*\right) \leq \|x^0 - x^*\|_2^2 - \|x^k - x^*\|_2^2 \tag{3}
$$

Convergence analysis

- Now we put Equation 2 to Equation 1:

$$
\begin{aligned}
f(x^{k+1}) \leq f(x^k) - \frac{\alpha}{2}\|\nabla f(x^k)\|^2 &\leq f^* + \langle \nabla f(x^k), x^k - x^* \rangle - \frac{\alpha}{2}\|\nabla f(x^k)\|^2 \\
&= f^* + \langle \nabla f(x^k), x^k - x^* - \frac{\alpha}{2}\nabla f(x^k) \rangle \\
&= f^* + \frac{1}{2\alpha}\left\langle \alpha \nabla f(x^k), 2\left(x^k - x^* - \frac{\alpha}{2}\nabla f(x^k)\right)\right\rangle
\end{aligned}
$$

Let $a = x^k - x^*$ and $b = x^k - x^* - \alpha \nabla f(x^k)$. Then $a + b = \alpha \nabla f(x^k)$ and $a - b = 2\left(x^k - x^* - \frac{\alpha}{2}\nabla f(x^k)\right)$.

$$
\begin{aligned}
f(x^{k+1}) &\leq f^* + \frac{1}{2\alpha}\left[\|x^k - x^*\|_2^2 - \|x^k - x^* - \alpha \nabla f(x^k)\|_2^2\right] \\
&\leq f^* + \frac{1}{2\alpha}\left[\|x^k - x^*\|_2^2 - \|x^{k+1} - x^*\|_2^2\right] \\
2\alpha\left(f(x^{k+1}) - f^*\right) &\leq \|x^k - x^*\|_2^2 - \|x^{k+1} - x^*\|_2^2
\end{aligned}
$$

- Now suppose, that the last line is defined for some index i and we sum over $i \in [0, k-1]$. Almost all summands will vanish due to the telescopic nature of the sum:

$$
2\alpha \sum_{i=0}^{k-1} \left(f(x^{i+1}) - f^*\right) \leq \|x^0 - x^*\|_2^2 \cancel{- \|x^k - x^*\|_2^2} \leq \|x^0 - x^*\|_2^2 = R^2 \tag{3}
$$

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

Convergence analysis

- Due to the monotonic decrease at each iteration $f(x^{i+1}) < f(x^i)$:

$$
kf(x^k) \leq \sum_{i=0}^{k-1} f(x^{i+1})
$$

к слагаемых

Convergence analysis

- Due to the monotonic decrease at each iteration $f(x^{i+1}) < f(x^i)$:

$$
kf(x^k) \leq \sum_{i=0}^{k-1} f(x^{i+1}) \quad \Rightarrow \quad f(x_k) \leq \frac{\sum_{i=0}^{k-1} f_i}{k}
$$

- Now putting it to Equation 3:

Convergence analysis

- Due to the monotonic decrease at each iteration $f(x^{i+1}) < f(x^i)$:

$$kf(x^k) \leq \sum_{i=0}^{k-1} f(x^{i+1})$$

- Now putting it to Equation 3:

$$2\alpha kf(x^k) - 2\alpha kf^* \leq 2\alpha \sum_{i=0}^{k-1} \left(f(x^{i+1}) - f^*\right) \leq \|x^0 - x^*\|_2^2$$

$$f(x^k) - f^* \leq \frac{\|x^0 - x^*\|_2^2}{2\alpha k} \leq \frac{LR^2}{2k}$$

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

Convergence analysis

- Due to the monotonic decrease at each iteration $f(x^{i+1}) < f(x^i)$:

$$kf(x^k) \leq \sum_{i=0}^{k-1} f(x^{i+1})$$

- Now putting it to Equation 3:

$$2\alpha k f(x^k) - 2\alpha k f^* \leq 2\alpha \sum_{i=0}^{k-1} \left(f(x^{i+1}) - f^*\right) \leq \|x^0 - x^*\|_2^2$$

$$f(x^k) - f^* \leq \frac{\|x^0 - x^*\|_2^2}{2\alpha k}$$

Convergence analysis

- Due to the monotonic decrease at each iteration $f(x^{i+1}) < f(x^i)$:

$$
kf(x^k) \leq \sum_{i=0}^{k-1} f(x^{i+1})
$$

- Now putting it to Equation 3:

$$
2\alpha k f(x^k) - 2\alpha k f^* \leq 2\alpha \sum_{i=0}^{k-1} \left(f(x^{i+1}) - f^* \right) \leq \|x^0 - x^*\|_2^2
$$

$$
f(x^k) - f^* \leq \frac{\|x^0 - x^*\|_2^2}{2\alpha k} \leq \frac{L\|x^0 - x^*\|_2^2}{2k}
$$

Q.T.D.