

**Gradient Descent** 



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

⊕ ი

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

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

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:

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

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

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  $\alpha \to 0$ :

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

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

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

 $Also\ from\ Cauchy-Bunyakovsky-Schwarz\ inequality:$ 

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

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  $\alpha \to 0$ :

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

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  $\alpha \to 0$ :

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

Also from Cauchy–Bunyakovsky–Schwarz inequality:

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

Thus, the direction of the antigradient

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

gives the direction of the  ${\bf steepest}$   ${\bf local}$  decreasing of the function f.

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:

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

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

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

⊕ ი დ

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

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

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  $\alpha$  step:

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

♥ C) Ø

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  $\alpha$  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 🖺

 $f \to \min_{x,y,z}$  Gradient Descent

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

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

and discretize it on a uniform grid with  $\alpha$  step:

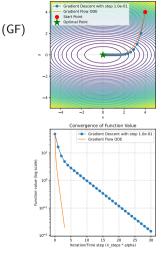
$$\frac{x_{k+1}-x_k}{\alpha}=-f'(x_k),$$

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 🖺



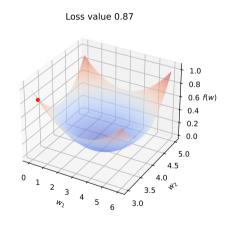
Trajectories with Contour Plot

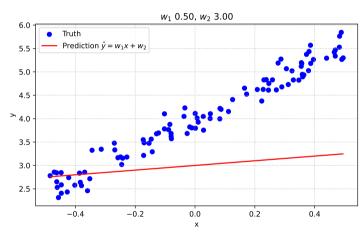
Рисунок 1: Gradient flow trajectory

Gradient Descent

## Convergence of Gradient Descent algorithm

Heavily depends on the choice of the learning rate  $\alpha$ :







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

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

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

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

Optimality conditions:

# Exact line search aka steepest descent

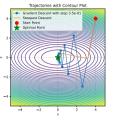
$$\alpha_k = \arg\min_{\alpha \in \mathbb{R}^+} f(x_{k+1}) = \arg\min_{\alpha \in \mathbb{R}^+} f(x_k - \alpha \nabla f(x_k))$$

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

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

Optimality conditions:

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



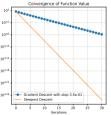


Рисунок 2: Steepest Descent

Open In Colab 🐥



**Strongly convex quadratics** 





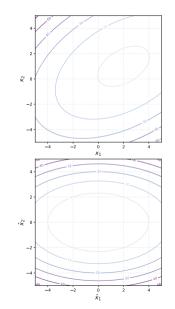
Consider the following quadratic optimization problem:

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

Consider the following quadratic optimization problem:

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

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

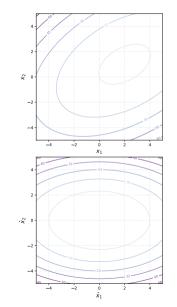




Consider the following quadratic optimization problem:

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

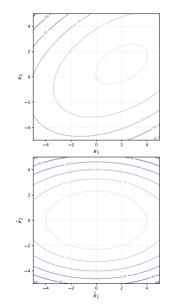
- ullet Firstly, without loss of generality we can set c=0, which will or affect optimization process.
- ullet Secondly, we have a spectral decomposition of the matrix  $A=Q\Lambda Q^T.$



Consider the following quadratic optimization problem:

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

- ullet 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  $A=Q\Lambda Q^T$ .
- 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  $x = Q\hat{x} + x^*$ .



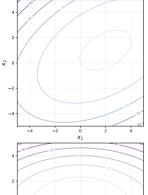
⊕ n ø

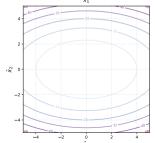
Consider the following quadratic optimization problem:

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

- 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  $A=Q\Lambda Q^T.$
- 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  $x = Q\hat{x} + x^*$ .

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





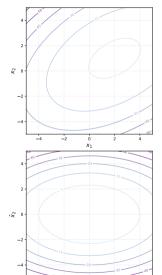
♥ ೧ •

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  $A=Q\Lambda Q^T$ .
- 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  $x = Q\hat{x} + x^*$ .

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



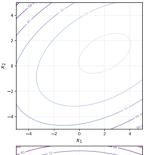


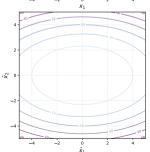
Consider the following quadratic optimization problem:

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

- ullet Firstly, without loss of generality we can set c=0, which will or affect optimization process.
- ullet Secondly, we have a spectral decomposition of the matrix  $A=Q\Lambda Q^T.$
- 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  $x = Q\hat{x} + x^*$ .

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





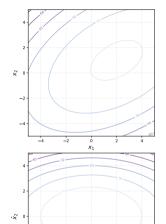
Strongly convex quadratics

Consider the following quadratic optimization problem:

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

- ullet 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  $A=Q\Lambda Q^T.$
- 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  $x = Q\hat{x} + x^*$ .

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



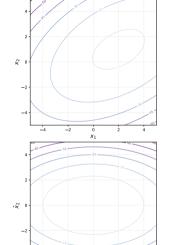


Consider the following quadratic optimization problem:

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

- ullet Firstly, without loss of generality we can set c=0, which will or affect optimization process.
- ullet Secondly, we have a spectral decomposition of the matrix  $A=Q\Lambda Q^T.$
- 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  $x = Q\hat{x} + x^*$ .

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



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  $\hat{x}$ )

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

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  $\hat{x}$ )

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

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  $\hat{x}$ )

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

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  $\hat{x}$ )

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

 $x_{(i)}^{k+1} = (1 - \alpha^k \lambda_{(i)}) x_{(i)}^k$  For *i*-th coordinate

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  $\hat{x}$ )

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

 $x_{(i)}^{k+1} = (1 - \alpha^k \lambda_{(i)}) x_{(i)}^k$  For *i*-th coordinate

$$x_{(i)}^{k+1} = (1-\alpha^k\lambda_{(i)})^kx_{(i)}^0$$



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  $\hat{x}$ )

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

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

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

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

 $f \to \min_{x,y}$ 

⊕ ი മ

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  $\hat{x}$ )

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

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

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

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

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

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  $\hat{x}$ )

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

$$x_{(i)}^{k+1} = (1-\alpha^k\lambda_{(i)})^kx_{(i)}^0$$

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

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

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

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

 $f \to \min_{x,y,z}$  Strongly convex quadratics

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  $\hat{x}$ )

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

$$x_{(i)}^{(i)} = (1 - \alpha^k \lambda_{(i)})^k x_{(i)}^0$$

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

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

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

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

$$-1 < 1 - \alpha \mu < 1$$
$$\alpha < \frac{2}{\mu} \qquad \alpha \mu > 0$$

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  $\hat{x}$ )

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

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

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

 $x_{(i)}^{k+1} = (1 - \alpha^k \lambda_{(i)})^k x_{(i)}^0$ 

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

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

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  $\hat{x}$ )

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

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

$$ho(lpha) = \max |1 - lpha \lambda_{(i)}| < 1$$

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

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

 $f \to \min_{x,y,z}$  Strongly convex quadratics

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  $\hat{x}$ )

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

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

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

 $x_{(i)}^{k+1} = (1 - \alpha^k \lambda_{(i)})^k x_{(i)}^0$ 

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

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

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  $\hat{x}$ )

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

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

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

 $x_{(i)}^{k+1} = (1 - \alpha^k \lambda_{(i)})^k x_{(i)}^0$ 

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

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

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  $\hat{x}$ )

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

$$x_{(i)}^{k+1}=(1-\alpha^k\lambda_{(i)})x_{(i)}^k \text{ For } i\text{-th coordinate}$$
 
$$x_{(i)}^{k+1}=(1-\alpha^k\lambda_{(i)})^kx_{(i)}^0$$
 Let's use constant stepsize  $\alpha^k=\alpha$ . Convergence

condition:  $\rho(\alpha) = \max|1 - \alpha\lambda_{(i)}| < 1$ 

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

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

$$r = r = r = r$$

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

$$\alpha &< \frac{2}{\mu} \quad \alpha \mu > 0 \qquad \alpha &< \frac{2}{L} \quad \alpha L > 0$$

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

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  $\hat{x}$ )

$$x^{k+1} = x^k - \alpha^k \nabla f(x^k) = x^k - \alpha^k \Lambda x^k$$
$$= (I - \alpha^k \Lambda) x^k$$
$$x_{(i)}^{k+1} = (1 - \alpha^k \lambda_{(i)}) x_{(i)}^k \text{ For } i\text{-th coordinate}$$

Let's use constant stepsize 
$$\alpha^k=\alpha.$$
 Convergence condition: 
$$\rho(\alpha)=\max|1-\alpha\lambda_{(i)}|<1$$

 $x_{(i)}^{k+1} = (1 - \alpha^k \lambda_{(i)})^k x_{(i)}^0$ 

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

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

$$\alpha < \frac{2}{\mu} \qquad \alpha \mu > 0 \qquad \qquad \alpha < \frac{2}{L} \qquad \alpha L > 0$$
 
$$\alpha < \frac{2}{L} \quad \text{is needed for convergence.}$$

Now we would like to tune  $\alpha$  to choose the best (lowest) convergence rate

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

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  $\hat{x}$ )

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

 $x_{(i)}^{k+1} = (1 - \alpha^k \lambda_{(i)})^k x_{(i)}^0$ 

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

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

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

$$|1-lpha\mu|<1$$
  $|1-lpha L|<1$ 

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

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

$$\alpha < \frac{2}{\mu} \quad \alpha \mu > 0$$

$$\alpha < \frac{2}{L} \quad \alpha L > 0$$

Now we would like to tune  $\alpha$  to choose the best (lowest) convergence rate

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

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

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  $\hat{x}$ )

$$x^{k+1} = x^k - \alpha^k \nabla f(x^k) = x^k - \alpha^k \Lambda x^k$$
$$= (I - \alpha^k \Lambda) x^k$$
$$x_{(i)}^{k+1} = (1 - \alpha^k \lambda_{(i)}) x_{(i)}^k \text{ For } i\text{-th coordinate}$$

 $x_{(i)}^{k+1} = (1 - \alpha^k \lambda_{(i)})^k x_{(i)}^0$ 

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

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

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

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

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

$$\alpha < \frac{2}{\mu} \qquad \alpha \mu > 0 \qquad \qquad \alpha < \frac{2}{L} \qquad \alpha L > 0$$

Now we would like to tune  $\boldsymbol{\alpha}$  to choose the best (lowest) convergence rate

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

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

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  $\hat{x}$ )

$$x^{k+1} = x^k - \alpha^k \nabla f(x^k) = x^k - \alpha^k \Lambda x^k$$
$$= (I - \alpha^k \Lambda) x^k$$
$$x_{(i)}^{k+1} = (1 - \alpha^k \lambda_{(i)}) x_{(i)}^k \text{ For } i\text{-th coordinate}$$

condition: 
$$\rho(\alpha) = \max |1 - \alpha \lambda_{(i)}| < 1$$

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

 $x_{(i)}^{k+1} = (1 - \alpha^k \lambda_{(i)})^k x_{(i)}^0$ 

$$|\lambda_{(i)}| < 1$$

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

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

$$\alpha < \frac{2}{\mu} \qquad \alpha \mu > 0 \qquad \qquad \alpha < \frac{2}{L} \qquad \alpha L > 0$$
 
$$\alpha < \frac{2}{L} \quad \text{is needed for convergence.}$$

Now we would like to tune  $\alpha$  to choose the best (lowest) convergence rate

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

 $|1 - \alpha u| < 1$ 

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  $\hat{x}$ )

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

$$x_{(i)}^{k+1}=(1-\alpha^k\lambda_{(i)})^kx_{(i)}^0$$
 Let's use constant stepsize  $\alpha^k=\alpha.$  Convergence

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

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

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

$$\begin{array}{ll} \alpha<\frac{2}{\mu} & \alpha\mu>0 & \alpha<\frac{2}{L} & \alpha L>0 \\ \alpha<\frac{2}{T} \text{ is needed for convergence.} \end{array}$$

Now we would like to tune  $\boldsymbol{\alpha}$  to choose the best (lowest) convergence rate

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

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

$$\alpha^* = \frac{2}{\mu + L}$$

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  $\hat{x}$ )

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

 $x_{(i)}^{k+1} = (1 - \alpha^k \lambda_{(i)})^k x_{(i)}^0$ 

Let's use constant stepsize 
$$\alpha^k=\alpha$$
. Convergence condition: 
$$\rho(\alpha)=\max|1-\alpha\lambda_{(i)}|<1$$

$$ho(lpha)=\max_i |1-lpha \chi_{(i)}|<1$$

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

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

Now we would like to tune  $\alpha$  to choose the best (lowest) convergence rate

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

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

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

 $f \to \min_{x,y,z}$  Strongly convex quadratics

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  $\hat{x}$ )

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

$$x_{(i)}^{k+1}=(1-\alpha^k\lambda_{(i)})^kx_{(i)}^0$$
 Let's use constant stepsize  $\alpha^k=\alpha$ . Convergence

condition:  $\rho(\alpha) = \max |1 - \alpha \lambda_{(\alpha)}| < 1$ 

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

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

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

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

Now we would like to tune  $\alpha$  to choose the best (lowest) convergence rate

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

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

$$x_{(i)}^{k+1} = \left(\frac{L-\mu}{L+\mu}\right)^k x_{(i)}^0$$

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  $\hat{x}$ )

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

Let's use constant stepsize 
$$\alpha^k = \alpha$$
. Convergence condition:  
 $a(\alpha) = \max |1 - \alpha \lambda_{\alpha}| < 1$ 

 $x_{(i)}^{k+1} = (1 - \alpha^k \lambda_{(i)})^k x_{(i)}^0$ 

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

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

the interpolation of 
$$\lambda_{\min} = \mu > 0$$
,  $\lambda_{\max} = D \ge \mu$ .

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

$$\alpha &< \frac{2}{\mu} \quad \alpha \mu > 0 \qquad \alpha &< \frac{2}{L} \quad \alpha L > 0$$

Now we would like to tune  $\alpha$  to choose the best (lowest) convergence rate

$$= \min_{\alpha} \left\{ |1 - \alpha \mu|, |1 - \alpha L| \right\}$$

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

 $\rho^* = \min \rho(\alpha) = \min \max |1 - \alpha \lambda_{(i)}|$ 

$$\alpha^* = \frac{2}{\mu + L} \quad \rho^* = \frac{L - \mu}{L + \mu}$$
$$x_{(i)}^{k+1} = \left(\frac{L - \mu}{L + \mu}\right)^k x_{(i)}^0$$

$$||x^{k+1}||_2 \le \left(\frac{L-\mu}{L+\mu}\right)^k ||x^0||_2$$

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

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  $\hat{x}$ )

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

$$\begin{split} x_{(i)}^{k+1} &= (1-\alpha^k\lambda_{(i)})x_{(i)}^k \text{ For } i\text{-th coordinate} \\ x_{(i)}^{k+1} &= (1-\alpha^k\lambda_{(i)})^kx_{(i)}^0 \end{split}$$

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

use constant stepsize 
$$lpha^k=lpha$$
. Convergence on: 
$$ho(lpha)=\max|1-lpha\lambda_{(i)}|<1$$

 $\rho(\alpha) = \max|1 - \alpha\lambda_{(i)}| < 1$ Remember, that  $\lambda_{\min} = \mu > 0, \lambda_{\max} = L \ge \mu$ .

$$\lim_{i} \left( 1 - \cos \lambda_{(i)} \right) < 1$$
  $\mu > 0, \lambda_{\mathsf{max}} = L \geq \mu.$ 

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

 $x_{(i)}^{k+1} = (1 - \alpha^k \lambda_{(i)})^k x_{(i)}^0$ 

Now we would like to tune  $\alpha$  to choose the best (lowest) convergence rate

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

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

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

$$\rho^* = \frac{r}{L + \mu}$$

$$+ \mu$$

 $x_{(i)}^{k+1} = \left(\frac{L-\mu}{L+\mu}\right)^k x_{(i)}^0$ 

$$x_{(i)}^{k+1} = \left(\frac{L-\mu}{L+\mu}\right) x_{(i)}^{0}$$

$$(L-\mu)^{k} x_{(i)} = (L-\mu)^{k} x_{(i)} + (L-\mu)^{k} x_{(i)} = (L-\mu)^{k} x_{(i)}$$

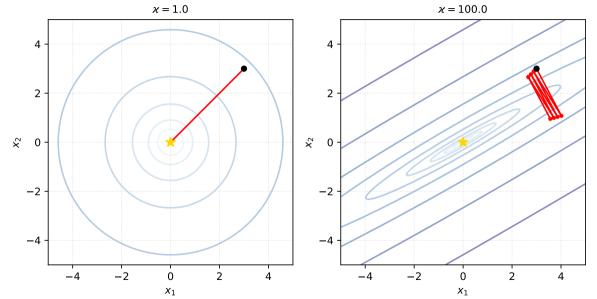
$$\|x^{k+1}\|_2 \le \left(\frac{L-\mu}{L+\mu}\right)^k \|x^0\|_2 \quad f(x^{k+1}) \le \left(\frac{L-\mu}{L+\mu}\right)^{2k} f(x^0)$$

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

N	ho	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



## **Condition number** $\varkappa$



Polyak-Lojasiewicz smooth case





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

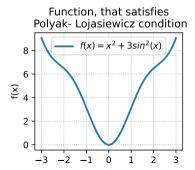
PL inequality holds if the following condition is satisfied for some  $\mu>0$ ,

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

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. **PLink** 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  $\mu > 0$ ,

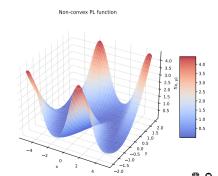
$$\|\nabla f(x)\|^2 \ge 2\mu(f(x) - f^*) \quad \forall x$$

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

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

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



#### i Theorem

Consider the Problem

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

and assume that f is  $\mu$ -Polyak-Lojasiewicz and L-smooth, for some  $L \ge \mu > 0$ .

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:

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



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

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



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

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

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

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

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

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



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

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

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

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

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

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

i Theorem

If a function f(x) is differentiable and  $\mu\text{-strongly convex, then it is a PL function.}$ 

#### Proof

By first order strong convexity criterion:

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

Putting  $y = x^*$ :

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

#### i Theorem

If a function f(x) is differentiable and  $\mu\text{-strongly convex}$ , then it is a PL function.

#### **Proof**

By first order strong convexity criterion:

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

Putting  $y = x^*$ :

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

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

#### 1 Theorem

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

#### Proof

By first order strong convexity criterion:

Polyak-Loiasiewicz smooth case

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

Putting  $y = x^*$ :

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

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

#### i Theorem

If a function f(x) is differentiable and  $\mu\text{-strongly}$  convex, then it is a PL function.

#### Proof

By first order strong convexity criterion:

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

Putting  $u = x^*$ :

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

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

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

$$f 
ightarrow \min_{x,y,z}$$
 Polyak-Lojasiewicz smooth case

#### i Theorem

If a function f(x) is differentiable and  $\mu\text{-strongly}$  convex, then it is a PL function.

#### Proof

By first order strong convexity criterion:

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

Putting  $u = x^*$ :

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

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

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

$$f o \min_{x,y,z}$$
 Polyak-Lojasiewicz smooth case

#### i Theorem

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

#### Proof

By first order strong convexity criterion:

$$=x^*$$
:

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

Putting  $y = x^*$ :

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

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

$$f \to \min_{x,y,z}$$
 Polyak-Lojasiewicz smooth case

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

#### 1 Theorem

If a function f(x) is differentiable and  $\mu$ -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  $y=x^*$ :

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

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

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

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

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



# Any $\mu$ -strongly convex differentiable function is a PL-function

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

## Any $\mu$ -strongly convex differentiable function is a PL-function

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

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

#### Smooth convex case





#### Smooth convex case

#### i Theorem

Consider the Problem

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

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  $0<\alpha\leq \frac{1}{L}$ . Then, for all  $x^*\in \operatorname{argmin} f$ , for all  $k\in\mathbb{N}$  we have that

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



As it was before, we first use smoothness:

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

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

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

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

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

$$(1)$$

Typically, for the convergent gradient descent algorithm the higher the learning rate the faster the convergence.

That is why we often will use  $\alpha = \frac{1}{L}$ .

As it was before, we first use smoothness:

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

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

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

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

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

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:

(2)

As it was before, we first use smoothness:

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

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

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

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

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

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:

$$f(y) \ge f(x) + \langle \nabla f(x), y - x \rangle$$

(2)

**⊕** 0 0

• As it was before, we first use smoothness:

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

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

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

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

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

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:

$$f(y) \ge f(x) + \langle \nabla f(x), y - x \rangle$$
 with  $y = x^*, x = x^k$ 

(2)

 $\xrightarrow{x,y,z}$  Smooth convex case

⊕ ი (

• As it was before, we first use smoothness:

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

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

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:

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

$$f(x^k) - f^* < \langle \nabla f(x^k), x^k - x^* \rangle$$
(2)

 $f \to \min_{x,y,z}$ 

⊕ ი დ

• Now we put Уравнение 2 to Уравнение 1:

⇔ ი

• Now we put Уравнение 2 to Уравнение 1:

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

• Now we put Уравнение 2 to Уравнение 1:

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

• Now we put Уравнение 2 to Уравнение 1:

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

 $f \to \min_{x,y,z}$  Smooth convex case

• Now we put Уравнение 2 to Уравнение 1:

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

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

 $f \to \min_{x,y,z}$  Smooth convex case

Now we put Уравнение 2 to Уравнение 1:

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

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

Now we put Уравнение 2 to Уравнение 1:

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

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

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

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

Now we put Уравнение 2 to Уравнение 1:

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

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

$$\begin{split} &-x^* - \alpha \nabla f(x^k). \text{ Then } a + b = \alpha \nabla f(x^k) \text{ and } a - b = 2\left(x^k - x^* - \frac{\alpha}{2} \nabla f(x^k) + f(x^{k+1})\right) \leq f^* + \frac{1}{2\alpha} \left[ \|x^k - x^*\|_2^2 - \|x^k - x^* - \alpha \nabla f(x^k)\|_2^2 \right] \end{split}$$

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

 $f \to \min_{x,y,z}$  Smooth convex case

Now we put Уравнение 2 to Уравнение 1:

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

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

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

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

• Now we put Уравнение 2 to Уравнение 1:

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

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

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

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

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

$$2\alpha \left( f(x^{n+1}) - f^{+} \right) \le \|x^{n} - x^{+}\|_{2}^{2} - \|x^{n+1} - x\|_{2}^{2}$$

• 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:

Now we put Уравнение 2 to Уравнение 1:

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

 $f = f^* + \frac{1}{2\alpha} \left\langle \alpha \nabla f(x^k), 2 \left( x^k - x^* - \frac{\alpha}{2} \nabla f(x^k) \right) \right\rangle$ Let  $a = x^k - x^*$  and  $b = x^k - x^* - \alpha \nabla f(x^k)$ . Then  $a + b = \alpha \nabla f(x^k)$  and  $a - b = 2(x^k - x^* - \frac{\alpha}{2} \nabla f(x^k))$ .

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

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

• 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:

will vanish due to the telescopic nature of the sum: 
$$2\alpha \sum^{k-1} \left(f(x^{i+1}) - f^*\right) \leq \|x^0 - x^*\|_2^2 - \|x^k - x^*\|_2^2 \tag{3}$$

Smooth convex case

(3)

• Now we put Уравнение 2 to Уравнение 1:

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

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

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

 $=f^*+\frac{1}{2\alpha}\left\langle\alpha\nabla f(x^k),2\left(x^k-x^*-\frac{\alpha}{2}\nabla f(x^k)\right)\right\rangle$  Let  $a=x^k-x^*$  and  $b=x^k-x^*-\alpha\nabla f(x^k)$ . Then  $a+b=\alpha\nabla f(x^k)$  and  $a-b=2\left(x^k-x^*-\frac{\alpha}{2}\nabla f(x^k)\right)$ .

$$2\alpha\left(f(x^{k+1})-f^*\right)\leq \|x^k-x^*\|_2^2-\|x^{k+1}-x^*\|_2^2$$
 Now suppose, that the last line is defined for some index  $i$  and we sum over  $i\in[0,k-1]$ . Almost all summands illustrate that the table part is a transfer of the result of the same for the same

will vanish due to the telescopic nature of the sum:

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

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

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



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

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

Now putting it to Уравнение 3:



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

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

• Now putting it to Уравнение 3:

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



vex case

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

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

Now putting it to Уравнение 3:

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

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

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

• Now putting it to Уравнение 3:

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



# **Summary**

Gradient Descent:

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

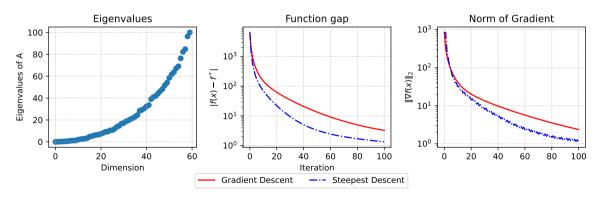
 $x^{k+1} = x^k - \alpha^k \nabla f(x^k)$ 

$\ \nabla f(x^k)\ ^2 \sim \mathcal{O}\left(\frac{1}{k}\right) \qquad \qquad f(x^k) - f^* \sim \mathcal{O}\left(\frac{1}{k}\right) \qquad \qquad \ x^k - x^*\ ^2 \sim \mathcal{O}\left(\left(1 - \frac{\mu}{L}\right)^k\right) \\ k_\varepsilon \sim \mathcal{O}\left(\frac{1}{\varepsilon}\right) \qquad \qquad k_\varepsilon \sim \mathcal{O}\left(\frac{1}{\varepsilon}\right) \qquad \qquad k_\varepsilon \sim \mathcal{O}\left(\varkappa \log \frac{1}{\varepsilon}\right)$	smooth (non-convex)	smooth & convex	smooth & strongly convex (or PL)
	(1)	(1)	( 1 ) (

Smooth convex case

$$f(x) = \frac{1}{2} x^T A x - b^T x \to \min_{x \in \mathbb{R}^n}$$

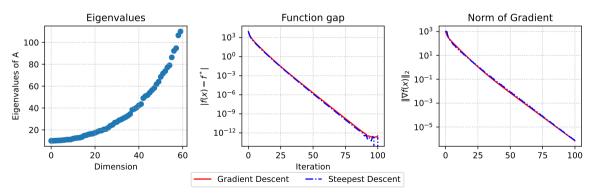
Convex quadratics. n=60, random matrix.





$$f(x) = \frac{1}{2} x^T A x - b^T x \to \min_{x \in \mathbb{R}^n}$$

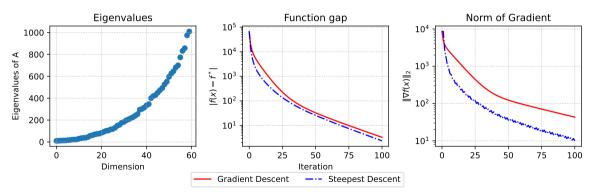
Strongly convex quadratics. n=60, random matrix.





$$f(x) = \frac{1}{2} x^T A x - b^T x \to \min_{x \in \mathbb{R}^n}$$

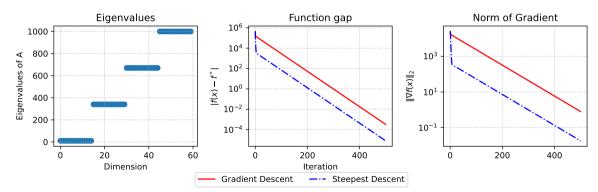
Strongly convex quadratics. n=60, random matrix.





$$f(x) = \frac{1}{2} x^T A x - b^T x \to \min_{x \in \mathbb{R}^n}$$

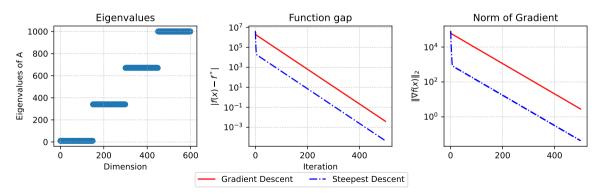
Strongly convex quadratics. n=60, clustered matrix.





$$f(x) = \frac{1}{2} x^T A x - b^T x \to \min_{x \in \mathbb{R}^n}$$

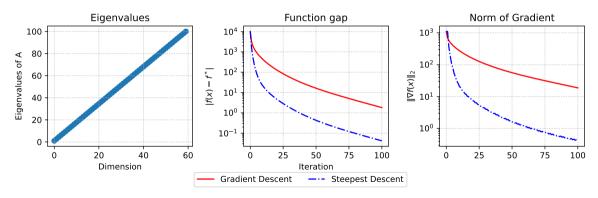
Strongly convex quadratics. n=600, clustered matrix.





$$f(x) = \frac{1}{2} x^T A x - b^T x \to \min_{x \in \mathbb{R}^n}$$

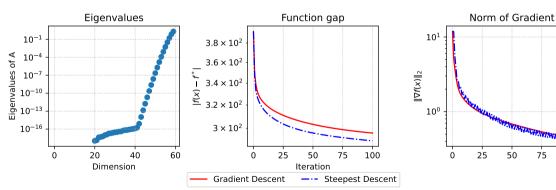
Strongly convex quadratics. n=60, uniform spectrum matrix.





$$f(x) = \frac{1}{2} x^T A x - b^T x \to \min_{x \in \mathbb{R}^n}$$

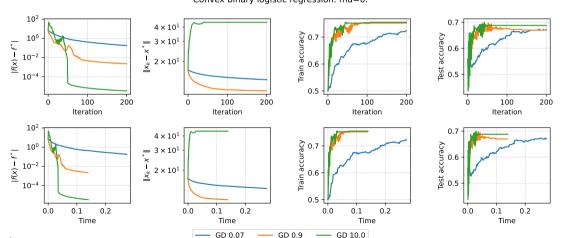
Strongly convex quadratics. n=60, Hilbert matrix.



100

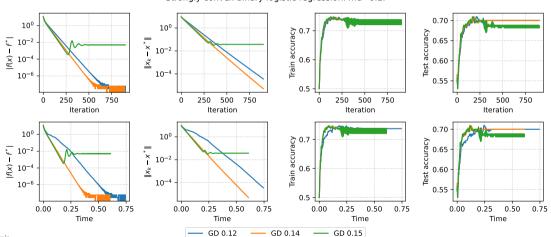
$$f(x) = \frac{\mu}{2} \|x\|_2^2 + \frac{1}{m} \sum_{i=1}^m \log(1 + \exp(-y_i \langle a_i, x \rangle)) \rightarrow \min_{x \in \mathbb{R}^n}$$

Convex binary logistic regression, mu=0.



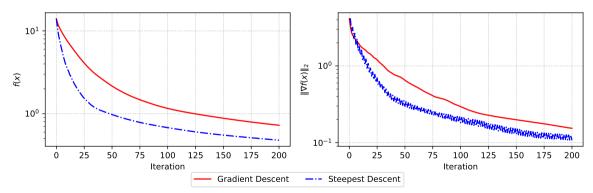
$$f(x) = \frac{\mu}{2} \|x\|_2^2 + \frac{1}{m} \sum_{i=1}^m \log(1 + \exp(-y_i \langle a_i, x \rangle)) \to \min_{x \in \mathbb{R}^n}$$

Strongly convex binary logistic regression. mu=0.1.



$$f(x) = \frac{\mu}{2} \|x\|_2^2 + \frac{1}{m} \sum_{i=1}^m \log(1 + \exp(-y_i \langle a_i, x \rangle)) \rightarrow \min_{x \in \mathbb{R}^n}$$

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





$$f(x) = \frac{\mu}{2} \|x\|_2^2 + \frac{1}{m} \sum_{i=1}^m \log(1 + \exp(-y_i \langle a_i, x \rangle)) \rightarrow \min_{x \in \mathbb{R}^n}$$

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

