

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

Daniil Merkulov

Optimization for ML. Faculty of Computer Science. HSE University



## Direction of local steepest descent

Let's consider a linear approximation of the differentiable function  $f$  along some direction

$h, \|h\|_2 = 1$ :

## Direction of local steepest descent

Let's consider a linear approximation of the differentiable function  $f$  along some direction

$h$ ,  $\|h\|_2 = 1$ :

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

## Direction of local steepest descent

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

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

We want  $h$  to be a decreasing direction:

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

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

## 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:

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

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

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

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

## Direction of local steepest descent

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

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

We want  $h$  to be a decreasing direction:

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

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

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

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

Also from Cauchy–Bunyakovsky–Schwarz inequality:

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

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

## 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  $\alpha \rightarrow 0$ :

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

Also from Cauchy–Bunyakovsky–Schwarz inequality:

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

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

Thus, the direction of the antigradient

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

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

## 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  $\alpha \rightarrow 0$ :

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

Also from Cauchy–Bunyakovsky–Schwarz inequality:

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

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

Thus, the direction of the antigradient

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

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

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



## Gradient flow ODE

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

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

## Gradient flow ODE

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

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

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

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

## Gradient flow ODE

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

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

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 

# Gradient flow ODE

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

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

(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 ♣



Figure 1: Gradient flow trajectory

# 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:

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

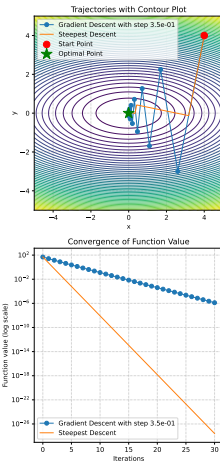


Figure 2: Steepest Descent

Open In Colab



## Coordinate shift

Consider the following quadratic optimization problem:

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

## 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 not affect optimization process.



## 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 not affect optimization process.
- Secondly, we have a spectral decomposition of the matrix  $A$ :

$$A = Q \Lambda Q^\top$$



## 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 not affect optimization process.
- Secondly, we have a spectral decomposition of the matrix  $A$ :

$$A = Q \Lambda Q^\top$$

- Let's show, that we can switch coordinates to make an analysis a little bit easier. Let  $\hat{x} = Q^\top(x - x^*)$ , where  $x^*$  is the minimum point of initial function, defined by  $Ax^* = b$ . At the same time  $x = Q\hat{x} + x^*$ .



## Coordinate shift

Consider the following quadratic optimization problem:

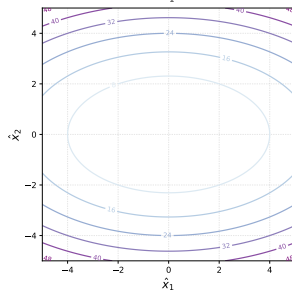
$$\min_{x \in \mathbb{R}^d} f(x) = \min_{x \in \mathbb{R}^d} \frac{1}{2} x^\top A x - b^\top x + c, \text{ where } A \in \mathbb{S}_{++}^d.$$

- Firstly, without loss of generality we can set  $c = 0$ , which will not affect optimization process.
- Secondly, we have a spectral decomposition of the matrix  $A$ :

$$A = Q \Lambda Q^\top$$

- Let's show, that we can switch coordinates to make an analysis a little bit easier. Let  $\hat{x} = Q^\top(x - x^*)$ , 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 not affect optimization process.
- Secondly, we have a spectral decomposition of the matrix  $A$ :

$$A = Q \Lambda Q^\top$$

- Let's show, that we can switch coordinates to make an analysis a little bit easier. Let  $\hat{x} = Q^\top(x - x^*)$ , where  $x^*$  is the minimum point of initial function, defined by  $Ax^* = b$ . At the same time  $x = Q\hat{x} + x^*$ .

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



# 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 not affect optimization process.
- Secondly, we have a spectral decomposition of the matrix  $A$ :

$$A = Q \Lambda Q^\top$$

- Let's show, that we can switch coordinates to make an analysis a little bit easier. Let  $\hat{x} = Q^\top(x - x^*)$ , where  $x^*$  is the minimum point of initial function, defined by  $Ax^* = b$ . At the same time  $x = Q\hat{x} + x^*$ .

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



## 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  $\hat{x}$ )

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



## 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  $\hat{x}$ )

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

## 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  $\hat{x}$ )

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

## 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  $\hat{x}$ )

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

$$= (I - \alpha^k \Lambda)x^k$$

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

## 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  $\hat{x}$ )

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

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

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

## 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  $\hat{x}$ )

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

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

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

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

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

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

## 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  $\hat{x}$ )

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

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

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

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

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

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

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

## 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  $\hat{x}$ )

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

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

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

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

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

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

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

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

## 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  $\hat{x}$ )

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

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

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

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

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

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

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

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

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



## 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  $\hat{x}$ )

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

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

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

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

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

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

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

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

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

## 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  $\hat{x}$ )

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

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

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

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

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

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

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

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

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

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

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

## 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  $\hat{x}$ )

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

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

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

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

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

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

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

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

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

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

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

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

## 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  $\hat{x}$ )

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

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

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

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

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

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

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

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

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

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

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

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

## 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  $\hat{x}$ )

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

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

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

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

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

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

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

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

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

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

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

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

$\alpha < \frac{2}{L}$  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  $\hat{x}$ )

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

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

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

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

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

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

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

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

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

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

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

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

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

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

$\alpha < \frac{2}{L}$  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  $\hat{x}$ )

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

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

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

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

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

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

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

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

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

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

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

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

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

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

$\alpha < \frac{2}{L}$  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  $\hat{x}$ )

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

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

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

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

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

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

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

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

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

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

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

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

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

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

$\alpha < \frac{2}{L}$  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  $\hat{x}$ )

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

## 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  $\hat{x}$ )

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

## 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  $\hat{x}$ )

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

## 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  $\hat{x}$ )

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

## 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  $\hat{x}$ )

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

## 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.

$\kappa$	$\rho$	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 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](#)

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



# 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.

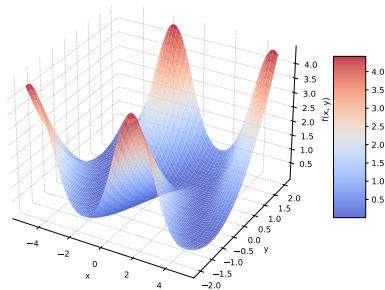
The following functions satisfy the PL condition but are not convex. [🔗Link to the code](#)

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



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

Non-convex PL function





# Convergence analysis

## Theorem

Consider the Problem

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

and assume that  $f$  is  $\mu$ -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:

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

## Convergence analysis

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

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

# Convergence analysis

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

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

# Convergence analysis

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

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

# Convergence analysis

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

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

# Convergence analysis

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

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

## Convergence analysis

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

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

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

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

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

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.



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

## 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  $y = x^*$ :

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

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

## 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  $y = x^*$ :

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

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

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

## 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  $y = x^*$ :

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

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

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

## 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  $y = x^*$ :

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

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

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

## 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  $y = x^*$ :

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

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

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

## 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:

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

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

Putting  $y = x^*$ :

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

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

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

## 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  $y = x^*$ :

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

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

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

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

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

## Smooth convex case

### Theorem

Consider the Problem

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

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

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

## Convergence analysis

- As it was before, we first use smoothness:

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

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

# Convergence analysis

- As it was before, we first use smoothness:

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

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:

(2)

# Convergence analysis

- As it was before, we first use smoothness:

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

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:

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

# Convergence analysis

- As it was before, we first use smoothness:

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

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:

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



# Convergence analysis

- As it was before, we first use smoothness:

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

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:

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

## Convergence analysis

- Now we put Equation 2 to Equation 1:

## Convergence analysis

- Now we put Equation 2 to Equation 1:

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

## Convergence analysis

- Now we put Equation 2 to Equation 1:

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

## Convergence analysis

- Now we put Equation 2 to Equation 1:

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

## 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  $b = x^k - x^* - \alpha \nabla f(x^k)$ .

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

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

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



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

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

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

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

## 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:

(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:

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

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

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

- 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:

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

# Convergence analysis

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

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

# 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:

# 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:

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



# 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:

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

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

# 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:

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

## How optimal is $\mathcal{O}\left(\frac{1}{k}\right)$ ?

- Is it somehow possible to understand, that the obtained convergence is the fastest possible with this class of problem and this class of algorithms?

## How optimal is $\mathcal{O}\left(\frac{1}{k}\right)$ ?

- Is it somehow possible to understand, that the obtained convergence is the fastest possible with this class of problem and this class of algorithms?
- The iteration of gradient descent:

$$\begin{aligned}x^{k+1} &= x^k - \alpha^k \nabla f(x^k) \\&= x^{k-1} - \alpha^{k-1} \nabla f(x^{k-1}) - \alpha^k \nabla f(x^k) \\&\vdots \\&= x^0 - \sum_{i=0}^k \alpha^{k-i} \nabla f(x^{k-i})\end{aligned}$$

## How optimal is $\mathcal{O}\left(\frac{1}{k}\right)$ ?

- Is it somehow possible to understand, that the obtained convergence is the fastest possible with this class of problem and this class of algorithms?
- The iteration of gradient descent:

$$\begin{aligned}x^{k+1} &= x^k - \alpha^k \nabla f(x^k) \\&= x^{k-1} - \alpha^{k-1} \nabla f(x^{k-1}) - \alpha^k \nabla f(x^k) \\&\vdots \\&= x^0 - \sum_{i=0}^k \alpha^{k-i} \nabla f(x^{k-i})\end{aligned}$$

- Consider a family of first-order methods, where

$$x^{k+1} \in x^0 + \text{span} \{ \nabla f(x^0), \nabla f(x^1), \dots, \nabla f(x^k) \} \quad (4)$$

## Smooth convex case

### Theorem

There exists a function  $f$  that is  $L$ -smooth and convex such that any method 4 satisfies

$$\min_{i \in [1, k]} f(x^i) - f^* \geq \frac{3L\|x^0 - x^*\|_2^2}{32(1+k)^2}$$

## Smooth convex case

### Theorem

There exists a function  $f$  that is  $L$ -smooth and convex such that any method 4 satisfies

$$\min_{i \in [1, k]} f(x^i) - f^* \geq \frac{3L\|x^0 - x^*\|_2^2}{32(1+k)^2}$$

- No matter what gradient method you provide, there is always a function  $f$  that, when you apply your gradient method on minimizing such  $f$ , the convergence rate is lower bounded as  $\mathcal{O}\left(\frac{1}{k^2}\right)$ .

## Smooth convex case

### Theorem

There exists a function  $f$  that is  $L$ -smooth and convex such that any method 4 satisfies

$$\min_{i \in [1, k]} f(x^i) - f^* \geq \frac{3L\|x^0 - x^*\|_2^2}{32(1+k)^2}$$

- No matter what gradient method you provide, there is always a function  $f$  that, when you apply your gradient method on minimizing such  $f$ , the convergence rate is lower bounded as  $\mathcal{O}\left(\frac{1}{k^2}\right)$ .
- The key to the proof is to explicitly build a special function  $f$ .



## Nesterov's worst function

- Let  $d = 2k + 1$  and  $A \in \mathbb{R}^{d \times d}$ .

$$\begin{bmatrix} 2 & -1 & 0 & 0 & \cdots & 0 \\ -1 & 2 & -1 & 0 & \cdots & 0 \\ 0 & -1 & 2 & -1 & \cdots & 0 \\ 0 & 0 & -1 & 2 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & 2 \end{bmatrix}$$

## Nesterov's worst function

- Let  $d = 2k + 1$  and  $A \in \mathbb{R}^{d \times d}$ .

$$\begin{bmatrix} 2 & -1 & 0 & 0 & \cdots & 0 \\ -1 & 2 & -1 & 0 & \cdots & 0 \\ 0 & -1 & 2 & -1 & \cdots & 0 \\ 0 & 0 & -1 & 2 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & 2 \end{bmatrix}$$

- Notice, that

$$x^T A x = x[1]^2 + x[d]^2 + \sum_{i=1}^{d-1} (x[i] - x[i+1])^2,$$

and, from this expression, it's simple to check  
 $0 \preceq A \preceq 4I$ .

## Nesterov's worst function

- Let  $d = 2k + 1$  and  $A \in \mathbb{R}^{d \times d}$ .

$$\begin{bmatrix} 2 & -1 & 0 & 0 & \cdots & 0 \\ -1 & 2 & -1 & 0 & \cdots & 0 \\ 0 & -1 & 2 & -1 & \cdots & 0 \\ 0 & 0 & -1 & 2 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & 2 \end{bmatrix}$$

- Notice, that

$$x^T A x = x[1]^2 + x[d]^2 + \sum_{i=1}^{d-1} (x[i] - x[i+1])^2,$$

and, from this expression, it's simple to check  $0 \preceq A \preceq 4I$ .

- Define the following  $L$ -smooth convex function

$$f(x) = \frac{L}{8} x^T A x - \frac{L}{4} \langle x, e_1 \rangle.$$

## Nesterov's worst function

- Let  $d = 2k + 1$  and  $A \in \mathbb{R}^{d \times d}$ .

$$\begin{bmatrix} 2 & -1 & 0 & 0 & \cdots & 0 \\ -1 & 2 & -1 & 0 & \cdots & 0 \\ 0 & -1 & 2 & -1 & \cdots & 0 \\ 0 & 0 & -1 & 2 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & 2 \end{bmatrix}$$

- Notice, that

$$x^T A x = x[1]^2 + x[d]^2 + \sum_{i=1}^{d-1} (x[i] - x[i+1])^2,$$

and, from this expression, it's simple to check  $0 \preceq A \preceq 4I$ .

- Define the following  $L$ -smooth convex function

$$f(x) = \frac{L}{8} x^T A x - \frac{L}{4} \langle x, e_1 \rangle.$$

- The optimal solution  $x^*$  satisfies  $Ax^* = e_1$ , and solving this system of equations gives

$$x^*[i] = 1 - \frac{i}{d+1},$$

## Nesterov's worst function

- Let  $d = 2k + 1$  and  $A \in \mathbb{R}^{d \times d}$ .

$$\begin{bmatrix} 2 & -1 & 0 & 0 & \cdots & 0 \\ -1 & 2 & -1 & 0 & \cdots & 0 \\ 0 & -1 & 2 & -1 & \cdots & 0 \\ 0 & 0 & -1 & 2 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & 2 \end{bmatrix}$$

- Notice, that

$$x^T A x = x[1]^2 + x[d]^2 + \sum_{i=1}^{d-1} (x[i] - x[i+1])^2,$$

and, from this expression, it's simple to check  $0 \preceq A \preceq 4I$ .

- Define the following  $L$ -smooth convex function

$$f(x) = \frac{L}{8} x^T A x - \frac{L}{4} \langle x, e_1 \rangle.$$

- The optimal solution  $x^*$  satisfies  $Ax^* = e_1$ , and solving this system of equations gives

$$x^*[i] = 1 - \frac{i}{d+1},$$

- And the objective value is

$$\begin{aligned} f(x^*) &= \frac{L}{8} x^{*T} A x^* - \frac{L}{4} \langle x^*, e_1 \rangle \\ &= -\frac{L}{8} \langle x^*, e_1 \rangle = -\frac{L}{8} \left( 1 - \frac{1}{d+1} \right). \end{aligned}$$