

## **Gradient Descent**



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

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

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

**♥ ೧ ⊘** 

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  $\alpha \to 0$ :

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

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  $\alpha \to 0$ :

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

Also from Cauchy–Bunyakovsky–Schwarz inequality:

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

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:

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:

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

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.

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  $\boldsymbol{h}$  to be a decreasing direction:

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

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

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

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

Also from Cauchy–Bunyakovsky–Schwarz inequality:

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

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

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

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:

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

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 &

 $f \rightarrow \min$ 

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

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

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

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

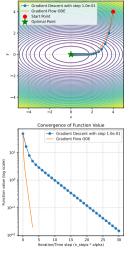
$$\frac{x_{k+1} - x_k}{\alpha} = -f'(x_k),$$

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.



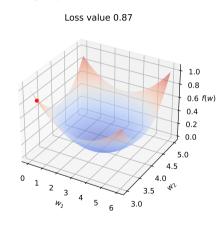
Trajectories with Contour Plot

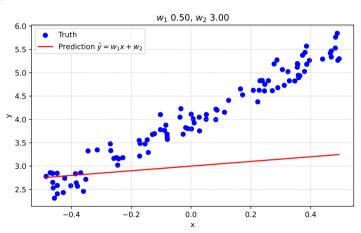
Figure 1: Gradient flow trajectory

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:

എ റ ഉ

## 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:

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

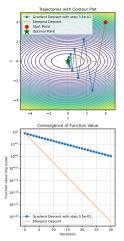


Figure 2: Steepest Descent

Open In Colab 🐥



**Strongly convex quadratics** 



Consider the following quadratic optimization problem:

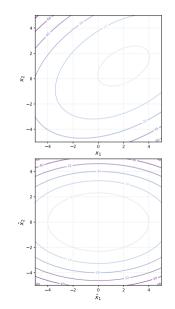
$$\min_{x \in \mathbb{R}^d} f(x) = \min_{x \in \mathbb{R}^d} \frac{1}{2} x^\top A x - b^\top x + c, \text{ where } A \in \mathbb{S}^d_{++}.$$

େ ଚ

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.



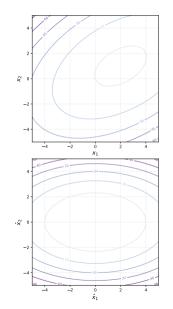
Strongly convex quadratics

Consider the following quadratic optimization problem:

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

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

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



**⊕ ೧ ⊘** 

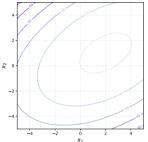
Consider the following quadratic optimization problem:

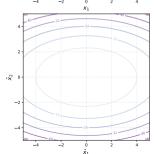
$$\min_{x \in \mathbb{R}^d} f(x) = \min_{x \in \mathbb{R}^d} \frac{1}{2} x^\top A x - b^\top x + c, \text{ where } A \in \mathbb{S}^d_{++}.$$

- $\bullet$  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:

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

• 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  $x = Q\hat{x} + x^*$ .





Consider the following quadratic optimization problem:

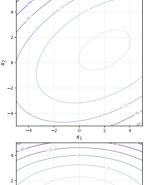
$$\min_{x \in \mathbb{R}^d} f(x) = \min_{x \in \mathbb{R}^d} \frac{1}{2} x^\top A x - b^\top x + c, \text{ where } A \in \mathbb{S}^d_{++}.$$

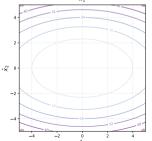
- ullet Firstly, without loss of generality we can set c=0, which will or affect optimization process.
- ullet Secondly, we have a spectral decomposition of the matrix A:

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

• 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  $x = Q\hat{x} + x^*$ .

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





Consider the following quadratic optimization problem:

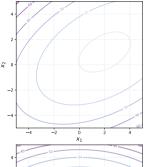
$$\min_{x \in \mathbb{R}^d} f(x) = \min_{x \in \mathbb{R}^d} \frac{1}{2} x^\top A x - b^\top x + c, \text{ where } A \in \mathbb{S}^d_{++}.$$

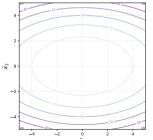
- ullet Firstly, without loss of generality we can set c=0, which will or affect optimization process.
- ullet Secondly, we have a spectral decomposition of the matrix A:

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

• 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  $x = Q\hat{x} + x^*$ .

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





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:

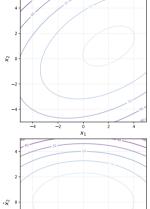
$$A = Q\Lambda Q^T$$

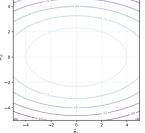
• Let's show, that we can switch coordinates to make an analysis a little bit easier. Let  $\hat{x} = Q^T(x - x^*)$ , where  $x^*$  is the minimum point of initial function, defined by  $Ax^* = b$ . At the same time  $x = Q\hat{x} + x^*$ .

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

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

$$= \frac{1}{2} \hat{x}^T \Lambda \hat{x}$$





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  $\hat{x}$ )

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

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  $\hat{x}$ )

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

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

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

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  $\hat{x}$ )

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

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  $\hat{x}$ )

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

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  $\hat{x}$ )

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

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

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

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

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  $\hat{x}$ )

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

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

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

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

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  $\hat{x}$ )

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

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

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

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

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

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 coordi

$$x_{(i)}^{k+1}=(1-lpha^k\lambda_{(i)})x_{(i)}^k$$
 For  $i$ -th coordinate  $x_{(i)}^{k+1}=(1-lpha^k\lambda_{(i)})^kx_{(i)}^0$ 

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

condition:

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

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

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

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

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  $\hat{x}$ )

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

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

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

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

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

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  $\hat{x}$ )

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

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

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

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

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

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  $\hat{x}$ )

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

condition:

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

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

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

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

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

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  $\hat{x}$ )

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

condition:

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

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

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

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

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

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  $\hat{x}$ )

$$x^{k+1} = x^k - \alpha^k \nabla f(x^k) = x^k - \alpha^k \Lambda x^k$$
$$= (I - \alpha^k \Lambda) x^k$$
$$x^{k+1} = (1 - \alpha^k \lambda x) x^k \text{. For } i \text{-th coording}$$

$$x_{(i)}^{k+1}=(1-lpha^k\lambda_{(i)})x_{(i)}^k$$
 For  $i$ -th coordinate 
$$x_{(i)}^{k+1}=(1-lpha^k\lambda_{(i)})^kx_{(i)}^0$$

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

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

$$P_{\text{composite}} = \frac{1}{i} \left( \frac{1}{i} \right)^{-1} = \frac{1}{i} \left( \frac{1}{i} \right)^{-$$

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

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

$$\begin{array}{ll} -1<1-\alpha\mu<1 & -1<1-\alpha L<1 \\ \alpha<\frac{2}{\mu} & \alpha\mu>0 & \alpha<\frac{2}{L} & \alpha L>0 \\ \alpha<\frac{2}{T} \text{ is needed for convergence.} \end{array}$$

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

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  $\hat{x}$ )

$$x_{(i)}^{k+1} = (1-\alpha^k\lambda_{(i)})^kx_{(i)}^0$$
 Let's use constant stepsize  $\alpha^k = \alpha$ . Convergence condition: 
$$\rho(\alpha) = \max_i |1-\alpha\lambda_{(i)}| < 1$$
 Remember, that  $\lambda_{\min} = \mu > 0, \lambda_{\max} = L \ge \mu$ .

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

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

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

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

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  $\hat{x}$ )

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

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

convergence rate

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

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

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

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

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

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

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

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

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

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

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

$$\alpha L < 1$$
 $\alpha L > 0$ 

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

condition:

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  $\hat{x}$ )

$$egin{aligned} x^{k+1} &= x^k - lpha^k 
abla f(x^k) = x^k - lpha^k \Lambda x^k \\ &= (I - lpha^k \Lambda) x^k \\ x^{k+1}_{(i)} &= (1 - lpha^k \lambda_{(i)}) x^k_{(i)} & ext{For $i$-th coordinate} \end{aligned}$$

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

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

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

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

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

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

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

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  $\hat{x}$ )

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

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

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

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

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

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

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

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

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

$$= \alpha L - 1$$

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

$$egin{aligned} x^{k+1} &= x^k - lpha^k 
abla f(x^k) = x^k - lpha^k \Lambda x^k \ &= (I - lpha^k \Lambda) x^k \ x_{(i)}^{k+1} &= (1 - lpha^k \lambda_{(i)}) x_{(i)}^k ext{ For $i$-th coordinate} \end{aligned}$$

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

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

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

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

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

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

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

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

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

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

$$egin{aligned} x^{k+1} &= x^k - lpha^k 
abla f(x^k) = x^k - lpha^k \Lambda x^k \ &= (I - lpha^k \Lambda) x^k \ x^{k+1}_{(i)} &= (1 - lpha^k \lambda_{(i)}) x^k_{(i)} & ext{For $i$-th coordinate} \end{aligned}$$

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

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

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

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

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

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

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

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

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

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

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

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

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:

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

convergence rate

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

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

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

$$\rho^* = \frac{L - L}{L + L}$$

$$=\frac{L-\mu}{L+\mu}$$

$$-\frac{1}{L+\mu}$$

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

$$+\mu$$

$$\vdash \mu$$
 $k+1$ 

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

$$f o \min$$
 Strongly convey guar

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.

$\kappa$	ho	Iterations to decrease domain gap $10\ \mathrm{times}$	Iterations to decrease function gap $10\ \mathrm{times}$
1.1	0.05	1	1
2	0.33	3	2
5	0.67	6	3
10	0.82	12	6
50	0.96	58	29
100	0.98	116	58
500	0.996	576	288
1000	0.998	1152	576



# Polyak-Lojasiewicz smooth case



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

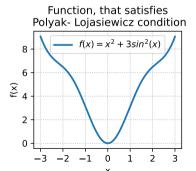
PL inequality holds if the following condition is satisfied for some  $\mu > 0$ ,

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

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

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



Polyak-Loiasiewicz smooth case



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

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

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

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

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

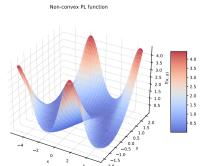
Function, that satisfies
Polyak- Lojasiewicz condition

8

6  $x = x^2 + 3sin^2(x)$ 0

-3 -2 -1 0 1 2 3

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



#### 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  $L \geq \mu > 0$ .

Consider  $(x^k)_{k\in\mathbb{N}}$  a sequence generated by the gradient descent constant stepsize algorithm, with a stepsize satisfying  $0 < \alpha \leq \frac{1}{T}$ . Then:

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



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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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:

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

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

i Theorem

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

### **Proof**

By first order strong convexity criterion:

Polyak-Loiasiewicz smooth case

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

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

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

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:

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

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

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

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

#### 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:

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

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

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

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

#### 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:

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

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

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

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

#### 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:

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

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

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

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

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

#### 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:

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

Putting  $y = x^*$ :

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

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

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

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

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



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



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



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

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

### Smooth convex case





### 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:

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

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

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

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

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

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

Typically, for the convergent gradient descent algorithm the higher the learning rate the faster the convergence. That is why we often will use  $\alpha = \frac{1}{4}$ .

• 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)

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

Typically, for the convergent gradient descent algorithm the higher the learning rate the faster the convergence. That is why we often will use  $\alpha = \frac{1}{4}$ .

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)

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

Typically, for the convergent gradient descent algorithm the higher the learning rate the faster the convergence. That is why we often will use  $\alpha = \frac{1}{4}$ .

After that we add convexity:

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

(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)

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

Typically, for the convergent gradient descent algorithm the higher the learning rate the faster the convergence. That is why we often will use  $\alpha = \frac{1}{4}$ .

After that we add convexity:

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

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

(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)

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

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

After that we add convexity:

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

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

• Now we put Equation 2 to Equation 1:





• Now we put Equation 2 to Equation 1:

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

Now we put Equation 2 to Equation 1:

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



Now we put Equation 2 to Equation 1:

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

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

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



Now we put Equation 2 to Equation 1:

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

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

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

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 Equation 2 to Equation 1:

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

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

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

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

• Now we put Equation 2 to Equation 1:

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

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

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

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

Let 
$$a = x^k - x^*$$
 and  $b = x^k - x^* - \alpha \nabla f(x^k)$ . Then  $a + b = \alpha \nabla f(x^k)$  and  $a - b = 2\left(x^k - x^* - \frac{\alpha}{2}\nabla f(x^k)\right)$ . 
$$f(x^{k+1}) \le f^* + \frac{1}{2}\left[\|x^k - x^*\|_2^2 - \|x^k - x^* - \alpha \nabla f(x^k)\|_2^2\right]$$

• Now we put Equation 2 to Equation 1:

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

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

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

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

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

• Now we put Equation 2 to Equation 1:

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

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

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

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

• Now we put Equation 2 to Equation 1:

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

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

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

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

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

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

• 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 Equation 2 to Equation 1:

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

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

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

Let  $a=x^k-x^*$  and  $b=x^k-x^*-\alpha\nabla f(x^k)$ . Then  $a+b=\alpha\nabla f(x^k)$  and  $a-b=2\left(x^k-x^*-\frac{\alpha}{2}\nabla f(x^k)\right)$ .  $f(x^{k+1}) \le f^* + \frac{1}{2\alpha} \left[ \|x^k - x^*\|_2^2 - \|x^k - x^* - \alpha \nabla f(x^k)\|_2^2 \right]$ 

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

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

 $2\alpha \sum (f(x^{i+1}) - f^*) \le ||x^0 - x^*||_2^2 - ||x^k - x^*||_2^2$ 

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

(3)

Now we put Equation 2 to Equation 1:

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

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

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

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

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

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

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

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

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

(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 Equation 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 Equation 3:

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



• 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 Equation 3:

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

• 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 Equation 3:

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

**⊕ ೧ Ø**