

# Gradient Descent. Convergence rates

## Seminar

Optimization for ML. Faculty of Computer Science. HSE University

# Gradient Descent

Suppose, we have a problem of minimization of a smooth function  $f(x) : \mathbb{R}^n \rightarrow \mathbb{R}$ :

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

# Gradient Descent

Suppose, we have a problem of minimization of a smooth function  $f(x) : \mathbb{R}^n \rightarrow \mathbb{R}$ :

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

One of the methods to solve this is **gradient descent**:

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

# Gradient Descent

Suppose, we have a problem of minimization of a smooth function  $f(x) : \mathbb{R}^n \rightarrow \mathbb{R}$ :

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

One of the methods to solve this is **gradient descent**:

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

The bottleneck (for almost all gradient methods) is choosing step-size, which can lead to the dramatic difference in method's behavior.

## How to choose step sizes

- One of the theoretical suggestions: choosing stepsize inversly proportional to the gradient Lipschitz constant

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

## How to choose step sizes

- One of the theoretical suggestions: choosing stepsize inversly proportional to the gradient Lipschitz constant

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

- **Backtracking line search.** Fix two parameters:  $0 < \beta < 1$  and  $0 < \alpha \leq 0.5$ . At each iteration, start with  $t = 1$ , and while

$$f(x_k - t\nabla f(x_k)) > f(x_k) - \alpha t \|\nabla f(x_k)\|_2^2,$$

shrink  $t = \beta t$ . Else perform Gradient Descent update  $x_{k+1} = x_k - t\nabla f(x_k)$ .

## How to choose step sizes

- One of the theoretical suggestions: choosing stepsize inversly proportional to the gradient Lipschitz constant

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

- **Backtracking line search.** Fix two parameters:  $0 < \beta < 1$  and  $0 < \alpha \leq 0.5$ . At each iteration, start with  $t = 1$ , and while

$$f(x_k - t\nabla f(x_k)) > f(x_k) - \alpha t \|\nabla f(x_k)\|_2^2,$$

shrink  $t = \beta t$ . Else perform Gradient Descent update  $x_{k+1} = x_k - t\nabla f(x_k)$ .

- **Exact line search.**

$$\eta_k = \arg \min_{\eta \geq 0} f(x_k - \eta \nabla f(x_k))$$

## 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

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

## Minimizer of Lipschitz parabola

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

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

## Minimizer of Lipschitz parabola

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

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

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

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

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



## Minimizer of Lipschitz parabola

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

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

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

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

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

Then

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

## Minimizer of Lipschitz parabola

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

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

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

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

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

Then

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

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

## Minimizer of Lipschitz parabola

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

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

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

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

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

Then

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

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



Figure 1: Illustration

## Minimizer of Lipschitz parabola

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

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

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

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

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

Then

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

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



Figure 1: Illustration

$$\nabla \phi_2(x) = 0$$

$$\nabla f(x_0) + L(x^* - x_0) = 0$$

$$x^* = x_0 - \frac{1}{L} \nabla f(x_0)$$

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

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

# Strongly convexity and Polyak - Łojasiewicz condition.

PL-condition:

$$\|\nabla f(x)\|^2 \geq 2\mu(f(x) - f^*) \quad \forall x \in \mathbb{R}^n, \mu > 0,$$

where  $f^* = f(x^*)$ ,  $x^* = \arg \min f(x)$

## Strongly convexity and Polyak - Lojasiewicz condition.

PL-condition:

$$\|\nabla f(x)\|^2 \geq 2\mu(f(x) - f^*) \quad \forall x \in \mathbb{R}^n, \mu > 0,$$

where  $f^* = f(x^*)$ ,  $x^* = \arg \min f(x)$

if  $f(x)$  is differentiable and  $\mu$  strongly convex then:

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

## Strongly convexity and Polyak - Lojasiewicz condition.

PL-condition:

$$\|\nabla f(x)\|^2 \geq 2\mu(f(x) - f^*) \quad \forall x \in \mathbb{R}^n, \mu > 0,$$

where  $f^* = f(x^*)$ ,  $x^* = \arg \min f(x)$

if  $f(x)$  is differentiable and  $\mu$  strongly convex then:

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

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

## Strongly convexity and Polyak - Lojasiewicz condition.

PL-condition:

$$\|\nabla f(x)\|^2 \geq 2\mu(f(x) - f^*) \quad \forall x \in \mathbb{R}^n, \mu > 0,$$

where  $f^* = f(x^*)$ ,  $x^* = \arg \min f(x)$

if  $f(x)$  is differentiable and  $\mu$  strongly convex then:

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

$$\begin{aligned} f(x) - f(x^*) &\leq \nabla f(x)^T(x - x^*) - \frac{\mu}{2}\|x^* - x\|^2 \leq \|\nabla f(x)\| \|x - x^*\| - \frac{\mu}{2}\|x^* - x\|^2 \\ &\leq [\text{parabola's top}] \leq \frac{\|\nabla f(x)\|^2}{2\mu} \end{aligned}$$

Thus, for a  $\mu$ -strongly convex function, the PL-condition is satisfied



## Exact line search aka steepest descent

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

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

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

Optimality conditions:

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

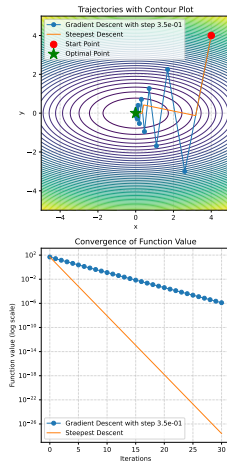



Figure 2: Steepest Descent

Open In Colab 

## Convergence analysis. Backtracking line search

Assume that  $f$  is convex, differentiable and Lipschitz gradient with constant  $L > 0$ .

### Theorem

Gradient descent with fixed step size  $t \leq 1/L$  satisfies

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

## Convergence analysis. Backtracking line search

Assume that  $f$  is convex, differentiable and Lipschitz gradient with constant  $L > 0$ .

### Theorem

Gradient descent with fixed step size  $t \leq 1/L$  satisfies

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

Let's show that the convergence rate for the Backtracking line search is no worse than  $O(1/k)$

## Convergence analysis. Backtracking line search

Assume that  $f$  is convex, differentiable and Lipschitz gradient with constant  $L > 0$ .

### Theorem

Gradient descent with fixed step size  $t \leq 1/L$  satisfies

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

Let's show that the convergence rate for the Backtracking line search is no worse than  $O(1/k)$

Since  $\nabla f$  is Lipschitz continuous with constant  $L > 0$ , we have

$$f(y) \leq f(x) + \nabla f(x)^T (y - x) + \frac{L}{2} \|y - x\|_2^2, \forall x, y$$

## Convergence analysis. Backtracking line search

Assume that  $f$  is convex, differentiable and Lipschitz gradient with constant  $L > 0$ .

### Theorem

Gradient descent with fixed step size  $t \leq 1/L$  satisfies

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

Let's show that the convergence rate for the Backtracking line search is no worse than  $O(1/k)$

Since  $\nabla f$  is Lipschitz continuous with constant  $L > 0$ , we have

$$f(y) \leq f(x) + \nabla f(x)^T (y - x) + \frac{L}{2} \|y - x\|_2^2, \forall x, y$$

Let  $y = x^+ = x - t\nabla f(x)$ , then:

$$f(x^+) \leq f(x) - \left(1 - \frac{Lt}{2}\right) t \|\nabla f(x)\|_2^2 \leq f(x) - \frac{1}{2L} \|\nabla f(x)\|_2^2$$

## Convergence analysis. Backtracking line search

Assume that  $f$  is convex, differentiable and Lipschitz gradient with constant  $L > 0$ .

### Theorem

Gradient descent with fixed step size  $t \leq 1/L$  satisfies

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

Let's show that the convergence rate for the Backtracking line search is no worse than  $O(1/k)$

Since  $\nabla f$  is Lipschitz continuous with constant  $L > 0$ , we have

$$f(y) \leq f(x) + \nabla f(x)^T (y - x) + \frac{L}{2} \|y - x\|_2^2, \forall x, y$$

Let  $y = x^+ = x - t\nabla f(x)$ , then:

$$f(x^+) \leq f(x) - \left(1 - \frac{Lt}{2}\right) t \|\nabla f(x)\|_2^2 \leq f(x) - \frac{1}{2L} \|\nabla f(x)\|_2^2$$

This recalls us the stopping condition in Backtracking line search when  $\alpha = 0.5, t = \frac{1}{L}$ . Hence, Backtracking line search with  $\alpha = 0.5$  plus condition of Lipschitz gradient will guarantee us the convergence rate of  $O(1/k)$ .

# Python Examples

Why convexity and strong convexity is important? Check the simple  code snippet.

Cool illustration of gradient descent 

Lipschitz constant for linear regression 