# Introduction to Optimization

**Lecture 05: Gradient descent.**
**Convex functions.**

university of
groningen

# Recall: descent methods

Most algorithms are based on the idea of (sufficient) descent: given $x_k$, find $x_{k+1}$ such that

$$f(x_{k+1}) \leq f(x_k) - \delta_k^2. \tag{1}$$

One way is to find $d_k \in \mathbb{R}^N$ and $\alpha_k > 0$, such that (1) holds with

$$x_{k+1} = x_k - \alpha_k d_k.$$

We say $-d_k$ is a descent direction, and $\alpha_k$ is the step size, step length or learning rate (in ML).

# $L$-smoothness

A differentiable function $f : A \subset \mathbb{R}^N \to \mathbb{R}$ is $L$-smooth, with $L > 0$, if

$$\|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\|$$

for all $x, y \in A$.

# $L$-smoothness

A differentiable function $f : A \subset \mathbb{R}^N \to \mathbb{R}$ is $L$-smooth, with $L > 0$, if

$$\|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\|$$

for all $x, y \in A$.

**Proposition (Descent Lemma)**

*If $f$ is $L$-smooth and $A$ is convex, then*

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

*for all $x, y \in A$.*

By the Descent Lemma, we have

$$f(x_n - \alpha_n d_n) \leq f(x_n) - \alpha_n \nabla f(x_n) \cdot d_n + \frac{\alpha_n^2 L}{2} \|d_n\|^2.$$

# Identifying descent directions

## The $L-$smooth case

By the Descent Lemma, we have

$$f(x_n - \alpha_n d_n) \leq f(x_n) - \alpha_n \nabla f(x_n) \cdot d_n + \frac{\alpha_n^2 L}{2} \|d_n\|^2.$$

It is therefore sufficient for $\alpha_n$ and $d_n$ to satisfy

$$0 < \alpha_n L \|d_n\|^2 < 2 \nabla f(x_n) \cdot d_n.$$

# Identifying descent directions

## The $L-$smooth case

By the Descent Lemma, we have

$$f(x_n - \alpha_n d_n) \leq f(x_n) - \alpha_n \nabla f(x_n) \cdot d_n + \frac{\alpha_n^2 L}{2} \|d_n\|^2.$$

It is therefore sufficient for $\alpha_n$ and $d_n$ to satisfy

$$0 < \alpha_n L \|d_n\|^2 < 2 \nabla f(x_n) \cdot d_n.$$

Gradient-consistency: $\tau \, \|d_n\|^2 \leq \|\nabla f(x_n)\|^2 \leq \sigma \, \nabla f(x_n) \cdot d_n.$

# Identifying descent directions

## The $L-$smooth case

By the Descent Lemma, we have

$$f(x_n - \alpha_n d_n) \leq f(x_n) - \alpha_n \nabla f(x_n) \cdot d_n + \frac{\alpha_n^2 L}{2} \|d_n\|^2.$$

It is therefore sufficient for $\alpha_n$ and $d_n$ to satisfy

$$0 < \alpha_n L \|d_n\|^2 < 2 \nabla f(x_n) \cdot d_n.$$

Gradient-consistency: $\tau \, \|d_n\|^2 \leq \|\nabla f(x_n)\|^2 \leq \sigma \, \nabla f(x_n) \cdot d_n$.

Sufficient condition for descent: $0 < \inf \alpha_n \leq \sup \alpha_n < \dfrac{2\tau}{\sigma L}$.

## Gradient-consistent methods

Constant stepsize $\alpha_n \equiv \alpha$, for simplicity

### Proposition

*Let $f$ be $L$-smooth and bounded from below. Iterate $x_{n+1} = x_n - \alpha d_n$, where $d_n$ is gradient-consistent and $0 < \alpha < \frac{2\tau}{\sigma L}$.*

# Gradient-consistent methods

Constant stepsize $\alpha_n \equiv \alpha$, for simplicity

## Proposition

*Let $f$ be L-smooth and bounded from below. Iterate $x_{n+1} = x_n - \alpha d_n$, where $d_n$ is gradient-consistent and $0 < \alpha < \frac{2\tau}{\sigma L}$. Then,*

1. $\exists \lim_{n\to\infty} f(x_n) \in \mathbb{R}$, *and* $\lim_{n\to\infty} \|\nabla f(x_n)\| = 0$.

# Gradient-consistent methods

Constant stepsize $\alpha_n \equiv \alpha$, for simplicity

## Proposition

*Let $f$ be L-smooth and bounded from below. Iterate $x_{n+1} = x_n - \alpha d_n$, where $d_n$ is gradient-consistent and $0 < \alpha < \frac{2\tau}{\sigma L}$. Then,*

1. $\exists \lim\limits_{n\to\infty} f(x_n) \in \mathbb{R}$, *and* $\lim\limits_{n\to\infty} \|\nabla f(x_n)\| = 0$.
2. *Cluster points are critical: if $x_{k_n} \to \hat{x}$, then $\nabla f(\hat{x}) = 0$.*

## Gradient-consistent methods

Constant stepsize $\alpha_n \equiv \alpha$, for simplicity

### Proposition

*Let $f$ be $L$-smooth and bounded from below. Iterate $x_{n+1} = x_n - \alpha d_n$, where $d_n$ is gradient-consistent and $0 < \alpha < \frac{2\tau}{\sigma L}$. Then,*

1. $\exists \lim_{n\to\infty} f(x_n) \in \mathbb{R}$, *and* $\lim_{n\to\infty} \|\nabla f(x_n)\| = 0$.
2. *Cluster points are critical: if $x_{k_n} \to \hat{x}$, then $\nabla f(\hat{x}) = 0$.*
3. *If $f$ has no critical points, then $\lim_{n\to\infty} \|x_n\| = +\infty$.*

# Gradient-consistent methods

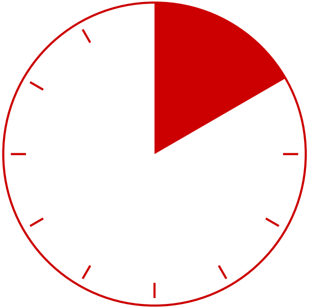
Constant stepsize $\alpha_n \equiv \alpha$, for simplicity

## Proposition

*Let $f$ be $L$-smooth and bounded from below. Iterate $x_{n+1} = x_n - \alpha d_n$, where $d_n$ is gradient-consistent and $0 < \alpha < \frac{2\tau}{\sigma L}$. Then,*

1. $\exists \lim_{n\to\infty} f(x_n) \in \mathbb{R}$, *and* $\lim_{n\to\infty} \|\nabla f(x_n)\| = 0$.
2. *Cluster points are critical: if $x_{k_n} \to \hat{x}$, then $\nabla f(\hat{x}) = 0$.*
3. *If $f$ has no critical points, then $\lim_{n\to\infty} \|x_n\| = +\infty$.*
4. *There is $C > 0$ such that* $\min\left\{\|\nabla f(x_i)\| : 1 \le i \le n\right\} \le \dfrac{C}{\sqrt{n}}$.

# Break



# Convex functions and the gradient method

**Theorem**

*Let $f$ be convex and $L$-smooth. Iterate* $x_{n+1} = x_n - \alpha \nabla f(x_n)$ *with* $0 < \alpha < \frac{2}{L}$.

# Convex functions and the gradient method

**Theorem**

*Let $f$ be convex and $L$-smooth. Iterate* $x_{n+1} = x_n - \alpha \nabla f(x_n)$ *with* $0 < \alpha < \frac{2}{L}$.

1. $\lim\limits_{n\to\infty} f(x_n) = \inf(f)$.

# Convex functions and the gradient method

## Theorem

*Let $f$ be convex and $L$-smooth. Iterate* $x_{n+1} = x_n - \alpha \nabla f(x_n)$ *with* $0 < \alpha < \frac{2}{L}$.

1. $\lim_{n \to \infty} f(x_n) = \inf(f)$.
2. *If $f$ has minimizers, then $x_n$ converges to one of them, and*

$$f(x_n) - \min(f) \leq \frac{D^2}{\alpha(2 - \alpha L)n},$$

*where $D$ is the distance from $x_0$ to its closest minimizer.*

*Moreover,* $\lim_{n \to \infty} n \left[ f(x_n) - \min(f) \right] = 0$.

# An important tool

**Proposition (Baillon-Haddad Lemma)**

*If $f$ is convex and $L$-smooth, then*

$$\frac{1}{L}|\nabla f(y) - \nabla f(x)|^2 \leq \big(\nabla f(y) - \nabla f(x)\big) \cdot (y - x)$$

*for all $x, y \in A$.*

This will be proved in the tutorial.

## Sketch of the proof

First, use

$$(2) \qquad \|x_{n+1} - u\|^2 = \alpha^2 \|\nabla f(x_n)\|^2 + \|x_n - u\|^2 - 2\alpha \nabla f(x_n) \cdot (x_n - u)$$

and Baillon-Haddad Lemma to show that

$$\frac{\alpha}{L}(2 - \alpha L) \sum_{n=0}^{k} \|\nabla f(x_n)\|^2 \leq \|x_0 - u\|^2.$$

Then, use (2) and convexity to prove that

$$2\alpha\big(f(x_n) - \min(f)\big) \leq \|x_n - u\|^2 - \|x_{n+1}\|^2 + \alpha^2 \|\nabla f(x_n)\|^2.$$

Sum over $n$, and combine the two inequalitites, to conclude that

$$2\alpha k\big(f(x_k) - \min(f)\big) \leq \|x_0 - u\|^2 \left(1 + \frac{\alpha L}{2 - \alpha L}\right).$$

# Sketch of the proof, continued

For the convergence, use that cluster points are critical, and that $\|x_n - u\|$ is nonincreasing, to deduce that $(x_n)$ cannot have more than one cluster point. This is sufficient for convergence because $(x_n)$ is bounded.

For the last statement, use the following lemma with $e_n = f(x_n) - \min(f)$:

**Lemma**

*Let $(e_n)$ be a positive, nonincreasing sequence such that $\sum_{n=0}^{\infty} e_n < +\infty$. Then $\lim_{n\to\infty} ne_n = 0$.*

# Other choices for $\alpha_n$

- Constant: $\alpha_n \equiv \alpha \in (0, 2/L)$.

# Other choices for $\alpha_n$

- Constant: $\alpha_n \equiv \alpha \in (0, 2/L)$.
- Exact minimization: $\min_{\alpha>0} f(x_n - \alpha d_n)$.

# Other choices for $\alpha_n$

- Constant: $\alpha_n \equiv \alpha \in (0, 2/L)$.
- Exact minimization: $\min_{\alpha>0} f(x_n - \alpha d_n)$.
- Limited minimization: $\min_{\alpha\in(0,A]} f(x_n - \alpha d_n)$.

# Other choices for $\alpha_n$

- Constant: $\alpha_n \equiv \alpha \in (0, 2/L)$.
- Exact minimization: $\min_{\alpha>0} f(x_n - \alpha d_n)$.
- Limited minimization: $\min_{\alpha\in(0,A]} f(x_n - \alpha d_n)$.
- Vanishing: $\alpha_n \to 0$, $\sum \alpha_n = \infty$.

# Other choices for $\alpha_n$

- Constant: $\alpha_n \equiv \alpha \in (0, 2/L)$.
- Exact minimization: $\min_{\alpha>0} f(x_n - \alpha d_n)$.
- Limited minimization: $\min_{\alpha\in(0,A]} f(x_n - \alpha d_n)$.
- Vanishing: $\alpha_n \to 0$, $\sum \alpha_n = \infty$.
- Backtracking (Armijo, Goldstein).

## Some remarks

### The simplest example, revisited

We had applied the gradient method to the function $f(x) = x^2$. Were the hypotheses on $\alpha$ and the rate of convergence consistent with the previous theorem?

# Some remarks

### The simplest example, revisited

We had applied the gradient method to the function $f(x) = x^2$. Were the hypotheses on $\alpha$ and the rate of convergence consistent with the previous theorem?

### Strong convexity

If the objective function is strongly convex, can we expect the gradient method to converge faster?