Optimization and Data Science

Lecture 10: Convergence of Descent Methods

Prof. Dr. Thomas Slawig

Kiel University - CAU Kiel
Dep. of Computer Science

Summer 2020

Contents

Recall: General descent method with efficient step-size

Algorithm (General descent method):

1. Choose initial guess $x_0 \in \mathbb{R}^n$.
2. For $k = 0, 1, \ldots$:
   1. Choose a descent direction $d_k \in \mathbb{R}^n$.
   2. Choose an efficient step-size $\rho_k > 0$, i.e., one that satisfies

$$f(x_k + \rho_k d_k) \leq f(x_k) - c_S \left(\frac{\nabla f(x_k)^\top d_k}{\|d_k\|} \right)^2$$

      with $c_S > 0$ independent of k.
   3. Set $x_{k+1} = x_k + \rho_k d_k$.

   until a stopping criterion is satisfied.

- $\nabla f(x_k) \neq 0 \rightsquigarrow$ negative gradient $d_k = -\nabla f(x_k)$ is a descent direction (gradient method).

Contents

For convergence: gradient-related directions

Definition

For a sequence $(x_k)_{k\in\mathbb{N}} \subset \mathbb{R}^n$ of iterates with $\nabla f(x_k) \neq 0$ for all $k \in \mathbb{N}$, the sequence of directions $(d_k)_{k\in\mathbb{N}} \subset \mathbb{R}^n$ is called **gradient-related**, if there exists $c_D > 0$ such that

$$-\frac{\nabla f(x_k)^\top d_k}{\|\nabla f(x_k)\|\|d_k\|} \geq c_D \quad \text{for all } k \in \mathbb{N}. \tag{1}$$

- Again the constant $c_D > 0$ has to be independent of k.
- The norm to be considered is arbitrary, since in $\mathbb{R}^n$ all norms are equivalent, i.e., for any two vector norms $\|\cdot\|_*, \|\cdot\|_\#$ there exist $c, C \in \mathbb{R}_{>0}$ with

$$c\|x\|_* \leq \|x\|_\# \leq C\|x\|_* \quad \text{for all } x \in \mathbb{R}^n.$$

- (1) implies that the directions are descent directions in the sense of our definiton. Why?

What does "gradient-related" mean?

- Let us consider the Euclidean norm $\|\cdot\|_2$ in the definition.
- Recall: Definition of the cosine of the angle $\phi := \angle(v, w)$ between two vectors $v, w \in \mathbb{R}^n$:

$$\cos\phi := \frac{v^\top w}{\|v\|_2 \|w\|_2}$$

- Sometimes used to define scalar product:

$$v^\top w = \|v\|_2 \|w\|_2 \cos\phi.$$

- Gradient-related:

$$\underbrace{-\frac{\nabla f(x_k)^\top d_k}{\|\nabla f(x_k)\|_2 \|d_k\|_2}}_{=\cos\angle(-\nabla f(x_k), d_k)} \geq c_D > 0 \quad \text{for all } k \in \mathbb{N}.$$

- When is

$$\cos\angle(-\nabla f(x_k), d_k) \geq c_D > 0 \quad \text{for all } k \in \mathbb{N}?$$

What does "gradient-related" mean?

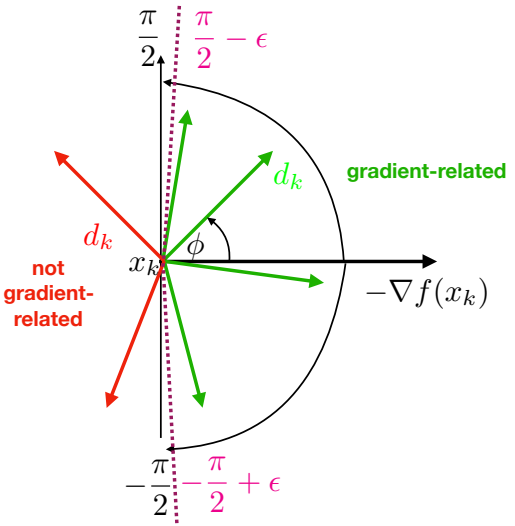
- Gradient-related means:

$$\exists c_D > 0 : \ \cos \angle(-\nabla f(x_k), d_k) \geq c_D \quad \text{for all } k \in \mathbb{N}.$$

- This can be expressed equivalently as:

$$\exists \epsilon > 0 : \ -\frac{\pi}{2} + \epsilon < \angle(-\nabla f(x_k), d_k) < \frac{\pi}{2} - \epsilon \quad \text{for all } k \in \mathbb{N}.$$

- $\rightsquigarrow$ Gradient-related directions d_k must not be orthogonal to the negative gradient, ...
- also not in the limit $k \to \infty$.



Contents

Convergence result for general descent method

Theorem

Let

- *f be differentiable and bounded from below,*
- *the sequence of directions $(d_k)_{k\in\mathbb{N}}$ be gradient-related,*
- *the sequence of step-sizes $(\rho_k)_{k\in\mathbb{N}}$ be efficient.*

Then the descent method generates a sequence $(x_k)_{k\in\mathbb{N}}$ that satisfies

- *either $\nabla f(x_k) = 0$ for some $k \in \mathbb{N}$*
- *or $\lim_{k\to\infty} \nabla f(x_k) = 0$.*

- Important: The theorem does not state that the sequence $(x_k)_{k\in\mathbb{N}}$ itself converges!
- With stronger assumptions we get a stronger result.

Proof (1)

- Since f is bounded from below, there exists $f^* \in \mathbb{R}$ with

$$-\infty < f^* \leq f(x_k) \text{ for all } k \in \mathbb{N}.$$

- Fix any $\ell \in \mathbb{N}$. Then

$$-\infty < f^* - f(x_0) \leq f(x_\ell) - f(x_0).$$

- We have

$$f(x_\ell) - f(x_0) = \sum_{k=0}^{\ell-1} (f(x_{k+1}) - f(x_k)) = \sum_{k=0}^{\ell-1} (f(x_k + \rho_k d_k) - f(x_k)).$$

- Since the step-sizes are efficient, there exist $c_S > 0$ independent of k(!!!) with

$$f^* - f(x_0) \leq f(x_\ell) - f(x_0) \leq -c_S \sum_{k=0}^{\ell-1} \left(\frac{\nabla f(x_k)^\top d_k}{\|d_k\|} \right)^2.$$

Proof (2)

- From

$$f^* - f(x_0) \leq f(x_\ell) - f(x_0) \leq -c_S \sum_{k=0}^{\ell-1} \left(\frac{\nabla f(x_k)^\top d_k}{\|d_k\|} \right)^2$$

- we get by division by $(-c_S) < 0$:

$$0 \leq \sum_{k=0}^{\ell-1} \left(\frac{\nabla f(x_k)^\top d_k}{\|d_k\|} \right)^2 \leq \frac{f^* - f(x_0)}{-c_S}.$$

- We pass to the limit $\ell \to \infty$:

$$\sum_{k=0}^{\infty} \left(\frac{\nabla f(x_k)^\top d_k}{\|d_k\|} \right)^2 \leq \underbrace{\frac{f^* - f(x_0)}{-c_S}}_{\text{constant w.r.t. } \ell} < \infty.$$

Proof (3)

- We have

$$\sum_{k=0}^{\infty} \left(\frac{\nabla f(x_k)^\top d_k}{\|d_k\|} \right)^2 < \infty.$$

- The infinite series converges, thus

$$\lim_{k\to\infty} \left(\frac{\nabla f(x_k)^\top d_k}{\|d_k\|} \right)^2 = 0 \text{ and, taking the square root: } \lim_{k\to\infty} \frac{\nabla f(x_k)^\top d_k}{\|d_k\|} = 0.$$

- The directions are gradient-related. Thus

$$\frac{\nabla f(x_k)^\top d_k}{\|d_k\|} = \underbrace{-\left(-\frac{\nabla f(x_k)^\top d_k}{\|\nabla f(x_k)\| \|d_k\|} \right)}_{\geq c_D > 0} \|\nabla f(x_k)\| \to 0 \text{ for } \ell \to \infty.$$

- Thus

$$\lim_{k\to\infty} \nabla f(x_k) = 0.$$

□

Slightly stronger convergence result for general descent method

Theorem

Let

- f *be continuously differentiable* ~~*and bounded from below*~~,
- x_0 *chosen such that the level set* $L(x_0) := \{x \in \mathbb{R}^n : f(x) \leq f(x_0)\}$ *is bounded,*
- *the sequence of directions* $(d_k)_{k\in\mathbb{N}}$ *be gradient-related,*
- *the sequence of step-sizes* $(\rho_k)_{k\in\mathbb{N}}$ *be efficient.*

Then the descent method generates a sequence $(x_k)_{k\in\mathbb{N}}$ *that satisfies*

- *either* $\nabla f(x_k) = 0$ *for some* $k \in \mathbb{N}$
- *or every accumulation point* $\bar{x}$ *satisfies* $\nabla f(\bar{x}) = 0$ *(first order necessary condition).*

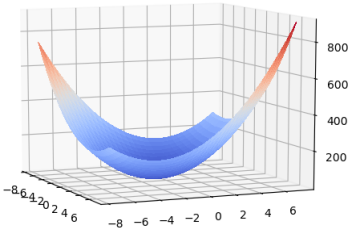
- Level set $L(x_0)$ is closed and bounded $\rightsquigarrow$ exists a minimum of f $\rightsquigarrow$ f bounded from below.
- $(x_k)_{k\in\mathbb{N}} \subset L(x_0)$ $\rightsquigarrow$ exists converging sub-sequence (to an accumulation point $\bar{x}$).
- f continuously differentiable $\rightsquigarrow$ $\lim\limits_{x_k \to \bar{x}} \nabla f(x_k) = \nabla f(\bar{x}) = 0$.

Contents

Optimality conditions for quadratic functions

- General quadratic function $f : \mathbb{R}^n \to \mathbb{R}$:

$$f(x) = \frac{1}{2} x^\top A x + b^\top x + c, \quad A \in \mathbb{R}^{n \times n}, b \in \mathbb{R}^n, c \in \mathbb{R}.$$



- A symmetric $\rightsquigarrow$ gradient: $\nabla f(x) = Ax + b$.
- $\rightsquigarrow$ first order necessary condition for a minimizer:

$$Ax + b = 0.$$

- A regular $\rightsquigarrow$ only candidate for a minimizer: $x^* = -A^{-1}b$.
- Hessian matrix:

$$\nabla^2 f(x) = A.$$

- If A is positive semi-definite, then there exists a minimum.
- If A is positive definite, then $x^* = -A^{-1}b$ is the unique minimizer.

Exact step-size for quadratic function

- For a quadratic function (with A symmetric), we can easily compute the exact step-size, i.e., the minimizer of $\Phi : \mathbb{R} \to \mathbb{R}$ defined by

$$\Phi(\rho) := f(x_k + \rho d_k)$$

- Derivative of Φ by the chain rule!

$$\Phi'(\rho) = \frac{d}{d\rho} f(x_k + \rho d_k) = \nabla f(x_k + \rho d_k)^\top d_k = (A(x_k + \rho d_k) + b)^\top d_k$$
$$= (Ax_k + b)^\top d_k + \rho d_k^\top A d_k$$

- Compute its root:

$$\rho_k = -\frac{(Ax_k + b)^\top d_k}{d_k^\top A d_k}.$$

- The exact step-size is well-defined if the denominator is $\neq 0$ which is the case if A is positive definite ($\rightsquigarrow f$ has a minimum).

Convergence Speed: Quadratic Function

- We consider the example

$$A = \begin{pmatrix} 2 & 0 \\ 0 & 2a \end{pmatrix}, b = 0, c = 0$$

  with some $a > 1$ ($\Rightarrow A$ positive definite).
- Compute the gradient:

$$\nabla f(x) = Ax = \begin{pmatrix} 2x_1 \\ 2ax_2 \end{pmatrix}$$

- Compute first search direction for the gradient method for initial guess $x_0 = (a, 1)^T$:

$$d_0 = -Ax_0 = \begin{pmatrix} -2a \\ -2a \end{pmatrix}.$$

Convergence Speed: Quadratic Function (2)

- Exact step-size for quadratic function:

$$\rho_k = -\frac{(Ax_k + b)^\top d_k}{d_k^\top A d_k} = \frac{d_k^\top d_k}{d_k^\top A d_k}.$$

- Compute the step-size in the first iteration,

$$A = \begin{pmatrix} 2 & 0 \\ 0 & 2a \end{pmatrix}, d_0 = \begin{pmatrix} -2a \\ -2a \end{pmatrix}.$$

- We get

$$d_0^\top d_0 = 8a^2, \quad Ad_0 = \begin{pmatrix} -4a \\ -4a^2 \end{pmatrix}, \quad d_0^\top A d_0 = 8(a^2 + a^3)$$

$$\rho_0 = \frac{d_0^\top d_0}{d_0^\top A d_0} = \frac{8a^2}{8(a^2 + a^3)} = \frac{1}{1 + a}$$

Convergence Speed: Quadratic Function (3)

- First iterate:

$$x_1 = \begin{pmatrix} a \\ 1 \end{pmatrix} + \frac{1}{1+a} \begin{pmatrix} -2a \\ -2a \end{pmatrix} = \frac{a-1}{a+1} \begin{pmatrix} a \\ -1 \end{pmatrix}$$

- Compute next search direction d_1:

$$d_1 = \frac{a-1}{a+1} \begin{pmatrix} -2a \\ 2a \end{pmatrix}$$

- Compute next step-size ρ_1:

$$\rho_1 = \frac{\|d_1\|_2^2}{d_1^T A d_1} = \frac{1}{1+a}$$

- Compute second iterate:

$$x_2 = \left(\frac{a-1}{a+1}\right)^2 \begin{pmatrix} a \\ 1 \end{pmatrix} = \left(\frac{a-1}{a+1}\right)^2 x_0$$

Convergence Speed: Quadratic Function (4)

- We have

$$x_0 = \begin{pmatrix} a \\ 1 \end{pmatrix}, \quad x_1 = \frac{a-1}{a+1}\begin{pmatrix} a \\ -1 \end{pmatrix}, \quad x_2 = \left(\frac{a-1}{a+1}\right)^2 \begin{pmatrix} a \\ 1 \end{pmatrix} = \left(\frac{a-1}{a+1}\right)^2 x_0$$

- We see (and can show by induction) that

$$x_k = \left(\frac{a-1}{a+1}\right)^k \begin{pmatrix} a \\ 1 \end{pmatrix}, \quad d_k = -Ax_k = \left(\frac{a-1}{a+1}\right)^k \begin{pmatrix} -2a \\ -2a \end{pmatrix}, \quad k \text{ even}$$

$$x_k = \left(\frac{a-1}{a+1}\right)^k \begin{pmatrix} a \\ -1 \end{pmatrix}, \quad d_k = -Ax_k = \left(\frac{a-1}{a+1}\right)^k \begin{pmatrix} -2a \\ 2a \end{pmatrix}, \quad k \text{ odd}$$

$$\rho_k = \frac{1}{1+a}$$

Convergence Speed: Quadratic Function (5)

- We have

$$x_k = \left(\frac{a-1}{a+1}\right)^k \begin{pmatrix} a \\ 1 \end{pmatrix}, k \text{ even}, \quad x_k = \left(\frac{a-1}{a+1}\right)^k \begin{pmatrix} a \\ -1 \end{pmatrix}, k \text{ odd},$$

- Because of

$$\left\| \begin{pmatrix} a \\ 1 \end{pmatrix} \right\|_2 = \sqrt{a^2+1} = \left\| \begin{pmatrix} a \\ -1 \end{pmatrix} \right\|_2$$

- ... we get:

$$\|x_k\|_2 = \left(\frac{a-1}{a+1}\right)^k \sqrt{a^2+1} \quad \text{for all } k \in \mathbb{N}.$$

- Moreover for $a > 1$:

$$\left(\frac{a-1}{a+1}\right)^k \to 0 \text{ for } k \to \infty \quad \Rightarrow \quad x_k \to x^* = 0 \text{ for } k \to \infty.$$

Convergence Speed: Quadratic Function (6)

- From
$$\|x_k\|_2 = \left(\frac{a-1}{a+1}\right)^k \sqrt{a^2+1} \quad \text{for all } k \in \mathbb{N}$$
- ... we get with $x^* = 0$:
$$\|x_{k+1} - x^*\|_2 = \frac{a-1}{a+1}\|x_k - x^*\|_2 \quad \text{for all } k \in \mathbb{N}$$
or
$$\frac{\|x_{k+1} - x^*\|_2}{\|x_k - x^*\|_2} = \frac{a-1}{a+1} \quad \text{for all } k \in \mathbb{N}$$
- We call this factor the Q-**factor** (quotient factor).
- How does the convergence speed depend on a in
$$A = \begin{pmatrix} 2 & 0 \\ 0 & 2a \end{pmatrix}?$$
- Fast if $a = 1 : Q = 0$, slow if $a \gg 1 : Q \approx 1$.

How can this be generalized (at first for general matrices)?

- Critical properties: eigenvalues of A.
- Recall: $\lambda \in \mathbb{R}$ is an **eigenvalue** of A, if there exists an **eigenvector** $x \in \mathbb{R}^n, x \neq 0$ with $Ax = \lambda x$.
- In our example: eigenvalues of $A = \begin{pmatrix} 2 & 0 \\ 0 & 2a \end{pmatrix}$?

$$\lambda_1 = 2, \lambda_2 = 2a.$$

- Consider $a > 1$:

$$\lambda_1 = 2 = \lambda_{min} := \min\{\lambda : \lambda \text{ eigenvalue of } A\},$$
$$\lambda_2 = 2a = \lambda_{max} := \max\{\lambda : \lambda \text{ eigenvalue of } A\}.$$

- Q-factor:

$$Q = \frac{a-1}{a+1} = \frac{2a-2}{2a+2} = \frac{\lambda_{max} - \lambda_{min}}{\lambda_{max} + \lambda_{min}}.$$

Convergence speed gradient method, quadratic functions

- For a general symmetric positive definite matrix, a similar estimate holds:

Theorem

For a quadratic function with symmetric positive definite matrix A the gradient method with exact step-size has the Q-factor (w.r.t. the norm $\|x\|_A := \sqrt{x^\top Ax}$):

$$Q_{\|\cdot\|_A} = \frac{\lambda_{max} - \lambda_{min}}{\lambda_{max} + \lambda_{min}} = \frac{cond(A) - 1}{cond(A) + 1},$$

where

- $\lambda_{min}, \lambda_{max}$ *are the smallest and biggest eigenvalue of A, respectively,*
- $cond(A) := \dfrac{\lambda_{max}(A)}{\lambda_{min}(A)} \geq 1$ *is the* **condition number** *of A.*

Contents

How can this be generalized for non-quadratic functions?

- Idea: Consider the approximation of f by a quadratic function, i.e. using Taylor formula:

$$f(x_k + s) = f(x_k) + \nabla f(x_k)^\top s + \frac{1}{2} s^\top \nabla^2 f(x_k) s + \mathcal{O}(\|s\|^3)$$

- with the Hessian matrix:

$$\nabla^2 f(x) := \left(\frac{\partial^2 f}{\partial x_i \partial x_j}(x) \right)_{i,j=1,\dots,n} \in \mathbb{R}^{n \times n}$$

- In the vicinity of x_k (with $\|s\|$ small), this is a good approximation of f.
- $\rightsquigarrow$ eigenvalues/condition number of $\nabla^2 f(x_k)$ are crucial.

What is important

- For the convergence of a general descent method, we need efficient step-sizes and gradient-related search directions.
- Gradient-related means that the search directions never become orthogonal to the negative gradient (also not in the limit).
- Then we obtain (under some assumptions on f) that the generated sequence satisfies $\nabla f(x_k) \to 0$ ...
- ... or (under a little bit more assumptions) that the sequence has accumulation points that satisfy the first order condition.
- To get convergence of the sequence of iterates, we need more assumptions.
- For quadratic functions with symmetric positive definite matrix, the convergence speed depends on the relation of biggest and smallest eigenvalue of the matrix.
- For general non-linear cost functions, the eigenvalues of the Hessian matrix are important.