Kutaisi International University | Introduction to Optimization

Prof. Dr. Dr. h.c. Florian Rupp
Akaki Matcharashvili

Spring Term
Week 4

Convergence of the Gradient Descent Method

This exercise sheet consists of three parts: at first problems for the Additional/ Central Exercise Problems class are given. Their solution will be provided and can serve you as further blueprints when solving similar tasks, e.g. for the homework assignment. Then, the actual Homework Assignments are stated that will be discussed during the TTF in the following week. Please, hand-in your results of these assignments through MSTeams at the date and time specified in MSTeams. Finally, the third part consists of Graded Homework Assignments that will be corrected and contribute to the continuous assessment of our course. Please, hand-in your results of these assignments as well through MSTeams at the date and time specified in MSTeams.

Central Exercise Problems:

Exercise 4.1: Steepest Descent w.r.t Different Norms — Consider a continuously differentiable function $f : \mathbb{R}^n \to \mathbb{R}$, a point $x \in \mathbb{R}^n$ with $\nabla f(x) \neq 0$ and a symmetric positive definite matrix $A \in \mathbb{R}^{n \times n}$. Moreover, consider the scalar product, which is induced by a A

$$\langle d_1 \, , \, d_2 \rangle_A \ := \ d_1^T A d_2 \qquad \text{for all } d_1, d_2 \in \mathbb{R}^n$$

and the corresponding norm $\|d\|_A := \sqrt{\langle d \, , \, d\rangle_A}$. Compute the *normalized direction of steepest descent of f in x with respect to the norm* $\| \cdot \|_A$, i.e. the solution to the problem

$$\min_{\|d\|_A = 1} \ \nabla f(x)^T d \, .$$

Exercise 4.2: The Compass-Search Algorithm — Let $f : \mathbb{R}^2 \to \mathbb{R}$ be continuously differentiable, a maximal step size $\sigma_0 > 0$ and a starting point $x^0 \in \mathbb{R}^2$ be given, such that the level set $\mathcal{N}_f(x_0)$ is compact. Moreover, let d_1, d_2, d_3, and d_4 be defined by

$$d_1 \ := \ \begin{pmatrix} 0 \\ 1 \end{pmatrix} , \quad d_2 \ := \ \begin{pmatrix} 1 \\ 0 \end{pmatrix} , \quad d_3 \ := \ \begin{pmatrix} 0 \\ -1 \end{pmatrix} , \quad d_4 \ := \ \begin{pmatrix} 0 \\ -1 \end{pmatrix} .$$

The goal of this exercise is to analyze the convergence behavior of the **Compass-Search Method**, which in pseudocode looks like this:

```
$x^0 \in \mathbb{R}^n$ and $\sigma_0 > 0$ are given.
for $k = 0, 1, 2, \ldots$ do
    Choose $j \in \{1, 2, 3, 4\}$ with $f(x^k + \sigma_k s^j) \leq f(x^k + \sigma_k s^i)$ for all $i \in \{1, 2, 3, 4\}$.
    if $f(x^k + \sigma_k s^j) < f(x^k)$ then
        Set $x^{k+1} := x^k + \sigma_k s^j$ and $\sigma_{k+1} := \min(\sigma_0, 2\sigma_k)$ (successful iterate).
    else
        Set $x^{k+1} := x^k$ and $\sigma_{k+1} := \sigma_k / 2$ (unsuccessful iterate).
    end if
end for
```

We first consider the case that the algorithm stagnates at some point:

a) Show: if there is a $K \in \mathbb{N}$, such that $x^K = x^k$ for all $k \geq K$, then the last computed iterate x^K is a stationary point of f.

Thus, if the algorithm stops at any point, we have found a stationary point of the function f. Now we consider the case that an infinite amount of iterations are computed.

b) Show: If there is no $K \in \mathbb{N}$ as in a), then the sequence of iterates $\{x^k\}$ is contained in $\mathcal{N}_f(x_0)$ and there is a sub-sequence $\{k_l\}$ with $\alpha_{k_l} \to 0$ for $l \to \infty$.
Hint: Argue with a proof by contradiction and use the update rule for σ_k.

c) Show using b): If there is no $K \in \mathbb{N}$ as in a), then there is a sub-sequence $\{k_l\}$ of unsuccessful iterations, such that along this sub-sequence, it holds

$$\sigma_{k_l} \to 0 \qquad \text{and} \qquad x_{k_l} \to x^*$$

for some $x^* \in \mathbb{R}^2$.

d) Show: If $\{k_l\}$ is a sub-sequence with the properties of c), then x^* is a stationary point.

This shows, that the Compass-Search method, at least along a sub-sequence converges to a stationary point of the objective function f.

Homework Assignment:

Problem 4.1: Accumulation Points of Descent Methods — Let $f : \mathbb{R}^n \to \mathbb{R}$ be a continuous function and $(x^k) \subset \mathbb{R}^n$ a sequence with $f(x^{k+1}) < f(x^k)$ for all $k \in \mathbb{N}_0$ (i.e. the sequence of iterates of the gradient descent method).

a) If $\overline{x}$ and $\tilde{x}$ are two accumulation points of (x^k), then it holds $f(\overline{x}) = f(\tilde{x})$.

b) If $\overline{x}$ is an accumulation point of (x^k), then $\overline{x}$ is no local maximum.

Problem 4.2: Comparing the Efficiency of Step Size Rules — Let us revisit problem 3.3 and again consider the quadratic minimization problem

$$\min_{x \in \mathbb{R}^5} x^T A x \,,$$

where $A = (a_{i,j})$ is the 5×5 Hilbert matrix defined by

$$a_{i,j} \;=\; \frac{1}{i+j-1}\,, \qquad \text{for } i,j = 1,2,3,4,5\,.$$

The matrix can be constructed via the SciLab command `A = testmatrix('hilb',5)`. Run the following methods and compare the number of iterations required by each of the methods when the initial vector is $x^0 = (1,2,3,4,5)^T$ to obtain a solution x with accuracy $\varepsilon = \|\nabla f(x)\| \leq 10^{-4}$:

a) Diagonally scaled Gradient Descent Method with diagonal elements $d_{i,i} = a_{i,i}^{-1}$, $i = 1,2,3,4,5$, and exact line search.

b) Diagonally scaled Gradient Descent Method with diagonal elements $d_{i,i} = a_{i,i}^{-1}$, $i = 1,2,3,4,5$, and Armijo step size selection with $\gamma = 0.1$, $\beta = 0.5$ and $S_0 = 1$.

Problem 4.3: Gradient Descent Method with the Armijo Step Size Rule — Use your implementations of the Gradient Descent Method with Armijo step size rule ($S_0 = 1$) to deal with the unrestricted minimization problems $\min_{x \in \mathbb{R}^2} f(x)$ for the following functions and the given additional inputs.

a) The *Himmelblau-function*[1]

$$f : \mathbb{R}^2 \to \mathbb{R}\,, \qquad f(x_1, x_2) \;=\; (x_1^2 + x_2 - 11)^2 + (x_1 + x_2^2 - 7)^2,$$

[1] Himmelblau's function is a multi-modal function, used to test the performance of optimization algorithms. The function is named after David Mautner Himmelblau (1924–2011), who introduced it in 1972.

with different initial points $x^0 = (-0.26, 0)^T$, $x^0 = (-0.27, 0)^T$, and $x^0 = (-0.28, 0)^T$ and the parameters $\texttt{maxit} = 1000$ (maximal number of iterations, otherwise terminate the algorithm), $\beta = 0.5$, $\gamma = 10^{-4}$, and $\varepsilon = 10^{-5}$. What do you notice?

Draw a contour plot of the Himmelblau-function and plot the sequence of iterates (x^k) computed by your method. What do you notice?

b) The *Rosenbrock-function*[2]

$$f : \mathbb{R}^2 \to \mathbb{R}\,, \qquad f(x_1, x_2) \;=\; 100(x_2 - x_1^2)^2 + (1 - x_1)^2,$$

with initial point $x^0 = (-1.2, 1)^T$ and the parameters $\texttt{maxit} = 100\,000$ (maximal number of iterations, otherwise terminate the algorithm), $\beta = 0.5$ and $\gamma = 10^{-4}$. Evaluate the method for $\varepsilon = 10^{-1}$, 10^{-2}, 10^{-3}, 10^{-4}, 10^{-5}. What do you notice?

Draw a contour plot of the Rosenbrock-function and plot the sequence of iterates (x^k) computed by your method. Can you explain, why the gradient descent method is so inefficient in this case.

Problem 4.4: Implementation of the Compass-Search Algorithm — Implement the Compass-Search algorithm of Exercise 4.2, and test it for the following benchmark functions. What are the strengths and weaknesses of this method? How could a termination criterion of the algorithm look like?

Benchmark functions:

- $f : \mathbb{R}^2_+ \to \mathbb{R}$ with

$$f(x_1, x_2) \;=\; 0.1 \cdot x_1 - 0.5 \cdot \ln(x_1) + 0.1 \cdot x_2 - 0.8 \cdot \ln(x_2)\,.$$

- Himmelblau-function (and compare with the corresponding results from Problem 4.3).
- Rosenbrock-function (and compare with the corresponding results from Problem 4.3).

Graded Homework Assignment:

Graded Problem IV.1: Diagonal Scaling and Improving the Condition Number — Consider the quadratic minimization problem

$$\min_{x \in \mathbb{R}^2} x^T Q x\,,$$

where $Q = (q_{i,j})$ is a positive definite 2×2-matrix. Suppose we use the diagonal scaling matrix

$$D \;=\; \begin{pmatrix} q_{1,1}^{-1} & 0 \\ 0 & q_{2,2}^{-1} \end{pmatrix}.$$

Show that the above scaling matrix improves the condition number of Q in the sense that

$$\kappa(D^{1/2} Q D^{1/2}) \;\leq\; \kappa(Q)\,.$$

Graded Problem IV.2: Transformation of Variables in the Gradient Descent Method — Let $f : \mathbb{R}^n \to \mathbb{R}$ be a $\mathcal{C}^2$-function and $T : \mathbb{R}^n \to \mathbb{R}^n$, $T(y) := By + b$ an affine-linear transformation of variables with an invertible matrix $B \in \mathbb{R}^{n \times n}$ and a vector $b \in \mathbb{R}^n$. Denote by $\hat{f}(y) := f(T(y))$ the function f transformed to y coordinates. Further let $x \in \mathbb{R}^n$ a point with $\nabla f(x) \neq 0$, and x_g denote the result of a gradient step for the solution of the problem $\min_{z \in \mathbb{R}^n} f(z)$ starting at the point $x \in \mathbb{R}^n$, i.e.

$$x_g \;:=\; x - \alpha \nabla f(x)\,,$$

for some step size $\alpha > 0$. Analogously let y_g the result of a gradient step for $\hat{f}$ starting at $y = T^{-1}(x)$, i.e.

$$y_g \;:=\; y - \alpha \nabla \hat{f}(y)\,.$$

[2] The Rosenbrock function is a non-convex function, introduced by Howard H. Rosenbrock (1920–2010) in 1960, which is used as a performance test problem for optimization algorithms.

a) Express $\nabla \hat{f}(y)$ and $H_{\hat{f}}(y)$ in terms of $\nabla f(x)$, $H_f(x)$, B and b. Write down your derivation.

b) Show that a step $y_g = y - \alpha \nabla \hat{f}(y)$ in the transformed space can be seen as a step in the original space with a search direction $s = T(y_g) - x$ which solves a linear system $Ms = -\nabla f(x)$. Here, M is a symmetric positive definite matrix.

Hint: Methods of this type are called Quasi-Newton methods and will be discussed later in the lecture.

c) For which class of matrices B is the Gradient Descent Method invariant under the transformation T, i.e. when does $T(y_g) = x_g$ hold (for the same step size α)?

d) Now consider specifically quadratic functions of the type

$$f(x) \;=\; c^T x + \tfrac{1}{2} x^T C x\,, \tag{1}$$

with $C \in \mathbb{R}^{n \times n}$ symmetric positive definite and $c \in \mathbb{R}^n$. In the lecture it was shown (or will be shown) that for the gradient descent method with the minimization rule for the step size, in order to minimize f of the form (1) we can expect the following rate of convergence:

$$\|x^k - \overline{x}\| \;\leq\; \sqrt{\frac{\lambda_{\max}(C)}{\lambda_{\min}(C)}} \left(\frac{\lambda_{\max}(C) - \lambda_{\min}(C)}{\lambda_{\max}(C) + \lambda_{\min}(C)} \right)^k \|x^0 - \overline{x}\|\,.$$

Argue, why it is useful to choose B in the transformation T in such a way, that the condition number of $B^T C B$ becomes as small as possible, such that the method converges quickly in the y-space. Why is a choice of B with $BB^T \approx C^{-1}$ especially useful?

Hint: the conditioning number $\kappa(A)$ of an symmetric positive definite (s.p.d.) matrix A satisfies: $\kappa(A) = \frac{\lambda_{\max}(A)}{\lambda_{\min}(A)}$.