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

The Wolfe-Powell Step Size Rule

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 6.1: Relation Between Armijo- & Wolfe-Powell Step Sizes — Let us consider the minimization problem $\min_{x\in\mathbb{R}^2} f(x)$ with $f : \mathbb{R}^2 \to \mathbb{R}$ given by

$$f(x) \;=\; \tfrac{1}{8}x_1^4 - \tfrac{3}{4}x_1^2 + x_1x_2 + x_2^2 + 2x_2\,.$$

In order to solve it, we use the general descent method with starting point $x^0 = (0,-1)^T$ and descent direction $s^k = -\nabla f(x^k)$.

a) Show, that the Armijo step size rule in the first iteration generates for any $\gamma \in (0,1)$ the step size $\sigma_0^{\mathrm{A}} = 1$.

b) Let now further $\gamma \in (0,\frac{1}{2})$ and $\eta \in (0,1)$ with $\gamma < \eta$ be given. Show, that the step size $\sigma_0^{\mathrm{A}} = 1$ does not satisfy the Wolfe-Powell conditions. Is there any step size $\sigma_0^{\mathrm{PW}} > 0$, which satisfies the Wolfe-Powell conditions? Explain your answer.

Exercise 6.2: Connecting the Wolfe-Powell Step Size Rule with the Angle Condition — Consider any iteration of a line search with $x^{k+1} = x^k + \sigma_k s^k$, where s^k is a descent direction and σ_k satisfies the Wolfe-Powell conditions. Suppose that f is bounded below in $\mathbb{R}^n$ and that f is continuously differentiable in an open set $U \subset \mathbb{R}^n$ containing the level set $\mathcal{N}(x^0) = \{x \in \mathbb{R}^n : f(x) \le f(x^0)\}$, where x^0 is the starting point of the iteration. Assume also that the gradient ∇f is Lipschitz continuous on U with a Lipschitz-constant $L > 0$, such that

$$\|\nabla f(x) - \nabla f(\tilde{x})\| \;\le\; L\cdot\|x-\tilde{x}\| \qquad \text{for all } x,\tilde{x}\in U\,.$$

Show that it then holds that

$$\sum_{k\ge 0} \cos^2\theta_k \|\nabla f_k\|^2 \;<\; \infty\,,$$

where θ_k is the angle between s^k and the steepest descent direction $-\nabla f(x^k)$:

$$\cos(\theta_k) \;=\; \frac{-\nabla f(x^k)^T s^k}{\|\nabla f(x^k)\|\cdot\|s^k\|}\,.$$

Remark: This is result is known as Zoutendijk's Theorem and implies $\cos^2(\theta_k)\|\nabla f(x^k)\| \to 0$. This limit can be used in turn to derive global convergence results for line search algorithms.

Homework Assignment:

Problem 6.1: Stopping Criterion — A question that arises in using an algorithm such as the Gradient Descent Method to minimize an objective function f is when to stop the iterative process, or, in other words, how can one tell when the current point x^k is close to a solution $\overline{x}$. If, as with steepest descent, it is known that convergence is linear, this knowledge can be used to develop a stopping criterion. Let $\{f(x^k)\}_{k=0}^{\infty}$ be the sequence of values obtained by the algorithm. We assume that $f(x^k) \to f(\overline{x})$ converges linearly with a rate $0 < \gamma < 1$,[1] but both $f(\overline{x})$ and the convergence ratio γ are unknown. However we know that, at least approximately,

$$f(x^{k+1}) - f(\overline{x}) = \gamma\left(f(x^k) - f(\overline{x})\right) \qquad \text{and} \qquad f(x^k) - f(\overline{x}) = \gamma\left(f(x^{k-1}) - f(\overline{x})\right).$$

These two equations can be solved for $f(\overline{x})$ and γ.

a) Show that

$$f(\overline{x}) = \frac{(f(x^k))^2 - f(x^{k-1})f(x^{k+1})}{2f(x^k) - f(x^{k-1}) - f(x^{k+1})} \qquad \text{and} \qquad \gamma = \frac{f(x^{k+1}) - f(x^k)}{f(x^k) - f(x^{k-1})}.$$

b) Motivated by the above we form the sequence $\{f_k^*\}_{k=0}^{\infty}$ defined by

$$f_k^* := \frac{(f(x^k))^2 - f(x^{k-1})f(x^{k+1})}{2f(x^k) - f(x^{k-1}) - f(x^{k+1})}$$

as the original sequence is generated. (This procedure of generating $\{f_k^*\}_{k=0}^{\infty}$ from $\{f(x^k)\}_{k=0}^{\infty}$ is called the Aitken δ^2-process.) If $\|f(x^k) - f(\overline{x})\| = \gamma^k + o(\gamma^k)$ show that $\|f_k^* - f(\overline{x})\| = o(\gamma^k)$ which means that $\{f_k^*\}_{k=0}^{\infty}$ converges to $f(\overline{x})$ faster than $\{f(x^k)\}_{k=0}^{\infty}$ does. The iterative search for the minimum of f can then be terminated when $f(x^k) - f_k^*$ is smaller than some prescribed tolerance.

Problem 6.2: Implementing Wolfe-Powell and Analyzing Step Sizes — The goal of this exercise is to implement the Wolfe-Powell step size rule and to compare it to other line search methods. The following algorithm to determine a Wolfe-Powell step size has been presented in the lecture and proven to return a step size satisfying the Wolfe-Powell conditions:

$$(*) \quad f(x + \sigma s) - f(x) \le \sigma\gamma\nabla f(x)^T s \qquad \text{and} \qquad (**) \quad \nabla f(x + \sigma s)^T s \ge \eta\nabla f(x)^T s\,.$$

```
1: if $\sigma = 1$ satisfies the Armijo-Condition $(*)$, go to step 3
2: Determine the largest $\sigma_- \in \{2^{-1}, 2^{-2}, ...\}$, such that $\sigma_-$ satisfies $(*)$.
   Set $\sigma_+ = 2\sigma_-$ and go to step 4.
3: Compute the smallest $\sigma_+ \in \{2, 2^2, 2^3, ...\}$ such that $(*)$ is violated. Set $\sigma_- = \sigma_+/2$.
4: while $\sigma = \sigma_-$ violates $(**)$ do
5:     Set $\sigma = \frac{\sigma_- + \sigma_+}{2}$
6:     If $\sigma$ satisfies $(*)$, set $\sigma_- = \sigma$, else set $\sigma_+ = \sigma$
7: end while
8: return $\sigma = \sigma_-$
```

a) Implement the Wolfe-Powell step size rule. During the algorithm, count all evaluations of the objective function and its gradient.

b) We consider the Himmelblau function (see problem 4.6:) and a general descent method with search directions $s^k := \delta\nabla f(x^k)$.

Use the Armijo-Backtracking (with $\beta = 0.5$, $\gamma = 10^{-4}$, maximal number of searches for the best exponent of β: `maxit` $= 20$) and the Wolfe-Powell line search (with $\gamma = 10^{-4}$ and $\eta =$

[1] A sequence $\{y^k\}_{k\in\mathbb{N}_0} \subset \mathbb{R}^n$ is said to **converge linearly** with a rate $0 < \gamma < 1$ to $\overline{y} \in \mathbb{R}^n$, if there is a $\ell \ge 0$ such that $\|y^{k+1} - \overline{y}\| \le \gamma\|y^k - \overline{y}\|$ for all $k \ge \ell$ ($\ell \in \mathbb{N}_0$). This implies $y^k \to \overline{y}$.

0.25) for the following values: $\delta = 1, 0.1, 0.01, 0.001$. Use the initial point $x_0 = (-0.26, 0)^T$, tolerance $\texttt{tol} = 10^{-5}$ and maximal number of iterations 100.

For which values of δ is the Armijo-rule more cost efficient, for which ones the Wolfe-Powell rule? What are the reasons for this? Do you note similarities to problem 5.2?

Note: With *cost* we mean the number of function evaluations and gradient evaluations.

c) Apply the gradient method with Armijo line search and Wolfe-Powell line search to the function $f(x) = a \cdot 0.5\|x\|^2$ for $a = 10, 1, 0.1, 0.01$. Use the starting point $x_0 = (-1, 0)^T$ and otherwise the same parameters as in b).

Analyze as in b) in which cases the Armijo and in which Wolfe-Powell is more cost efficient. Do you notice parallels to problem 5.2?

d) Apply the gradient descent method with a constant step size of $\sigma = \sigma_k = \text{const.} = 0.005, 0.01, 0.019, 0.03$ to the Himmelblau function. Use the tolerance $\texttt{tol} = 10^{-5}$, the maximal number of iterations 100 and starting point $x^0 = (-0.26, 0)^T$. What do you observe?

Graded Homework Assignment:

Graded Problem VI.1: Minimization and the Armijo Step Size Rule — Let $f(x) = \frac{1}{2}x^T Cx + c^T x$ with a symmetric and positive matrix $C \in \mathbb{R}^{n\times n}$ and $c \in \mathbb{R}^n$. Moreover, let $s \in \mathbb{R}^n$ be a descent direction of f at a point $x \in \mathbb{R}^n$ and $\sigma^* \geq 0$ be the exact line search step size, i.e. $f(x + \sigma^*) = \min_{\sigma \geq 0} f(x + \sigma s)$.

a) Show that f is strictly convex.

b) Show that $\sigma^* > 0$ holds.

c) What form (linear, quadratic, ...) does the function $\phi(\sigma) = f(x + \sigma s)$ have? Use a Taylor expansion of ϕ about $\sigma = 0$ and deduct that σ^* is well defined and uniquely determined.

d) Show that for all $\gamma \in (0, \frac{1}{2}]$ the choice $\sigma = \sigma^*$ satisfies the sufficient decrease condition

$$f(x + \sigma s) - f(x) \;\leq\; \gamma\sigma\nabla f(x)^T s\,,$$

though, that this is not the case for $\gamma > \frac{1}{2}$.

e) Sketch the graph of ϕ and use it illustrate the statement discussed in d).

Graded Problem VI.2: Applications of the Wolfe-Powell Rule

a) The first Wolfe-Powell condition (sufficient decrease condition) requires

$$f(x^k + \sigma_k s^k) - f(x^k) \;\leq\; \gamma\sigma_k\nabla f(x^k)^T s^k\,.$$

What is the maximum step length σ_k that satisfies this condition, given that $f(x) = 5 + x_1^2 + x_2^2$, $x^k = (-1, -1)^T$, $s^k = (1, 0)^T$, and $\gamma = 10^{-4}$?

b) Given a general descent algorithm with the Wolfe-Powell step size rule. Provide an example to show that the set

$$\left\{ t \in \mathbb{R} : \begin{array}{c} \nabla f(x + \sigma s)^T s \geq \eta\nabla f(x)^T s \quad \text{and} \\ f(x + \sigma s) - f(x) \;\leq\; \gamma\sigma\nabla f(x)^T s \end{array} \right\}$$

may be empty if $0 < \gamma < \eta < 1$. I.e., that under these conditions no step size can be found.

Hint: Think of a one-dimensional example.

Graded Problem VI.3: Gradient Descent Along Eigenvectors — Let $f(x) = \frac{1}{2}x^T Cx + c^T x + \gamma$ with a symmetric and positive definite matrix $C \in \mathbb{R}^{n\times n}$, $c \in \mathbb{R}^n$, and $\gamma \in \mathbb{R}$. Show that Gradient Descent Method with exact line search step size reaches the global minimum $\overline{x} = C^{-1}c$ in exactly one step if the initial point x^0 is chosen such that $\nabla f(x^0)$ is an eigenvector of C. What does this imply for a strategy to choose the initial point in a diagonally scaled Gradient Descent Method?