Control and Cybernetics

vol. **40** (2011) No. 4

Nonmonotone line searches for optimization algorithms*

by

Ekkehard W. Sachs[1] and Stephen M. Sachs[2]

[1] FB IV - Mathematik, Universität Trier, Germany
[2] Graduate School of Computational Engineering,
Technische Universität Darmstadt, Germany

Abstract: In this paper we develop a general convergence theory for nonmonotone line searches in optimization algorithms. The advantage of this theory is that it is applicable to various step size rules that have been published in the past decades. This gives more insight into the structure of these step size rules and points to several relaxations of the hypotheses. Furthermore, it can be used in the framework of discretized infinite-dimensional optimization problems like optimal control problems and ties the discretized problems to the original problem formulation.

Keywords: line search, nonmonotone, Armijo's rule.

1. Descent methods

In this paper, we consider the generalization of step size rules for unconstrained optimization problems in Hilbert space.

Let $f : X \to I\!R$ with Hilbert space X and inner product $\langle \cdot, \cdot \rangle$ be given and let f be Fréchet-differentiable where the derivative $f'(x)v$ of f at $x \in X$ is denoted by the gradient $\nabla f(x) \in X$ as

$$f'(x)v = \langle \nabla f(x), v \rangle$$

for all $v \in X$. The minimization problem is formulated as

$$\min_{x \in X} f(x) \qquad \text{or find } x_* \text{ such that} \quad f(x_*) \leq f(x) \quad \text{for all} \quad x \in X.$$

Algorithms for finding x_* are often based on the following concept.

Given $x_k \in X$ we call $d_k \in X$ a **descent direction**, if $\langle \nabla f(x_k), d_k \rangle < 0$. In a descent method the next iterate is determined by

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

*Submitted: February 2011; Accepted: September 2011

with an appropriate **step size** $\alpha_k \in (0, \infty)$. It is obvious that a sufficiently small step size α_k combined with the descent direction yields a decrease in function value, i.e. $f(x_{k+1}) < f(x_k)$. This yields a monotone decreasing sequence $f(x_k)$ and is often a building block in a subsequent convergence proof.

However, it has been noted that such a monotonicity property can be a burden from a numerical point of view, i.e. it could lead to fairly small step sizes even when far away from the minimum. Therefore, several papers have appeared in the past, where this requirement has been weakened to allow nonmonotone behavior in the function values.

The determination of an appropriate step size rule has been the topic of numerous research articles and books starting with the early days of numerical optimization. Here, we concentrate on step size rules which do not guarantee that the function value of the new point is smaller than the function value of the original point, so called nonmonotone step size rules. Fixed step sizes given by sequences which satisfy certain convergence properties were already considered fifty years ago, likewise those step size rules that are fractions of the norm of the gradients. For a general theory on step size rules see e.g. Warth and Werner (1997). In this context we also refer to the paper by Hüther (2002).

L. Grippo, F. Lampariello and S. Lucidi (1986, 1989) published a nonmonotone version of the popular Armijo's rule. Shi and Shen (2006) extended this approach to a descent term including a quadratic term. Zhang and Hager (2004) gave a different version of a nonmonotone Armijo's rule. Another well known descent method with a nonmonotone step size rule is the Barzilai-Borwein method (1988). The approaches in Barzilai and Borwein (1988), Grippo, Lampariello and Lucidi (1986, 1989) are not covered in this paper and are the topic of future research efforts to include these also in a general theory of nonmonotone line searches.

The goal of this paper is to develop a general framework for a convergence theory in connection with step size rules which do not require a monotone descent property and allow also for perturbed descent directions. The advantage of such a theory is its versatility in applying it to various scenarios of step size rules. The general setting and the convergence statement for the general case can be found in Section 2.

As a special case, it includes the theory of a-priori fixed sequences of step size rules which have been quite popular with a number of authors. These step size rules lead automatically to nonmonotone behavior of the function values. In Section 3 we show how the general theorem can be applied to this special case of step size rules.

Armijo's step size rule is one of the most popular and most efficient step size rules, since its behavior is - in contrast to the rules in Section 3 - problem dependent. In Section 4 we formulate an extension of the classical Armijo rule which allows for a nonmonotone behavior in the resulting function values $f(x_k)$. Again, the general convergence theory in Section 2 plays an instrumental role in the proof of convergence of the gradients.

Another version of step size rules is based on the step size being a fixed portion of the norm of the current gradient. Also this rule cannot lead to a necessarily monotone descent property for the function values. In this case, we can apply the general theory and interpret this rule in Section 5 as a special case.

Zhang and Hager (2004) considered a nonmonotone version of Armijo's rule which can be interpreted as a rule taking the average of previous function value iterates into account as a reference instead of the function value of the last iterate. It is interesting to see that also here the general convergence theory, when specialized appropriately, leads in Section 6 to a convergence statement similar to the one stated in Zhang and Hager (2004).

In the final Section 7 we give some comments on how the perturbation term in Armijo's rule introduced in Section 4 can be interpreted as a discretization error for a discretized problem. If the discretization is refined in the course of the iterations, this can lead to a convergence statement for a sequence of iterates in the infinite dimensional setting.

2. General setting and convergence theorem

Before we start with nonmonotone line searches, let us quote an elementary Lemma, which will be used in the sequel.

LEMMA 1 *Let sequences $\{a_k\}$ and $\{\varepsilon_k\}$ be given with*

$$a_k \geq a_{min} \qquad \varepsilon_k \geq 0, \qquad \sum_{k=1}^{\infty} \varepsilon_k < \infty.$$

Assume that

$$a_{k+1} \leq a_k + \varepsilon_k,$$

then there exists $a_* \geq a_{min}$ with

$$\lim_{k\to\infty} a_k = a_*.$$

A proof can be found, e.g., in Kaplan, Tichatschke (1994, Lemma 4.1).

Let us start with a general theorem on the convergence of function values. In the assumptions the usual conditions on the descent properties of the search directions are relaxed. The theorem deals not only with the convergence of the function values, but yields also a convergence statement, the Zoutendijk condition, about a forcing function

$$\sigma : X \to I\!R$$

which will be specified more precisely below. The most common example of a forcing function is $\sigma(\cdot) = \|\nabla f(\cdot)\|$ such that the convergence of $\sigma(x_k)$ to zero forces the gradients of f also to converge to zero.

Note that the condition (1) which relates the descent direction and the forcing function $\sigma(\cdot)$ contains a perturbation term λ. Even more important is the fact that equation (2) does not require a descent property, even if the term $\langle \nabla f(x_k), d_k \rangle$ is negative. Even in this case, the perturbation term ν_k allows for nonmonotone behavior in the function values.

THEOREM 1 *Let $f : X \to I\!R$ be Fréchet-differentiable and let f be bounded from below on X. Furthermore, let $\sigma : X \to I\!R_+$ be given.*
Assume that the sequences $x_k, d_k \in X$ and $\alpha_k, \lambda_k, \nu_k \in I\!R$ satisfy $\alpha_k, \lambda_k, \nu_k \geq 0$ and

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

with

$$\lambda_k - \langle \nabla f(x_k), d_k \rangle \geq \sigma(x_k) \geq 0 \tag{1}$$

and for some $\varrho > 0$

$$f(x_{k+1}) - f(x_k) \leq \varrho \, \alpha_k \, \langle \nabla f(x_k), d_k \rangle + \nu_k. \tag{2}$$

If

$$\sum_{k=1}^{\infty} \lambda_k \alpha_k < \infty \quad \textit{and} \quad \sum_{i=1}^{\infty} \nu_k < \infty, \tag{3}$$

then there exists $f_* \in I\!R$ with

$$\lim_{k \to \infty} f(x_k) = f_*$$

and we have for the forcing terms

$$\sum_{k=1}^{\infty} \alpha_k \, \sigma(x_k) < \infty. \tag{4}$$

Proof. Let f_m denote the lower bound of f on X.
From (1) and (2) we obtain

$$\begin{aligned} (f(x_{k+1}) - f_m) \quad &- \quad (f(x_k) - f_m) = f(x_{k+1}) - f(x_k) \\ &\leq \quad \varrho \, \alpha_k \langle \nabla f(x_k), d_k \rangle + \nu_k \leq \varrho \, \alpha_k \, \lambda_k + \nu_k. \end{aligned}$$

We choose $a_k = f(x_k) - f_m, \;\; \varepsilon_k = \varrho \, \alpha_k \, \lambda_k + \nu_k$ and use assumption (3) in

$$\sum_{k=1}^{\infty} \varepsilon_k = \varrho \, \sum_{k=1}^{\infty} \alpha_k \, \lambda_k + \sum_{k=1}^{\infty} \nu_k < \infty$$

so that Lemma 1 yields the convergence of the function values $f(x_k) \to f_*$. The inequalities (1) and (2) imply

$$\alpha_k \sigma(x_k) \leq \alpha_k \lambda_k - \alpha_k \langle \nabla f(x_k), d_k \rangle \leq \alpha_k \lambda_k + (f(x_k) - f(x_{k+1}) + \nu_k)/\varrho.$$

Summation over k

$$\sum_{k=1}^{j} \alpha_k \sigma(x_k) \leq \sum_{k=1}^{j} \alpha_k \lambda_k + \frac{1}{\varrho} \sum_{k=1}^{j} \nu_k + \frac{1}{\varrho} (f(x_1) - f(x_{j+1}))$$

and passing to the limit gives the estimate (3)

$$\sum_{k=1}^{\infty} \alpha_k \sigma(x_k) \leq \sum_{k=1}^{\infty} \alpha_k \lambda_k + \frac{1}{\varrho} \sum_{k=1}^{\infty} \nu_k + \frac{1}{\varrho} (f(x_1) - f_*) < \infty. \qquad \blacksquare$$

3. Predetermined sequences of step sizes

One of the oldest step size concepts are those, where the step sizes are predetermined sequences which have to satisfy certain summability requirements. Those schemes are independent of the underlying function f to be minimized and therefore do not necessarily yield a descent in each iteration. Our general framework is set up in such a way that it covers this scenario also as a special case.

In comparison with classical convergence statements in this case, we allow the descent direction d_k to deviate from the steepest descent $-\nabla f(x_k)$ by a vector quantity r_k. We give conditions in (7) which still guarantee certain convergence results as shown below.

Theorem 2 *Let $f : X \to I\!R$ be Fréchet-differentiable and let f be bounded from below on X. Furthermore, let $\nabla f(x)$ be bounded and Lipschitz-continuous on X with Lipschitz-constant L.*
Assume that the sequences $x_k, d_k, r_k \in X, \alpha_k \in I\!R$ satisfy $\alpha_k \geq 0$ and

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

with

$$d_k = -\nabla f(x_k) + r_k \tag{5}$$

with

$$\sum_{k=1}^{\infty} \alpha_k^2 < \infty \tag{6}$$

and

$$\sum_{k=1}^{\infty} \alpha_k \|r_k\| < \infty, \qquad \|r_k\| \leq r_*. \tag{7}$$

Then there exists $f_* \in I\!R$ with

$$\lim_{k\to\infty} f(x_k) = f_* \quad \textit{and} \quad \sum_{k=1}^{\infty} \alpha_k \, \|\nabla f(x_k)\|^2 < \infty. \tag{8}$$

Proof. Note that with

$$\lambda_k = |\langle \nabla f(x_k), r_k\rangle| \quad \text{and} \quad \sigma(x) = \|\nabla f(x)\|^2, \tag{9}$$

together with (5), we obtain

$$\begin{aligned} \lambda_k - \langle \nabla f(x_k), d_k\rangle &= |\, \langle \nabla f(x_k), r_k\rangle \,| - \langle \nabla f(x_k), -\nabla f(x_k) + r_k\rangle \\ &\geq \|\nabla f(x_k)\|^2 = \sigma(x_k) \geq 0. \end{aligned} \tag{10}$$

To determine the perturbation ν_k estimate

$$\begin{aligned} f(x_{k+1}) \;-\; & f(x_k) = f(x_k + \alpha_k d_k) - f(x_k) = \int_0^1 \langle \nabla f(x_k + \tau\alpha_k d_k), \alpha_k d_k\rangle \, d\tau \\ =\;& \alpha_k \langle \nabla f(x_k), d_k\rangle + \int_0^1 \langle \nabla f(x_k + \tau\alpha_k d_k) - \nabla f(x_k), \alpha_k d_k\rangle d\tau \\ \leq\;& \alpha_k \langle \nabla f(x_k), d_k\rangle + \frac{L}{2}\alpha_k^2 \|d_k\|^2. \end{aligned} \tag{11}$$

By assumption, we have numbers g_*, r_* such that

$$\|d_k\| \leq \|\nabla f(x_k)\| + \|r_k\| \leq g_* + r_*. \tag{12}$$

Hence, (2) is satisfied with

$$\varrho = 1, \qquad \nu_k = L\alpha_k^2 \|d_k\|^2/2$$

and, moreover, by (6)

$$\sum_{k=1}^{\infty} \nu_k \leq \sum_{k=1}^{\infty} L\alpha_k^2 \|d_k\|^2/2 \leq L(g_* + r_*)^2/2 \sum_{k=1}^{\infty} \alpha_k^2 < \infty.$$

Likewise, it is satisfied by (7)

$$\sum_{k=1}^{\infty} \lambda_k \alpha_k = \sum_{k=1}^{\infty} \alpha_k |\langle \nabla f(x_k), r_k\rangle| \leq g_* \sum_{k=1}^{\infty} \alpha_k \|r_k\| < \infty. \tag{13}$$

Since (3) is satisfied, we deduce from Theorem 1

$$\lim_{k\to\infty} f(x_k) = f_*, \qquad \sum_{k=1}^{\infty} \alpha_k \|\nabla f(x_k)\|^2 < \infty. \qquad \blacksquare$$

If we also assume a lower estimate on the step sizes, then we obtain the convergence of the gradients to zero.

Corollary 1 *In addition to the assumptions of Theorem 2 let*

$$\sum_{k=1}^{\infty} \alpha_k = \infty. \tag{14}$$

Then

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

Proof. First, we show

$$\liminf_{k\to\infty} \|\nabla f(x_k)\| = 0. \tag{15}$$

If (15) does not hold, then there exists $\varepsilon > 0$ such that $\|\nabla f(x_k)\| \geq \varepsilon$ for all $k \geq k_0$ and hence

$$\sum_{k=1}^{\infty} \alpha_k \|\nabla f(x_k)\|^2 \geq \sum_{k=k_0}^{\infty} \alpha_k \varepsilon^2 = \infty,$$

a contradiction to (8).

In order to show $\nabla f(x_k) \to 0$ we assume, on the contrary, that there exists a subsequence x_{k_j} such that

$$\|\nabla f(x_{k_j})\| \geq \varepsilon > 0.$$

From (15) it follows that for each j there exists a smallest $l_j > k_j$ with

$$\|\nabla f(x_{l_j})\| \leq \frac{\varepsilon}{2},$$

i.e.

$$\|\nabla f(x_k)\| > \frac{\varepsilon}{2} \quad \text{for} \quad k = k_j, \ldots, l_j - 1. \tag{16}$$

Using the Lipschitz continuity of $\nabla f(\cdot)$ and (12) we get

$$\begin{aligned} \frac{\varepsilon}{2} &= \varepsilon - \frac{\varepsilon}{2} \leq \|\nabla f(x_{k_j})\| - \|\nabla f(x_{l_j})\| \leq \|\nabla f(x_{k_j}) - \nabla f(x_{l_j})\| \\ &\leq L\|x_{k_j} - x_{l_j}\| = L\| \sum_{k=k_j}^{l_j-1} \alpha_k d_k \| \leq L \sum_{k=k_j}^{l_j-1} \alpha_k \|d_k\| \leq L(g_* + r_*) \sum_{k=k_j}^{l_j-1} \alpha_k \end{aligned}$$

and hence

$$\sum_{k=k_j}^{l_j-1} \alpha_k \geq \frac{\varepsilon}{2L(g_* + r_*)} > 0. \tag{17}$$

For the function values we obtain using (10), (11) and (12)

$$\begin{aligned} f(x_{l_j}) - f(x_{k_j}) &\leq \sum_{k=k_j}^{l_j-1} (f(x_{k+1}) - f(x_k)) \\ &\leq \sum_{k=k_j}^{l_j-1} (\alpha_k \langle \nabla f(x_k), d_k \rangle + \frac{L}{2} \alpha_k^2 \|d_k\|^2) \\ &\leq \sum_{k=k_j}^{l_j-1} \big(- \alpha_k \|\nabla f(x_k)\|^2 + \alpha_k \lambda_k + \frac{L}{2}(g_* + r_*)^2 \alpha_k^2 \big). \end{aligned}$$

We use (17) to estimate

$$\begin{aligned} f(x_{l_j}) - f(x_{k_j}) \quad &\leq \quad - \min_{k_j \leq k \leq l_j - 1} \|\nabla f(x_k)\|^2 \frac{\varepsilon}{2L(g_* + r_*)} \\ &\qquad + \sum_{k=k_j}^{\infty} \alpha_k \lambda_k + \frac{L}{2}(g_* + r_*)^2 \sum_{k=k_j}^{\infty} \alpha_k^2. \end{aligned}$$

Since the function values are converging, see (8), and the series (13) and (6) are convergent, we obtain for $j \longrightarrow \infty$

$$0 \leq - \lim_{j \to \infty} \min_{k_j \leq k \leq l_j - 1} \|\nabla f(x_k)\|^2 \ \frac{\varepsilon}{2L(g_* + r_*)}$$

and therefore

$$\lim_{j \to \infty} \min_{k_j \leq k \leq l_j - 1} \|\nabla f(x_k)\|^2 = 0,$$

a contradiction to (16). ■

4. Nonmonotone Armijo's rule

One of the most popular step size rules is Armijo's step size rule or the backtracking rule. Here, we give a version which allows also for an increase in the function value from one iteration to the next.

Nonmonotone Armijo step size rule
Let $\beta, \varrho \in (0,1)$ be fixed and $x_k \in X$, $d_k \in X$ with

$$\langle \nabla f(x_k), d_k \rangle < 0 \tag{18}$$

be given. Furthermore, let ν_k be a sequence of non-negative numbers and $\alpha_{max} > 0$.
If

$$f(x_k + \alpha_{max} d_k) - f(x_k) \leq \varrho \alpha_{max} \langle \nabla f(x_k), d_k \rangle + \nu_k \tag{19}$$

set

$$\alpha_k = \alpha_{max},$$

otherwise find the smallest $l_k \in I\!N$ with

$$f(x_k + \alpha_{max}\beta^{l_k} d_k) - f(x_k) \leq \varrho\alpha_{max}\beta^{l_k}\langle\nabla f(x_k), d_k\rangle + \nu_k \tag{20}$$

and set

$$\alpha_k = \alpha_{max}\beta^{k_l}.$$

If the perturbation quantity ν_k is set to zero, the usual Armijo step size rule is as listed above. Note that in the case $\nu_k > 0$, although d_k is a descent direction, the value of $f(x_k + \alpha_k d_k)$ does not need to be smaller than $f(x_k)$, leading to a nonmonotone line search.

Lemma 2 *The nonmonotone Armijo's rule is well defined, provided ∇f is Lipschitz-continuous.*

Proof. If (20) does not hold, then there exists a subsequence $l_{k_j} \to \infty$ such that

$$\begin{aligned} f(x_k + \alpha_{max}\beta^{l_{k_j}}\, d_k) - f(x_k) &> \varrho\alpha_{max}\beta^{l_{k_j}}\, \langle\nabla f(x_k), d_k\rangle + \nu_k \\ &\geq \varrho\alpha_{max}\beta^{l_{k_j}}\, \langle\nabla f(x_k), d_k\rangle. \end{aligned}$$

If L is the Lipschitz constant, Taylor's expansion yields

$$\begin{aligned} \langle\nabla f(x_k), d_k\rangle\alpha_{max}\beta^{l_{k_j}}\, + \frac{L}{2}(\alpha_{max}\beta^{l_{k_j}}\, \|d_k\|)^2 &\geq f(x_k + \alpha_{max}\beta^{l_{k_j}}\, d_k) - f(x_k) \\ &> \varrho\alpha_{max}\beta^{l_{k_j}}\, \langle\nabla f(x_k), d_k\rangle. \end{aligned}$$

A division by $\beta^{l_{k_j}}$ and taking the limit for $j \to \infty$ lead to the estimate

$$\alpha_{max}\langle\nabla f(x_k), d_k\rangle \geq \varrho\alpha_{max}\langle\nabla f(x_k), d_k\rangle.$$

From this and $\varrho \in (0,1)$ we conclude that $\langle\nabla f(x_k), d_k\rangle \geq 0$, contradicting the assumption (18). ■

In the following theorem we prove that we obtain the same convergence estimates as for the original Armijo rule, if we assume that the perturbation parameters ν_k form a summable series.

Theorem 3 *Let f be bounded from below by f_m and let ∇f be Lipschitz continuous. Furthermore we assume*

$$\sum_{k=1}^{\infty} \nu_k < \infty$$

and let $x_{k+1} = x_k + \alpha_k d_k$ satisfy the nonmonotone Armijo rule. Then there exists f_* such that

$$\lim_{k\to\infty} f(x_k) = f_*$$

and

$$\min\left\{ -\langle \nabla f(x_k), d_k\rangle, \left(\frac{\langle \nabla f(x_k), d_k\rangle}{\|d_k\|}\right)^2 \right\} \longrightarrow 0. \tag{21}$$

Proof. Since $\nabla f(\cdot)$ is Lipschitz continuous, the nonmonotone Armijo rule is well defined according to Lemma 2 and we have the following inequality for all k

$$f(x_{k+1}) - f(x_k) \le \varrho\alpha_k\langle \nabla f(x_k), d_k\rangle + \nu_k.$$

If we set $\lambda_k = 0$ and $\sigma(x_k) = -\langle \nabla f(x_k), d_k\rangle$ we see that the assumptions (1) and (3) in Theorem 1 are satisfied. The conclusions of Theorem 1 are the convergence of the function values $f(x_k)$ to some limit f_* and

$$\sum_{k=1}^{\infty} \alpha_k(-\langle \nabla f(x_k), d_k\rangle) < \infty. \tag{22}$$

If the case (19) is satisfied, we have

$$\alpha_k(-\langle \nabla f(x_k), d_k\rangle) = -\alpha_{max}\langle \nabla f(x_k), d_k\rangle. \tag{23}$$

Otherwise, due to the choice of l_k we have for $\frac{\alpha_k}{\beta} = \alpha_{max}\beta^{l_k-1}$ the reverse of (20)

$$f(x_k + \frac{\alpha_k}{\beta}d_k) - f(x_k) > \varrho\frac{\alpha_k}{\beta}\langle \nabla f(x_k), d_k\rangle + \nu_k \ge \varrho\frac{\alpha_k}{\beta}\langle \nabla f(x_k), d_k\rangle.$$

As usual, Taylor's expansion yields

$$\frac{\alpha_k}{\beta}\ \langle \nabla f(x_k), d_k\rangle\ +\ \frac{L}{2}\ (\frac{\alpha_k}{\beta}\ \|d_k\|)^2 > \varrho\ \frac{\alpha_k}{\beta}\ \langle \nabla f(x_k), d_k\rangle$$

and finally

$$\alpha_k\ (-\langle \nabla f(x_k), d_k\rangle) \ge (1-\varrho)\ \frac{2\beta}{L}\frac{(-\langle \nabla f(x_k), d_k\rangle)^2}{\|d_k\|^2}.$$

Together with (23) we obtain

$$\alpha_k(-\langle \nabla f(x_k), d_k\rangle) \ge \min\{-\alpha_{max}\langle \nabla f(x_k), d_k\rangle, (1-\varrho)\ \frac{2\beta}{L}\left(\frac{\langle \nabla f(x_k), d_k\rangle}{\|d_k\|}\right)^2\} \ge 0$$

where the left side tends to zero due to (22) and (21) is true. ■

The following three corollaries illustrate some of the choices for descent directions.

Corollary 2 *Let $d_k = -\nabla f(x_k)$. Then*

$$\nabla f(x_k) \longrightarrow 0.$$

Corollary 3 *Let d_k be such that $\|d_k\| = 1$ and for some $c > 0$*

$$\langle \nabla f(x_k), d_k \rangle \leq -\ c\ \|\nabla f(x_k)\| \tag{24}$$

be given. Then we have

$$\nabla f(x_k) \longrightarrow 0.$$

Proof. Note that

$$-\langle \nabla f(x_k), d_k \rangle \geq c\|\nabla f(x_k)\|$$

and

$$\left(\frac{\langle \nabla f(x_k), d_k \rangle}{\|d_k\|}\right)^2 \geq c^2\|\nabla f(x_k)\|^2$$

hold. From (21) follows the convergence of $\nabla f(x_k)$ to zero. ■

Corollary 4 *Let d_k be such that for some $a_1, a_2 > 0$*

$$\langle \nabla f(x_k), d_k \rangle \leq -a_1\|\nabla f(x_k)\|^2, \qquad \|d_k\| \leq a_2\|\nabla f(x_k)\|.$$

Then we have

$$\nabla f(x_k) \longrightarrow 0.$$

Proof. Note that

$$-\langle \nabla f(x_k), d_k \rangle \geq a_1\|\nabla f(x_k)\|^2$$

and

$$-\frac{\langle \nabla f(x_k), d_k \rangle}{\|d_k\|} \geq \frac{a_1}{a_2}\|\nabla f(x_k)\|,$$

which shows the convergence of $\nabla f(x_k)$ to zero using (21). ■

Conditions on the relation between the gradient $\nabla f(x_k)$ and the descent direction d_k in Corollaries 3 and 4 as well as in the subsequent Theorems 4 and 5 are closely related to conditions on the boundedness away from zero of the angle $\cos(-\nabla f(x_k), d_k)$.

5. A gradient norm based step size rule

In some applications, function evaluations are quite expensive. In such a case, an iterative procedure to determine a proper step size like Armijo's rule can be quite time consuming. A possibility to avoid such an iterative procedure is the choice of a step size as a fraction of the norm of the gradient. Convergence for such a rule can be proven, if the Lipschitz constant of the gradient of f is known. We show that this case is also covered by the general approach we outlined in the second section of this paper.

Theorem 4 *Let $f(x) \geq f_m$ and let ∇f be Lipschitz continuous with Lipschitz constant L. Choose $0 < \alpha_- < \alpha_+ < \frac{2}{L}c$ with some constant $c > 0$. Let d_k be descent directions satisfying*

$$\|d_k\| = 1 \quad \textit{and} \quad \langle \nabla f(x_k), d_k \rangle \leq -c\|\nabla f(x_k)\|.$$

Set

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

where

$$\alpha_k = \overline{\alpha}_k \|\nabla f(x_k)\|, \qquad \overline{\alpha}_k \in [\alpha_-, \alpha_+]. \tag{25}$$

Then

$$\lim_{k\to\infty} f(x_k) = f_* \quad \textit{and} \quad \lim_{k\to\infty} \|\nabla f(x_k)\| = 0.$$

Proof. Taylor's expansion and the assumptions on d_k yield

$$\begin{aligned} f(x_{k+1}) - f(x_k) &\leq \alpha_k \langle \nabla f(x_k), d_k \rangle + \frac{L}{2}\, \alpha_k^2\, \|d_k\|^2 \\ &= \alpha_k \left(\langle \nabla f(x_k), d_k \rangle + \frac{L}{2}\, \overline{\alpha}_k\, \|\nabla f(x_k)\| \right) \\ &\leq \alpha_k \langle \nabla f(x_k), d_k \rangle \left(1 - \frac{L}{2}\, \overline{\alpha}_k\, \frac{1}{c}\right) \\ &\leq \alpha_k \langle \nabla f(x_k), d_k \rangle \left(1 - \frac{L}{2}\, \alpha_+\, \frac{1}{c}\right). \end{aligned}$$

Hence (2) holds with $\nu_k = 0$ and $\varrho = 1 - \frac{L}{2c}\alpha_+ \in (0, 1]$. We choose $\lambda_k = 0$ and $\sigma(x_k) = \|\nabla f(x_k)\|\ c$, such that (1) holds. Then Theorem 1 yields the convergence of the function values $f(x_k)$ to some value f_* and by (4)

$$\begin{aligned} \infty > \sum_{k=1}^{\infty} \alpha_k \sigma(x_k) &= \sum_{k=1}^{\infty} \alpha_k \|\nabla f(x_k)\|\ c \\ &= c \sum_{k=1}^{\infty} \overline{\alpha}_k\ \|\nabla f(x_k)\|^2\ \geq c\alpha_- \sum_{k=1}^{\infty} \|\nabla f(x_k)\|^2 \end{aligned}$$

and we have $\|\nabla f(x_k)\| \longrightarrow 0.$ ■

The main disadvantage of Theorem 4 is the fact that an explicit knowledge of L is required in order to estimate α_+.

6. Nonmonotone step size rule by Zhang and Hager

In a recent paper, Zhang and Hager (2004) derived some results for nonmonotone line searches in connection with a modified Armijo's rule or Wolfe's rule.

In the case of Armijo's rule the setup is identical with the one in the Armijo's algorithm outlined in the previous section. The difference is related to the parameter ν_k. In Zhang and Hager (2004) it is assumed that the iterates x_k and directions d_k satisfy the descent condition $\langle \nabla f(x_k), d_k \rangle < 0$. If $\alpha_{max} > 0$ satisfies for some $\varrho > 0$

$$f(x_k + \alpha_{max} d_k) \leq c_k + \varrho\, \alpha_{max}\, \langle \nabla f(x_k), d_k \rangle,$$

then set $\alpha_k = \alpha_{max}$, otherwise find the smallest $l_k \in I\!N$ such that for $\alpha_k = \alpha_{max}\beta^{l_k}$ we have

$$f(x_{k+1}) = f(x_k + \alpha_k d_k) \leq c_k + \varrho\, \alpha_k\, \langle \nabla f(x_k), d_k \rangle. \tag{26}$$

The sequence c_k is updated according to the rule

$$c_{k+1} = (\eta_k q_k c_k + f(x_{k+1}))/q_{k+1}, \qquad q_{k+1} = \eta_k q_k + 1, \tag{27}$$

where $\eta_k \in [\eta_{min}, \eta_{max}]$ with numbers $0 \leq \eta_{min} \leq \eta_{max}$. As initial values we choose $c_0 = f(x_0)$ and $q_0 = 1$.

Note that, if $\eta_k = 0$, the classical Armijo-rule occurs as a special case. On the other hand, if $\eta_k > 0$, then c_{k+1} is a convex combination of c_k and $f(x_{k+1})$. Since c_k is also constructed as a convex combination, c_{k+1} can be regarded as a convex combination of all previous function values $f(x_{k+1})$:

$$c_{k+1} = \gamma_0 f(x_0) + \gamma_1 f(x_1) + ... + \gamma_{k+1} f(x_{k+1}).$$

Note that in (26), in comparison with (2), the value of $f(x_k)$ is replaced by c_k, thus allowing for a nonmonotone behavior of the function values.

It is obvious that by setting $\nu_k = c_k - f(x_k)$ we are in the framework of Section 4. Then the following theorem can be proved based on the results from the previous section. For Armijo's rule it coincides with the result in Zhang and Hager (2004, Theorem 2.2).

Theorem 5 *Let the assumption on the function f of Theorem 3 hold. If, in addition, for the sequence $x_{k+1} = x_k + \alpha_k d_k$ produced by Zhang and Hager's Algorithm there exist constants $a_1, a_2 > 0$ such that*

$$\langle \nabla f(x_k), d_k \rangle \leq -a_1 \|\nabla f(x_k)\|^2, \qquad \|d_k\| \leq a_2 \|\nabla f(x_k)\| \tag{28}$$

and if $\eta_{max} < 1$, then

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

Proof. We can easily apply Theorem 3 and Corollary 4, if we can prove for the quantities $\nu_k = c_k - f(x_k)$ that $\sum_{k=1}^{\infty} \nu_k < \infty$ holds.

Using (28) we obtain from (26) that

$$f(x_{k+1}) \leq c_k \tag{29}$$

holds.

We obtain from (27) that

$$c_{k+1} = \frac{\eta_k q_k c_k + f(x_{k+1})}{q_{k+1}} = \frac{(q_{k+1} - 1)c_k + f(x_{k+1})}{q_{k+1}} = c_k + \frac{f(x_{k+1}) - c_k}{q_{k+1}}. \tag{30}$$

With (29) we have

$$0 \leq \frac{c_k - f(x_{k+1})}{q_{k+1}} = c_k - c_{k+1}.$$

Using a telescope series argument and (29) again we obtain

$$0 \leq \sum_{k=0}^{n} \frac{c_k - f(x_{k+1})}{q_{k+1}} = c_0 - c_{n+1} \leq c_0 - f(x_{n+2}) \leq c_0 - f_m. \tag{31}$$

On the other hand equation (30) implies

$$\nu_{k+1} = c_{k+1} - f(x_{k+1}) = (1 - \frac{1}{q_{k+1}})(c_k - f(x_{k+1})) = (q_{k+1} - 1)\frac{c_k - f(x_{k+1})}{q_{k+1}} \geq 0,$$

because $q_k > 1$ by definition and $c_k \geq f(x_{k+1})$ as shown in (29). The sequence of q_k is bounded, because $\eta_k \leq \eta_{max} < 1$

$$q_{k+1} = \eta_k q_k + 1 \leq \eta_{max} q_k + 1 \leq ... \leq \eta_{max}^{k+1} + \eta_{max}^{k} + ... + 1 \leq \frac{1}{1 - \eta_{max}}.$$

From (31)

$$\begin{aligned} 0 \leq \sum_{k=1}^{\infty} \nu_k &= \sum_{k=1}^{\infty} (q_k - 1)\frac{c_{k-1} - f(x_k)}{q_k} \leq \frac{\eta_{max}}{1 - \eta_{max}} \sum_{k=1}^{\infty} \frac{c_{k-1} - f(x_k)}{q_k} \\ &\leq \frac{\eta_{max}}{1 - \eta_{max}}(c_0 - f_m) < \infty. \end{aligned}$$

■

7. Discretization of optimization problems

The here developed theory can also be applied to finite-dimensional discretizations of infinite-dimensional optimization problems. A whole class of examples are PDE-constrained optimization problems as outlined in Tröltzsch (2010). In order to indicate some potential applications, we take a quick look at an idealized situation:

Consider a problem in Hilbert space X with $f : X \to I\!R$

$$\min_{x \in X} f(x).$$

Let X_n be a sequence of nested finite-dimensional subspaces of X

$$X_1 \subset ... \subset X_n \subset X_{n+1} \subset ... \subset X. \tag{32}$$

For functions $f^n : X_n \to I\!R$ we can consider a sequence of finite-dimensional optimization problems

$$\min_{x \in X_n} f^n(x).$$

Suppose that we refine the meshes as the iteration progresses, i.e. we have a sequence $n(k) \to \infty$ for $k \to \infty$. If we choose a fixed predetermined step size rule, it would yield the same sequence of numbers α_k for all levels of discretization. Then the implementable finite-dimensional version of a steepest descent method is given by

$$x^n_{k+1} = x^n_k - \alpha_k \nabla f^n(x^n_k).$$

For each fixed discretization level n, the standard theory would yield some convergence statement of the type $\nabla f^n(x^n_k) \to 0$ for $k \to \infty$ and n fixed. It would be more meaningful to have a statement, where the original gradients tend to zero, like $\nabla f(x^{n(k)}_k) \to 0$ for $k \to \infty$.

We see below how this can be achieved as a simple application of the general theory developed above.

Theorem 6 *Let $f : X \to I\!R, f^n : X_n \to I\!R$ be a family of functions defined on a sequence of nested subspaces X_n of X as in (32). Let $n(k)$ be a sequence with $n(k) \to \infty$ for $k \to \infty$. Define a sequence $x^{n(k)}_k \in X_{n(k)}$ through a steepest descent method where the discretization is refined at each step*

$$x^{n(k+1)}_{k+1} = x^{n(k)}_k - \alpha_k \nabla f^{n(k)}(x^{n(k)}_k). \tag{33}$$

Let the family of functions satisfy the estimate

$$\|\nabla f^{n(k)}(x) - \nabla f(x)\| \leq \tau_k \qquad \text{for all} \quad x \in S, \tag{34}$$

where $S \subset X$ contains all iterates. If the sequences α_k, τ_k are chosen such that

$$\sum_{k=1}^{\infty} \alpha_k^2 < \infty, \qquad \sum_{k=1}^{\infty} \alpha_k = \infty, \qquad \sum_{k=1}^{\infty} \alpha_k \tau_k < \infty, \tag{35}$$

then

$$\nabla f(x^{n(k)}_k) \to 0 \qquad k \to \infty. \tag{36}$$

Proof. Due to the nested subspaces, we can consider the sequence $x_k^{n(k)}$ as a sequence in X. Then from (33)

$$x_{k+1}^{n(k+1)} = x_k^{n(k)} + \alpha_k d_k, \quad d_k = -\nabla f^{n(k)}(x_k^{n(k)}).$$

This is the setting of Theorem 2, where by (34)

$$\|r_k\| = \|d_k + \nabla f(x_k^{n(k)})\| = \|\nabla f^{n(k)}(x_k^{n(k)}) - \nabla f(x_k^{n(k)})\| \leq \tau_k.$$

Considering assumption (35) we see that in Theorem 2 the assumptions (6) and (7) are satisfied. Hence the conclusions are true and since the assumption (14) in Corollary 1 also holds, we obtain its conclusion, which is exactly (36). ■

Usually, the discretization is parameterized by a mesh size parameter h and for the gradient errors one might have $\tau_k = O(h_k^p)$ for some $p > 0$. A typical choice for the step sizes would be $\alpha_k = 1/k$ such that (35) is satisfied if

$$\sum_{k=1}^{\infty} \alpha_k h_k^p < \infty$$

holds.

A typical application of such a theorem is shown in the following situation. If we can extract a convergent subsequence $x_{k_j}^{n(k_j)} \to x_*$, then the statement (36) yields immediately $\nabla f(x_*) = 0$, the stationarity of x_* for the original infinite-dimensional problem.

8. Conclusions

We consider a general setup for nonmonotone step size rules which allows to show global convergence in the sense that the gradients of the objective functions evaluated at the iterates converge to zero. We show how this can be applied to various instances of nonmonotone step size rules. In particular, this yields a new proof of an Armijo-type step size rule developed by Hager and Zhang. Furthermore, the potential application in discretization schemes of optimization with PDEs is indicated.

References

Barzilai, J. and Borwein, J. (1988) Two-point step size gradient methods. *IMA Journal Numer. Anal.*, **8** , 141–148.

Grippo, L., Lampariello, F. and Lucidi, S. (1986) A nonmonotone line search technique for Newton's method. *SIAM Journal on Numerical Analysis*, **23**, 707–716.

Grippo, L., Lampariello, F. and Lucidi, S. (1989) A truncated Newton method with nonmonotone line search for unconstrained optimization. *Journal of Optimization Theory and Applications*, **60**, 401–419.

Hüther, B. (2002) Global convergence of algorithms with nonmonotone line search strategy in unconstrained optimization. *Results Math.*, **41**, 320–333.

Kaplan, A. and Tichatschke, R. (1994) *Stable Methods for Ill-Posed Variational Problems.* Akademie-Verlag, Berlin.

Shi, Z.J. and Shen, J. (2006) Convergence of nonmonotone line search method. *Journal of Computational and Applied Mathematics*, **193**, 397–412.

Tröltzsch, F. (2010) *Optimal Control of Partial Differential Equations.* AMS.

Warth, W. and Werner, J. (1977) Effiziente Schrittweitenfunktionen bei unrestringierten Optimierungsaufgaben. *Computing*, **19**, 59–72.

Zhang, H. and Hager, W.W. (2004) A nonmonotone line search technique and its application to unconstrained optimization. *SIAM Journal on Optimization*, **14**, 1043–1046.