Line search and convergence in bound-constrained optimization

Arnold Neumaier

Fakultät für Mathematik, Universität Wien
Oskar-Morgenstern-Platz 1, A-1090 Wien, Austria
email: Arnold.Neumaier@univie.ac.at
WWW: `http://www.mat.univie.ac.at/~neum`

Behzad Azmi

Johann Radon Institute for Computational and Applied Mathematics (RICAM)
Austrian Academy of Sciences
Altenbergerstraße 69, A-4040 Linz, Austria
email: behzad.azmi@ricam.oeaw.ac.at
WWW: `https://people.ricam.oeaw.ac.at/b.azmi/`

March 27, 2019

Abstract. The first part of this paper discusses convergence properties of a new line search method for the optimization of continuously differentiable functions with Lipschitz continuous gradient. The line search uses (apart from the gradient at the current best point) function values only. After deriving properties of the new, in general curved, line search, global convergence conditions for an unconstrained optimization algorithm are derived and applied to prove the global convergence of a new nonlinear conjugate gradient (CG) method. This method works with the new, gradient-free line search – unlike traditional nonlinear CG methods that require line searches satisfying the Wolfe condition.

In the second part, a class of algorithms is developed for bound constrained optimization. The new scheme uses the gradient-free line search along bent search paths. Unlike traditional algorithms for bound constrained optimization, our algorithm ensures that the reduced gradient becomes arbitrarily small. It is also proved that all strongly active variables are found and fixed after finitely many iterations.

A Matlab implementation of a bound constrained solver based on the new theory is discussed in the companion paper *LMBOPT – a limited memory program for bound-constrained optimization* by M. Kimiaei and the present authors.

Contents

1 Introduction

This paper discusses theoretical properties of line search methods for the bound constrained optimization problem

$$\begin{array}{ll} \min & f(x) \\ \text{s.t.} & x \in \mathbb{R}^n, \quad \underline{x} \leq x \leq \overline{x}, \end{array} \tag{1}$$

where f is continuously differentiable with Lipschitz continuous gradient. This contains the unconstrained case when all bounds are infinite.

Convergence results for both the unconstrained and bound constrained have a long history. For complete references we refer to the book by Nocedal & Wright [14]; here we concentrate on tracing only those references relevant for the present work that are aimed at

an improved convergence theory for line search methods valid under significantly weaker assumptions than before.

Most line search algorithms are based on satisfying the Wolfe conditions (Wolfe [17]), which require a gradient evaluation at every trial point but match the requirements in typical modern convergence proofs. We shall instead focus on (possibly curved) line searches that use (apart from the gradient at the current best point) function values only. In the past, such line searches were based on satisfying the Goldstein conditions (Goldstein [8]), which are known to behave poorly in strongly nonconvex regions. Our new line search (presented in Section 3) is based on satisfying a new sufficient descent condition (9) that removes this weakness and still produces efficient steps (Theorem 3.1) in the sense of Warth & Werner [16].

We prove in Theorem 4.1 a variant of their result that this condition together with a weak condition on the search directions suffices for global convergence. This global convergence result is used to prove in Theorem 7.3 the global convergence of a new nonlinear conjugate gradient (CG) method that – unlike traditional nonlinear CG methods that need line searches satisfying the Wolfe condition – works with the new, gradient-free line search. The new CG method is motivated by the desire to reduce the inefficiency of line search methods due to zigzagging of the search directions, discussed in Section 6. The search direction is therefore chosen by minimizing (Theorem 7.1) a preconditioned distance to the previous search direction.

In the second part we show how to utilize the new line search to solve the bound constrained optimization problem. Because of the bound constraints, the search path must now be bent (Bertsekas [2]) in order to produce feasible points only. We therefore discuss in detail (Section 9) the properties of bent search paths that are instrumental for an analysis of the descent properties. We then formulate (Algorithm 9.1) a generic algorithm for solving bound constrained optimization problems using a gradient-free bent line search.

Global convergence of this generic algorithm is then proved in Section 11. Its good theoretical properties with regard to zigzagging are established by proving that all strongly active variables are found and fixed after finitely many iterations (Theorem 8.2). Thus the new algorithm shares this property with the traditional active set methods by Bertsekas [2], Conn et al. [5], and Hager & Zhang [10].

An implementation of a bound constrained solver based on the new theory must take care of many other questions not covered by the theory, in particular regarding finite precision effects. A discussion of such implementation questions, details for a particular implementation in Matlab and Java, and a thorough comparison with other state of the art solvers is given in the companion paper Kimiaei et al. [13].

Acknowledgments. Earlier versions of this paper benefitted from discussions with Waltraud Huyer, Morteza Kimiaei, Hermann Schichl.

2 Smooth bound-constrained optimization problems

Inequalities between vectors or matrices are interpreted component-wise. For an arbitrary norm $\|\cdot\|$, the **dual norm** $\|\cdot\|_*$ is defined by

$$\|y\|_* := \sup_{x \neq 0} \frac{y^T s}{\|s\|},$$

so that the **generalized Cauchy–Schwarz inequality**

$$|y^T s| \leq \|y\|_* \|s\|$$

holds. To be numerically appropriate, the norm must give a sensible measure of distance between points where we evaluate functions. For example, one could use a **scaled 1-norm**, with

$$\|s\| := \sum_k \left| \frac{s_k}{w_k} \right|,$$

where $w_k > 0$ is a weight specifying the typical magnitude x_k of the kth component of a trial point. In this case, the dual norm is a **scaled maximum norm**, with

$$\|y\|_* := \max_k w_k |y_k|.$$

Another useful pair of norms are the **ellipsoidal norms**

$$\|p\| = \sqrt{p^T B p}, \quad \|g\|_* = \sqrt{g^T B^{-1} g} \tag{2}$$

defined in terms of a symmetric positive definite matrix $B \in \mathbb{R}^{n \times n}$. Using a Cholesky factorization $B = R^T R$ and a linear transformation $p' = Rp$, $g' = R^{-T} g$, where R^{-T} denotes the transposed inverse of R, it is easy to check that these indeed form a pair of dual norms, so that

$$|g^T p| \leq \|g\|_* \|p\|.$$

For the identity matrix $B = I$, (2) becomes the standard **Euclidean norm** $\|s\|_2 := \sqrt{s^T s}$, which is its own dual. At times we assume that a norm is **monotone**, i.e., it satisfies

$$|s| \leq |s'| \quad \Rightarrow \quad \|s\| \leq \|s'\|, \tag{3}$$

and hence also $\|s\| = \| \, |s| \, \|$. (Here, as always later, the absolute value of and inequalities between vectors are componentwise.) This is the case for scaled 1-norms, scaled maximum norms, the Euclidean norm. But ellipsoidal norms are monotone only if B is a diagonal matrix.

We consider the **bound-constrained optimization problem**

$$\begin{array}{ll} \min & f(x) \\ \text{s.t.} & x \in \mathbf{x}, \end{array} \tag{4}$$

where the **objective function** $f : C \subseteq \mathbb{R}^n \to \mathbb{R}$ is a continuously differentiable function, and

$$\mathbf{x} := [\underline{x}, \overline{x}] := \{x \in \mathbb{R}^n \mid \underline{x} \le x \le \overline{x}\}$$

is a bounded or unbounded **box** in $\mathbb{R}^n$ describing the bounds on the variables. One-sided or missing bounds are accounted for by allowing components of the vector $\underline{x}$ of lower bounds to take the value $-\infty$ and components of the vector $\overline{x}$ of upper bounds to take the value ∞. A point x is called **feasible** if it belongs to the box $\mathbf{x}$. To have a well-defined optimization problem, the box $\mathbf{x}$ must be part of the domain C of definition[1] of f. We assume that the **gradient**

$$g(x) := \partial f(x)/\partial x = f'(x)^T \in \mathbb{R}^n.$$

is **Lipschitz continuous** in the feasible domain, i.e.,

$$\|g(x') - g(x)\|_* \le \overline{\gamma}\|x' - x\| \quad \text{for } x, x' \in \mathbf{x}. \tag{5}$$

The **Lipschitz constant** $\overline{\gamma}$ depends on the norm used, but not the notion of Lipschitz continuity, since all norms in $\mathbb{R}^n$ are equivalent.

The optimization methods discussed here improve an initial feasible point x^0 by constructing a sequence $x^0, x^1, x^2, \ldots$ of feasible points with decreasing function values. To ensure this, we search in each iteration along an appropriate search path $x(\alpha)$ starting at the current point $x(0) = x^\ell$, and take $x^{\ell+1} = x(\alpha_\ell)$ where α_ℓ is determined by a line search (discussed in Section 3) based on function values only. If the iteration index ℓ is fixed, we simply write x for the current point x^ℓ.

The actual optimization typically proceeds through three phases with distinct characteristics. In the initial phase, one moves down into a valley; the search direction is of minor importance, and most activities are correctly adjusted if appropriately bent search paths (see Section 9) are used. In the second, intermediate phase, one moves along the valley towards the minimizer. This phase may be long if the valley is long, steep and curved, or short and even absent if the valley is fairly round. To be sure to come close to the minimizer, the search directions must conform to conditions that allow one to prove convergence of the method; see Sections 4, 5, and 11. To be efficient in this phase, one also needs to take measures against various forms of zigzagging; see Section 6.

In the final phase, one is close to the minimizer but has to locate it to the desired accuracy. Here a good choice of search direction is essential. As near a minimizer the function is typically almost quadratic, a good method must select in this phase search direction and step sizes in a way that a good behavior on quadratic functions is guaranteed. We shall utilize for this purpose approximate conjugate directions; see Section 7.

[1] In practice, one may allow a smaller domain of definition if f satisfies the **coercivity condition** that, as x approaches the boundary of the domain of definition, $f(x)$ exceeds the function value $f(x^0)$ at a known starting point x^0. Also, Lipschitz continuity may be relaxed to local Lipschitz continuity if all evaluation points remain in a bounded region.

3 The line search

A **line search** proceeds by searching points $x(\alpha)$ on a curve of feasible points parameterized by a **step size** $\alpha > 0$ starting at the current point $x = x(0)$. The goal is to find a value for the step size such that $f(x(\alpha))$ is sufficiently smaller than $f(x)$, with a notion of "sufficiently" to be made precise. If the gradient $g = g(x)$ is nonzero, the existence of such an $\alpha > 0$ is guaranteed if the tangent vector

$$p := x'(0) \tag{6}$$

exists and satisfies

$$g^T p < 0; \tag{7}$$

a vector p with (7) is called a **descent direction**. In the unconstrained case, the curve is frequently taken to be a ray from x in a descent direction p, giving $x(\alpha) = x + \alpha p$. In this case, we call the line search **straight**; if the curve is a piecewise-linear path, **bent**; and otherwise **curved**.

A good and computationally useful measure of progress of a line search is the **Goldstein quotient** (first considered by Goldstein [9])

$$\mu(\alpha) := \frac{f(x(\alpha)) - f(x)}{\alpha g(x)^T p} \quad \text{for} \quad \alpha > 0. \tag{8}$$

μ can be extended to a continuous function on $[\,0, \infty]$ by defining $\mu(0) := 1$ since, by l'Hôpital's rule,

$$\lim_{\alpha \to 0} \mu(\alpha) = \lim_{\alpha \to 0} \frac{f'(x(\alpha))x'(\alpha)}{g(x)^T p} = \frac{f'(x)x'(0)}{g^T p} = 1.$$

Since by assumption $g^T p < 0$, we have $f(x(\alpha)) < f(x)$ iff $\alpha > 0$ and $\mu(\alpha) > 0$. Restrictions on the values of the Goldstein quotient define regions where sufficient descent is achieved. We consider here the **sufficient descent condition**

$$\mu(\alpha)|\mu(\alpha) - 1| \geq \beta \tag{9}$$

with fixed $\beta > 0$. This condition requires $\mu(\alpha)$ to be both not too close to one, forbidding steps that are too short, and sufficiently positive, typically forbidding steps that are too long by forcing $f(x(\alpha)) < f(x)$. The condition is closely related to the so-called **Goldstein condition**

$$f(x) + \alpha\mu'' g^T p \leq f(x(\alpha)) \leq f(x) + \alpha\mu' g^T p, \tag{10}$$

where $0 < \mu' < \mu'' < 1$. Indeed, (10) is equivalent to

$$\mu' \leq \mu(\alpha) \leq \mu'', \tag{11}$$

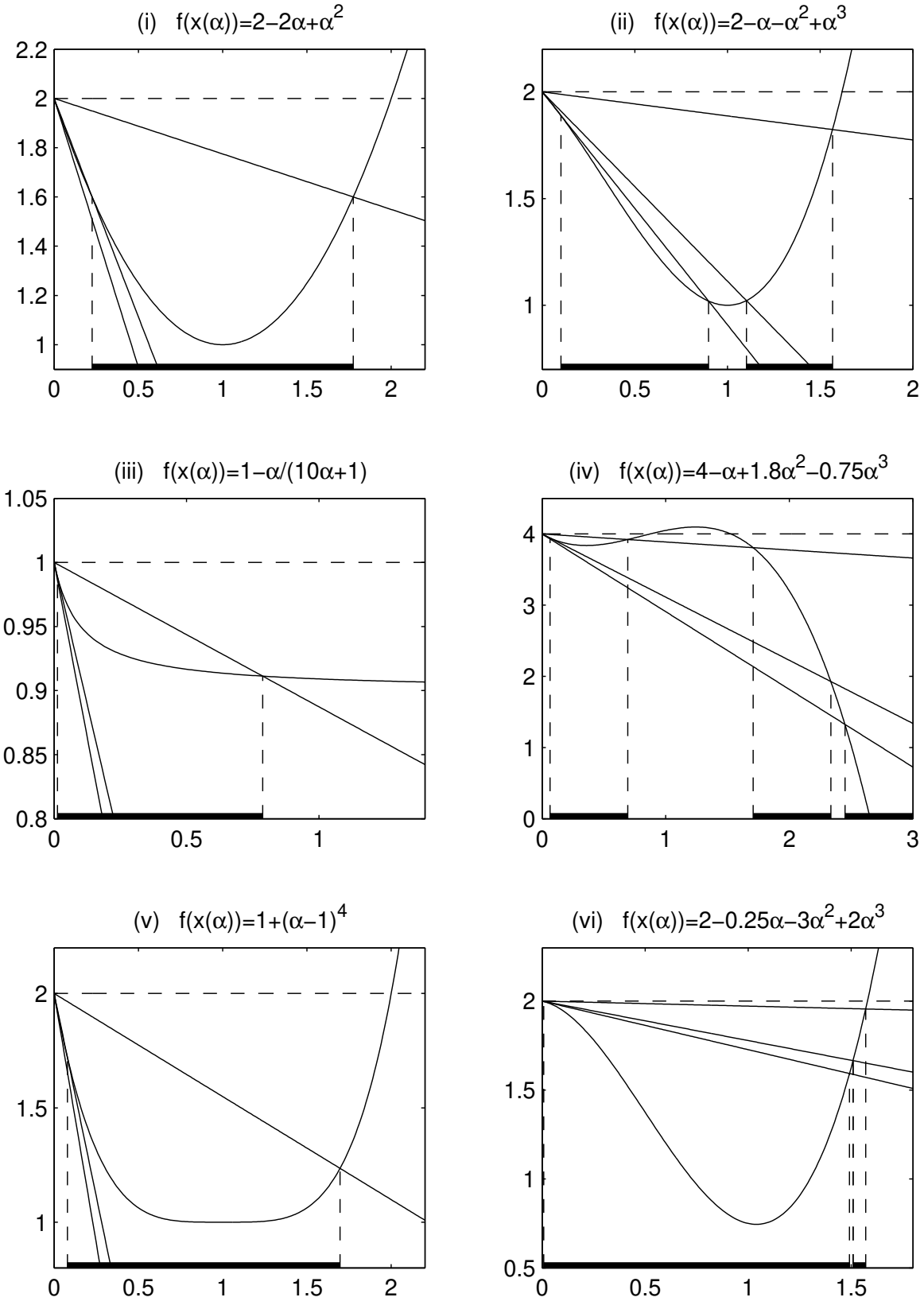
hence (9) holds with

$$\beta = \mu'(1 - \mu'') > 0.$$

Conversely, (9) implies that either (10) holds with

$$\mu' = \frac{2\beta}{1 + \sqrt{1 - 4\beta}}, \quad \mu'' = \frac{1 + \sqrt{1 - 4\beta}}{2},$$

Figure 1: The Goldstein condition with tuning parameter $\beta = 0.1$. Drawn in each case are the lines with slopes $\mu' g^T p$, $\mu'' g^T p$, $\mu''' g^T p$, and the resulting set of acceptable step sizes.



or the alternative **fast descent condition**

$$\mu(\alpha) \geq \mu''' \tag{12}$$

holds with

$$\mu''' = \frac{1+\sqrt{1+4\beta}}{2}.$$

The Goldstein condition (11) can be interpreted geometrically: In the graph of $f(x(\alpha))$, the cone defined by the two lines through $(0, f)$ with slopes $\mu' g^T p$ and $\mu'' g^T p$ cuts out a section of the graph, which defines the admissible step size parameters. Similarly, equality in (12) defines another line that determines the boundary of another section of the graph leading to admissible step size parameters. Some illustrative examples are given in Figure 1.

By the preceding discussion, satisfying the sufficient descent condition (9) guarantees a sensible decrease in the objective function. Indeed, we shall prove an explicit bound on the gain. It will be essential to get later global convergence statements.

3.1 Theorem. *Suppose that the function f has a continuous gradient g satisfying the Lipschitz condition*

$$\|g(x) - g(x')\|_* \leq \overline{\gamma}\|x - x'\|.$$

If the restriction of the search path to $[0, \alpha]$ is a ray and $\alpha > 0$ satisfies the sufficient descent condition (9) *then*

$$(f(x) - f(x(\alpha'))\frac{\|p\|^2}{(g(x)^T p)^2} \geq \frac{2\beta}{\overline{\gamma}} \tag{13}$$

holds for any step size α' with $f(x(\alpha')) \leq f(x(\alpha))$.

Proof. By assumption,

$$x(\alpha') = x + \alpha' p \quad \text{for } 0 \leq \alpha' \leq \alpha.$$

The function ψ defined by

$$\psi(\alpha) := f(x + \alpha p) - \alpha g(x)^T p$$

satisfies

$$\psi'(\alpha) = g(x + \alpha p)^T p - g(x)^T p = (g(x + \alpha p) - g(x))^T p.$$

The generalized Cauchy–Schwarz inequality gives

$$|\psi'(\alpha)| \leq \|g(x + \alpha p) - g(x)\|_* \|p\| \leq \overline{\gamma}\|\alpha p\|\|p\| = \overline{\gamma}\alpha \, \|p\|^2,$$

hence

$$\begin{aligned} |f(x + \alpha p) - f(x) - \alpha g(x)^T p| &= |\psi(\alpha) - \psi(0)| = \Big| \int_0^\alpha \psi'(t)dt \Big| \\ &\leq \int_0^\alpha |\psi'(t)|dt \leq \int_0^\alpha \overline{\gamma} t \|p\|^2 dt = \frac{\overline{\gamma}\alpha^2}{2}\|p\|^2. \end{aligned}$$

Therefore

$$\frac{\|p\|^2}{|g(x)^T p|} \geq \frac{2}{\alpha\overline{\gamma}}\Big|\frac{f(x+\alpha p)-f(x)-\alpha g(x)^T p}{\alpha g(x)^T p}\Big| = \frac{2}{\alpha\overline{\gamma}}|\mu(\alpha)-1|.$$

On the other hand, since $g(x)^T p < 0$,

$$\frac{f(x)-f(x(\alpha))}{|g(x)^T p|} = \frac{f(x(\alpha))-f(x)}{g(x)^T p} = \alpha\mu(\alpha).$$

Taking the product, we conclude that

$$(f(x)-f(x(\alpha'))\frac{\|p\|^2}{(g(x)^T p)^2} \geq \alpha\mu(\alpha)\frac{2}{\alpha\overline{\gamma}}|\mu(\alpha)-1| = \frac{2\mu(\alpha)|\mu(\alpha)-1|}{\overline{\gamma}} \geq \frac{2\beta}{\overline{\gamma}}.$$

□

A line search satisfying the conclusion of Theorem 3.1 is called an **efficient line search** (Warth & Werner [16]).

Near a local minimizer, twice continuously differentiable functions are bounded from below and, because of Taylor's theorem, almost quadratic. For a linear search path and a stricly convex quadratic function,

$$f(x+\alpha p) = f(x) + \alpha g(x)^T p + \frac{\alpha^2}{2} p^T G(x) p =: f + a\alpha + b\alpha^2 = f - \frac{a^2}{4b} + b(\alpha - \widehat{\alpha})^2$$

with

$$a < 0 < b, \quad \widehat{\alpha} = -\frac{a}{2b} > 0.$$

This implies that

$$\mu(\alpha) = 1 + \frac{b\alpha}{a} = 1 - \frac{\alpha}{2\widehat{\alpha}} < 1$$

for $\alpha > 0$. In particular, $\mu(\widehat{\alpha}) = \frac{1}{2}$, and the minimizer

$$\widehat{\alpha} = \frac{\alpha}{2(1-\mu(\alpha))} \tag{14}$$

along the search ray can be computed from any $\alpha > 0$.

A step size satisfying the sufficient descent condition (9) can be found constructively, when the objective function is bounded below.

3.2 Theorem. *Let $\beta \in]0, \frac{1}{4}[$, $g^T p < 0$. If $f(x(\alpha))$ is bounded below then the equation $\mu(\widehat{\alpha}) = \frac{1}{2}$ has a solution $\widehat{\alpha} > 0$, and any α sufficiently close to $\widehat{\alpha}$ satisfies* (9).

Proof. Let $\underline{f} := \inf_{\alpha\geq 0} f(x(\alpha))$ and $\mu_0 := \inf_{\alpha\geq 0} \mu(\alpha)$. If $\mu_0 > 0$ then (8) implies for $\alpha > 0$ the inequality

$$\underline{f} - f(x) \leq f(x(\alpha)) - f(x) = \alpha g^T p \mu(\alpha) \leq \alpha g^T p \mu_0, \tag{15}$$

but since $g^T p < 0$, this is impossible for sufficiently large α. Therefore $\mu_0 \leq 0$. By continuity, $\mu(\widehat{\alpha}) = \frac{1}{2}$ has a solution $\widehat{\alpha} > 0$. Since $\mu(\widehat{\alpha})|\mu(\widehat{\alpha}) - 1| = \frac{1}{4} > \beta$, (9) holds for all α sufficiently close to $\widehat{\alpha}$. □

Based on the theory so far it is natural to attempt to find a step size α with $\mu(\alpha) \approx \frac{1}{2}$. This can be done by a simple bisection procedure.

3.3 Algorithm. Curved line search (CLS)

Purpose: Finds a step size α with $\mu(\alpha)|\mu(\alpha) - 1| \geq \beta$
Input: $x(\alpha)$ (search path), $f_0 = f(x(0))$, $\nu = -g(x(0))^T x'(0)$
α_{init} (initial step size), $\alpha_{\max}$ (maximal step size),
Requirements: $\nu > 0$, $0 < \alpha_{\text{init}} \leq \alpha_{\max} \leq \infty$
Parameters: $\beta \in]0, \frac{1}{4}[$, $q > 1$

```
first=1;
α̲ = 0; ᾱ = ∞; α = α_init;
while 1,
   μ = (f_0 − f(x(α)))/(αν);
   if μ|μ − 1| ≥ β, break; end;
   if μ ≥ 1/2, α̲ = α;
   elseif α = α_max, break;
   else ᾱ = α; % linear decrease or more
   end;
   if first,
      first=0;
      if μ < 1, α = (1/2)α/(1 − μ); else α = αq; end;
   else
      if ᾱ = ∞,     α = αq;
      elseif α̲ = 0, α = (1/2)α/(1 − μ);
      else          α = sqrt(α̲ ᾱ);
      end;
   end;
   α = min(α, α_max);
end;
return α;
```

The Boolean variable `first` in the while loop ensures that the quadratic case will be optimally handled. In the first iteration we use the formula (14) whenever $\mu(\alpha) < 1$. If the resulting next value for μ does not satisfy the sufficient descent condition (9), the function is far from quadratic and bounded, and we proceed with a simple bisection scheme: Until we know bounds $\widehat{\alpha} \in [\underline{\alpha}, \overline{\alpha}]$ with $\underline{\alpha} > 0$ and $\overline{\alpha} < \infty$, we interpolate with (14), but we

extrapolate with a constant factor $q > 1$. Once such a bracket $[\underline{\alpha}, \overline{\alpha}]$ is found we use geometric mean steps since the bracket may span several orders of magnitude. However, we quit the line search once the stopping test is satisfied and return the final step size α.

Because of Theorem 3.1, Algorithm 3.3 defines (for $\alpha_{\max} = \infty$) an efficient line search and achieves a well-defined minimal reduction in the function value.

Note that in a computer implementation, this idealized line search needs an extra stopping test to ensure that it ends after finitely many steps even when f is unbounded below along the search curve. In addition, one need to take measures that make the line search robust in the presence of rounding errors by forbidding steps that are so small that the change in function value is dominated by rounding errors. Details are discussed in the companion paper Kimiaei et al. [13].

In practice, one may use a small value such as $\beta = 0.02$, and a large value of q such as $q = 25$. The best values depend on the particular algorithm calling the line search, and must be determined by calibration on a set of test problems.

In all cases where for small α, the graph of $f(x(\alpha))$ is – as in Figure 1(vi) – concave and fairly flat, while for larger α, $f(x(\alpha))$ is strongly increasing, the traditional Goldstein condition (10) allows – unlike the present sufficient descent condition – only a tiny and inefficient range of step sizes. This is one of the reasons why many currently used line searches also involve the so-called Wolfe condition, which needs gradient evaluations during the line search.

The present line search is gradient-free but still avoids this defect of the Goldstein condition. Indeed, in the above cases, the range allowed by the sufficient descent condition (9) is considerably larger than that of the Goldstein condition, since it includes the values where (12) holds.

4 Convergence – unconstrained case

For any sequence $x^0, x^1, x^2, \ldots$ of feasible points (typically generated by an optimization algorithm) and $\ell = 0, 1, 2, \ldots$, we write

$$f_\ell := f(x^\ell), \quad g^\ell := g(x^\ell),$$

$$s^\ell := x^{\ell+1} - x^\ell, \quad y^\ell := g^{\ell+1} - g^\ell. \tag{16}$$

We refer to the s^ℓ as **steps**. If straight line searches in the directions p^ℓ are used then $s^\ell = \alpha_\ell p^\ell$, where $\alpha_\ell > 0$ is the step length accepted in the ℓth step.

The following is a variant of a convergence result by Warth & Werner [16].

4.1 Theorem. *Suppose that, for* $\ell = 1, 2, 3, \ldots$,

$$\sigma_\ell := |(g^\ell)^T s^\ell| > 0, \tag{17}$$

$$\sup_\ell \|g^\ell\|_*^2 \Big(\frac{\|s^\ell\|^2}{\sigma_\ell^2} - \frac{\|s^{\ell-1}\|^2}{\sigma_{\ell-1}^2} \Big) < \infty. \tag{18}$$

$$\inf_\ell (f_\ell - f_{\ell+1}) \frac{\|s_\ell\|^2}{\sigma_\ell^2} > 0, \tag{19}$$

Then

$$\inf_\ell \|g^\ell\|_* = 0 \quad \textit{or} \quad \lim_{\ell\to\infty} f_\ell = -\infty. \tag{20}$$

Proof. Suppose that $\inf \|g^\ell\|_* = \varepsilon > 0$; we need to show that $f_\ell \to -\infty$ for $\ell \to \infty$. We write κ for the supremum in (18), and find

$$\frac{\|s^\ell\|^2}{\sigma_\ell^2} - \frac{\|s^{\ell-1}\|^2}{\sigma_{\ell-1}^2} \le \frac{\kappa}{\|g^\ell\|_*^2} \le \frac{\kappa}{\varepsilon^2} \le \kappa' := \max\Big(\frac{\kappa}{\varepsilon^2}, \frac{\|s^1\|^2}{\sigma_1^2}\Big)$$

Summation over all steps gives

$$\frac{\|s^\ell\|^2}{\sigma_\ell^2} \le \ell\kappa' \quad \text{for } \ell > 0.$$

If we write δ for the infimum in (19), we may therefore conclude that

$$f_\ell - f_{\ell+1} \ge \frac{\delta\sigma_\ell^2}{\|s^\ell\|^2} \ge \frac{\delta}{\ell\kappa'}.$$

Summation over all steps gives $f_\ell \le f_0 - \dfrac{\delta}{\kappa'} \displaystyle\sum_{i=1}^{\ell-1} \frac{1}{i} \to -\infty$ as $\ell \to \infty$. □

According to the theorem, if all three conditions (17)–(19) hold, we either find arbitrarily large negative function values, or we come arbitrarily close to a stationary point – typically a local minimizer (unless we accidently hit directly on a stationary point or a symmetry-protected saddle leading to such a point). In both cases, the unconstrained optimization problem $f(x) = \min!$ may be considered as solved.[2]

To obtain results about the speed of convergence we need to make stronger assumptions. We call a point $\widehat{x} \in \mathbb{R}^n$ a **strong local minimizer** of f if f is twice continuously differentiable in a neighborhood of $\widehat{x}$, the gradient $g(\widehat{x})$ of f at $\widehat{x}$ vanishes, and the Hessian $G(\widehat{x})$ of f at $\widehat{x}$ is positive definite.

[2]To be completely sure, we would have to verify the second order sufficient optimality conditions. However, this would require Hessian information, which is often not available. As a consequence, any optimization method using only first order information and finite-precision arithmetic may get stuck in nearly flat regions where rounding errors dominate and produce spurious apparent local minimizers. If Hessian information is available, it can be used to check if it is positive semidefinite (indicating within the numerical accuracy the prsence of a local minimizer). If this is not the case, one can find a direction of negative curvature along which descent is generally possible even in finite precision arithmetic.

4.2 Theorem. *If the x^ℓ converge to a strong local minimizer $\widehat{x}$ and if*

$$\inf_\ell \frac{f_\ell - f_{\ell+1}}{\|g^\ell\|_*^2} > 0 \tag{21}$$

then there are constants $q \in]0,1[$ and $c, c' > 0$ such that, for all sufficiently large ℓ,

$$f_{\ell+1} - f(\widehat{x}) \le q^2(f_\ell - f(\widehat{x})), \tag{22}$$

$$\|x^\ell - \widehat{x}\| \le cq^\ell, \quad \|g^\ell\|_* \le c'q^\ell. \tag{23}$$

(22) and (23) are conventionally expressed by saying that convergence is **locally linear**.

Proof. Since the eigenvalues of a positive definite matrix are positive, the requirements on $\widehat{x}$ imply that there are positive constants $\underline{\gamma}, \overline{\gamma}$ and a ball C around $\widehat{x}$ such that for all $x \in C$, the eigenvalues of the Hessian $G(x)$ are in $[\underline{\gamma}, \overline{\gamma}]$. The remainder form of Taylor's theorem now implies that for $x, x' \in C$,

$$\|g(x') - g(x)\|_* \le \overline{\gamma}\|x' - x\|. \tag{24}$$

$$\frac{1}{2}\underline{\gamma}\|x' - x\|^2 \le f(x') - f(x) - g(x)^T(x' - x) \le \frac{1}{2}\overline{\gamma}\|x' - x\|^2. \tag{25}$$

(For example, (25) follows by a simple modification of an argument in the proof of Theorem 3.1.) Interchanging x and x' in the first inequality of (25), adding the two formulas, and applying the generalized Cauchy–Schwarz inequality gives

$$\underline{\gamma}\|x' - x\|^2 \le (g(x') - g(x))^T(x' - x) \le \|g(x') - g(x)\|_*\|x' - x\|, \tag{26}$$

so that

$$\|x' - x\| \le \underline{\gamma}^{-1}\|g(x') - g(x)\|_*. \tag{27}$$

Since $g(\widehat{x})$=0, (25) and (27) imply

$$\frac{1}{2}\underline{\gamma}\|x^\ell - \widehat{x}\|^2 \le f(x^\ell) - f(\widehat{x}) \le \frac{1}{2}\overline{\gamma}\|x^\ell - \widehat{x}\|^2 \le \kappa\|g^\ell\|_*^2, \tag{28}$$

where $\kappa = \overline{\gamma}/2\underline{\gamma}^2$. Writing $\widehat{f} := f(\widehat{x})$, we find

$$\frac{f_{\ell+1} - \widehat{f}}{f_\ell - \widehat{f}} = 1 - \frac{f_\ell - f_{\ell+1}}{f_\ell - \widehat{f}} \le 1 - \frac{f_\ell - f_{\ell+1}}{\kappa\|g^\ell\|_*^2} \le q^2 := 1 - \inf \frac{f_\ell - f_{\ell+1}}{\kappa\|g^\ell\|_*^2} < 1; \tag{29}$$

the last inequality holds by (21). Therefore (22) holds, and by induction

$$f_\ell - \widehat{f} \le c_0 q^{2\ell}$$

for some constant $c_0 > 0$. Now the first inequality in (28) gives $\|x^\ell - \widehat{x}\| \le cq^\ell$, and then (24) gives $\|g^\ell\|_* \le c'q^\ell$. □

We now discuss how to satisfy the convergence conditions. In the unconstrained case, (17) is usually accomplished by considering methods with

$$x^{\ell+1} = x^\ell + \alpha_\ell p^\ell, \quad \alpha_\ell > 0, \tag{30}$$

where p^ℓ is a descent direction, so that $(g^\ell)^T s^\ell = \alpha_\ell (g^\ell)^T p^\ell < 0$. In this case, $s^\ell = \alpha_\ell p^\ell$, and (17) holds. For this class of methods, we give in the following sections recipes for satisfying (18) in terms of particular choices for the descent directions. The remaining convergence condition (19) is satisfied by Theorem 3.1 if we employ the line search from Algorithm 3.3 with linear search paths.

5 The angle condition

As a special case of our convergence results we obtain the following classical result.

5.1 Theorem. *An optimization method that computes its points by* (30)*, where the search directions satisfy the* ***angle condition***[3]

$$\sup_\ell \frac{(g^\ell)^T p^\ell}{\|g^\ell\|_* \|p^\ell\|} < 0, \tag{31}$$

and uses a linear line search (e.g., Algorithm 3.3) enforcing the sufficient descent condition (9) *satisfies*

$$\inf_\ell \|g^\ell\|_* = 0 \quad \text{or} \quad \lim_{\ell\to\infty} f_\ell = -\infty.$$

Moreover, if the x^ℓ converge to a strong local minimizer then convergence is locally linear.

Proof. Since $\sigma_\ell > 0$ for all l, the angle condition gives (17). Writing $-\delta$ for the value of the supremum, it also implies that

$$\sigma_\ell^2 = \Big((g^\ell)^T s^\ell\Big)^2 \geq \delta^2 \|g^\ell\|_*^2 \|s^\ell\|^2 > 0.$$

Hence (17) holds and

$$\|g^\ell\|_*^2 \frac{\|s^\ell\|^2}{\sigma_\ell^2} \leq \frac{1}{\delta^2}.$$

This implies (18) since the negative terms can be discarded. Since Theorem 3.1 gives (19), Theorem 4.1 applies and shows that either $\inf_\ell \|g^\ell\|_* = 0$ or $\lim_\ell f_\ell = -\infty$. Moreover,

$$\frac{f_\ell - f_{\ell+1}}{\|g^\ell\|_*^2} = (f_\ell - f_{\ell+1}) \frac{\|p^\ell\|^2}{((g^\ell)^T p^\ell)^2} \Big(\frac{(g^\ell)^T p^\ell}{\|g^\ell\|_* \|p^\ell\|}\Big)^2,$$

and both factors are separately bounded below by positive constants. Hence (21) holds, and Theorem 4.2 implies the final claim. □

For example, the **steepest descent directions**

$$p^\ell := -g^\ell$$

[3] If the norms are Euclidean, the ratio is the cosine of the angle between g^ℓ and p^ℓ.

satisfy the angle condition in the Euclidean norm with $\delta = 1$, hence lead to global convergence if used together with an efficient line search.

Enforcing the angle condition. In the following, $B \in \mathbb{R}^{n\times n}$ is a fixed but arbitrary symmetric, positive definite matrix, called the **preconditioner**. In practice, B is the identity matrix, a multiple of it, diagonal scaling matrix, or a matrix with the property that linear systems with coefficient matrix B are easy to compute. B is considered as a (more or less good) constant approximation of the Hessian matrix for the objective function. We may then apply the preceding with the ellipsoidal norms defined by (2). The **simplified Newton directions**

$$p^\ell := -B^{-1}g^\ell$$

satisfies the angle condition in the norms (2) with $\delta = 1$, hence leads to global convergence if used together with an efficient line search.

More generally, we may modify an arbitrary direction q by adding a multiple of the simplified Newton direction to get a direction

$$p := q - \lambda B^{-1}g \tag{32}$$

that satisfies the angle condition for a proper choice of the factor λ. Clearly it is enough to discuss the case $q \neq 0$. If

$$c := \frac{g^T q}{\sqrt{g^T B^{-1} g \cdot q^T B q}}$$

satisfies $c \leq -\delta$ we can take $\lambda = 0$ and $p := q$ satisfies the bounded angle condition. If this is not the case, we may use the following result.

5.2 Proposition. *Suppose that $g \neq 0$ and let $q \in \mathbb{R}^n \setminus \{0\}$, $0 < \delta < 1$. Put*

$$\pi_1 := g^T B^{-1} g > 0, \quad \pi_2 := q^T B q > 0, \quad \pi := g^T q. \tag{33}$$

Then

$$c = \frac{\pi}{\sqrt{\pi_1 \pi_2}} \in [-1, 1], \quad w := \frac{\pi_1 \pi_2 (1 - c^2)}{1 - \delta^2} \geq 0, \tag{34}$$

and (32) *satisfies the angle condition*

$$\frac{g^T p}{\sqrt{(g^T B^{-1} g)(p^T B p)}} = -\delta < 0 \tag{35}$$

when λ is chosen as

$$\lambda := \frac{\pi + \delta\sqrt{w}}{\pi_1}. \tag{36}$$

Proof. (34) follows directly from the generalized Cauchy–Schwarz inequality. In terms of the π_i, the angle condition (35) reads

$$\frac{\pi - \lambda\pi_1}{\sqrt{\pi_1(\pi_2 - 2\lambda\pi + \lambda^2\pi_1)}} = -\delta. \tag{37}$$

Squaring, multiplying with the denominator, and subtracting $\delta^2(\pi - \lambda\pi_1)^2$ gives

$$(1-\delta^2)(\pi-\lambda\pi_1)^2 = \delta^2\pi_1(\pi_2 - 2\lambda\pi + \lambda^2\pi_1) - \delta^2(\pi-\lambda\pi_1)^2 = \delta^2(\pi_1\pi_2 - \pi^2) = \delta^2\pi_1\pi_2(1-c^2).$$

Since $\pi - \lambda\pi_1$ is negative by (37), we need $\pi - \lambda\pi_1 = -\delta\sqrt{w}$, hence (36). By construction, this choice indeed satisfies (37) and hence (35). □

In finite precision arithmetic, rounding errors occasionally result in a value of $c^2 > 1$. Therefore one should compute w from

$$w := \frac{\pi_1\pi_2 \max\{\varepsilon, 1-c^2\}}{1-\delta^2},$$

where ε is the machine precision.

6 Zigzagging

The following examples show that unless the search paths are chosen with some care, convergence may be extremely slowed down by zigzagging.

(i) For the unconstrained problem

$$\begin{array}{ll} \min & f(x) = (x_1 - x_2)^2 + \varepsilon x_2^2 \\ \text{s.t.} & x \in \mathbb{R}^2 \end{array}$$

with small $\varepsilon > 0$, we have

$$g(x) = 2\begin{pmatrix} x_1 - x_2 \\ (1+\varepsilon)x_2 - x_1 \end{pmatrix}.$$

The Hessian matrix

$$G(x) = \begin{pmatrix} 2 & -2 \\ -2 & 2+2\varepsilon \end{pmatrix}$$

is constant and has condition number $\mathrm{cond}_\infty(G) = 4\varepsilon^{-1} + 4 + \varepsilon \to \infty$ for $\varepsilon \to 0$.

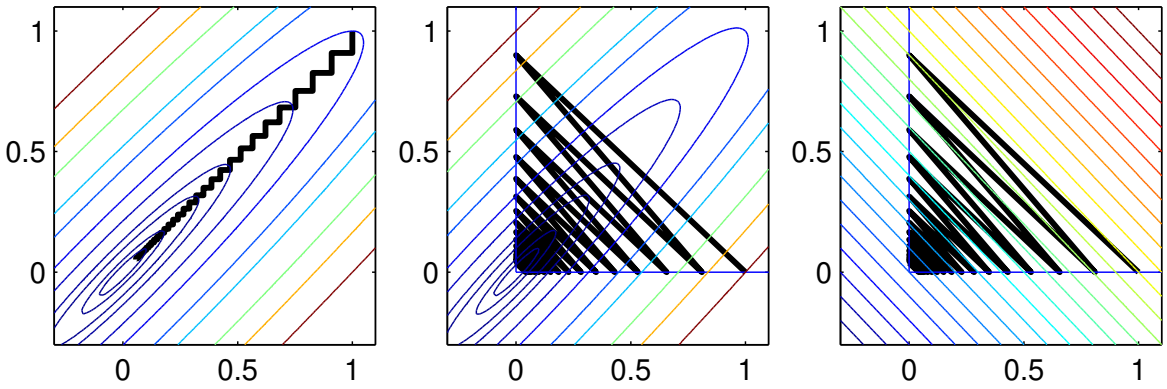
Starting with $x^0 = \binom{\xi}{\xi}$, we find for the steepest descent method ($p^\ell = -g^\ell$) with exact line searches the sequence

$$x^{2\ell} = \xi(1+\varepsilon)^{-\ell}\begin{pmatrix}1\\1\end{pmatrix}, \quad g^{2\ell} = 2\xi\varepsilon(1+\varepsilon)^{-\ell}\begin{pmatrix}0\\1\end{pmatrix},$$

$$x^{2\ell+1} = \xi(1+\varepsilon)^{-\ell-1}\begin{pmatrix}1+\varepsilon\\1\end{pmatrix}, \quad g^{2\ell+1} = 2\xi\varepsilon(1+\varepsilon)^{-\ell-1}\begin{pmatrix}1\\0\end{pmatrix},$$

with for $\varepsilon \to 0$ arbitrarily slow linear convergence to the solution at zero.

Figure 2: Inefficient zigzagging for convergence
(i) to an interior solution (left),
(ii) to an unconstrained minimizer in corner (middle), and
(iii) to a constrained minimizer in a corner (right).



Thus the global convergence of the steepest descent method does not rule out extremely slow convergence. (The same example also shows that extremely slow convergence is possible for another simple globally convergent method, namely **coordinate descent**, which uses as search directions the coordinate axes $\pm e^{(i)}$ in a cyclic (or more arbitrary) fashion.

(ii) For the bound constrained optimization problem

$$\begin{array}{ll} \min & \frac{1}{2}(x_1 - x_2)^2 + \varepsilon x_1 x_2 \\ \text{s.t.} & x_1 \geq 0, \quad x_2 \geq 0 \end{array}$$

with small $\varepsilon > 0$, we have

$$g(x) = 2 \begin{pmatrix} x_1 - (1-\varepsilon)x_2 \\ x_2 - (1-\varepsilon)x_1 \end{pmatrix}.$$

Started with $x^0 = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$, the search directions

$$p^{2\ell} = \begin{pmatrix} -1 \\ 1-\varepsilon \end{pmatrix}, \quad p^{2\ell+1} = \begin{pmatrix} 1-\varepsilon \\ -1 \end{pmatrix}$$

are scaled steepest descent directions, and may produce with an inexact line search the sequence

$$x^{2\ell} = \begin{pmatrix} (1-\varepsilon)^{2\ell} \\ 0 \end{pmatrix}, \quad x^{2\ell+1} = \begin{pmatrix} 0 \\ (1-\varepsilon)^{2\ell+1} \end{pmatrix},$$

with arbitrarily slow linear convergence to the solution at zero.

(iii) For the optimization of the bound constrained problem

$$\begin{array}{ll} \min & x_1 + x_2 \\ \text{s.t.} & x_1 \geq 0, \quad x_2 \geq 0, \end{array}$$

started with $x^0 = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$, the search directions

$$p^{2\ell} = \begin{pmatrix} -1 \\ 1-\varepsilon \end{pmatrix}, \quad p^{2\ell+1} = \begin{pmatrix} 1-\varepsilon \\ -1 \end{pmatrix}$$

($0 < \varepsilon < 1$ fixed) are descent directions and produce the sequence

$$x^{2\ell} = \begin{pmatrix} (1-\varepsilon)^{2\ell} \\ 0 \end{pmatrix}, \quad x^{2\ell+1} = \begin{pmatrix} 0 \\ (1-\varepsilon)^{2\ell+1} \end{pmatrix},$$

with arbitrarily slow linear convergence to the zero solution. Moreover,

$$g_{\rm red}(x^{2\ell}) = \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \quad g_{\rm red}(x^{2\ell+1}) = \begin{pmatrix} 0 \\ 1 \end{pmatrix},$$

so that $g_{\rm red}(x^\ell)$ does not converge to zero.

Thus zigzagging is a possible source of inefficiency. Good optimization methods should therefore be designed to eliminate zigzagging behavior as far as possible.

7 A nonlinear conjugate gradient method

The first example in Section 6 shows that the steepest descent method, which uses $p^\ell = -g^\ell$ as descent direction suffers from zigzagging although it trivially satisfies the angle condition.

Better methods attempt to reduce this effect. In order to avoid zigzagging we choose the search direction p as a vector satisfying $g^T p < 0$ that is closest to the previous search direction $p_{\rm old}$, with respect to the distance in the ellipsoidal norm (2) associated with a fixed symmetric and positive definite preconditioner B. Note that the results of this chapter are useful even when working with the 2-norm. This is the special case **without preconditioning**, where $B = I$.

In order that it is meaningful to compare two different search directions we note that for a sufficiently small step size α we obtain a gain in the function value of $f(x) - f(x + \alpha p) = -\alpha g^T p + o(\alpha)$. Hence the infinitesimal quality of a direction is fully characterized by $\nu := g^T p$. We therefore compare only directions with the same value of ν; this is no restriction of generality since we may rescale an arbitrary direction to match any given value of ν.

7.1 Theorem. *Among all $p \in \mathbb{R}^n$ with $g^T p = -\nu < 0$, the squared preconditioned distance $(p - p_{\rm old})^T B (p - p_{\rm old})$ becomes minimal for*

$$p = p_{\rm old} - \lambda B^{-1} g, \tag{38}$$

where

$$\lambda = \frac{\nu + g^T p_{\rm old}}{g^T B^{-1} g}. \tag{39}$$

Proof. This optimization problem can be solved using Lagrange multipliers. We have to find a stationary point of the Lagrange function

$$L(p) := \frac{1}{2}(p - p_{\text{old}})^T B(p - p_{\text{old}}) + \lambda g^T p,$$

giving the condition $B(p - p_{\text{old}}) + \lambda g = 0$, hence (38) holds. The Lagrange multiplier λ is determined from the constraint $g^T p = -\nu$, and yields (39). □

Note that Pardalos & Kovoor [15] show that for diagonal B, a bound-constrained version of the no-zigzag direction of Theorem 7.1 is computable in $O(n)$ steps.

For a search direction of the form

$$p^\ell = \rho_\ell p^{\ell-1} - \lambda_\ell B^{-1} g^\ell \tag{40}$$

we need

$$0 < \nu_\ell := -(g^\ell)^T p^\ell = -\rho_\ell (g^\ell)^T p^{\ell-1} + \lambda_\ell (g^\ell)^T B^{-1} g^\ell,$$

hence

$$\lambda_\ell = \frac{\nu_\ell + \rho_\ell (g^\ell)^T p^{\ell-1}}{(g^\ell)^T B^{-1} g^\ell}. \tag{41}$$

For $\rho_\ell = 1$, this agrees with the direction derived from the zigzagging avoiding argument; for $\rho_\ell = 0$, we get the simplified Newton direction. Thus search directions of the form (40) look like a flexible choice.

7.2 Theorem. *Suppose that* (40) *and* (41) *hold for all sufficiently large* ℓ *with*

$$\nu_\ell > 0, \quad |\rho_\ell| \le \frac{\nu_\ell}{\nu_{\ell-1}}.$$

Then

$$\frac{(p^\ell)^T B p^\ell}{\nu_\ell^2} - \frac{(p^{\ell-1})^T B p^{\ell-1}}{\nu_{\ell-1}^2} \le \frac{1}{(g^\ell)^T B^{-1} g^\ell}. \tag{42}$$

Moreover, if an efficient line search – e.g., Algorithm 3.3 – is used, we have

$$\inf_\ell \|g^\ell\|_* = 0 \quad \text{or} \quad \lim_{\ell\to\infty} f_\ell = -\infty.$$

Proof. We have

$$\begin{aligned}
(p^\ell)^T B p^\ell &= \rho_\ell^2 (p^{\ell-1})^T B p^{\ell-1} - 2\rho_\ell \lambda_\ell (g^\ell)^T p^{\ell-1} + \lambda_\ell^2 (g^\ell)^T B^{-1} g^\ell \\
&\le \frac{\nu_\ell^2}{\nu_{\ell-1}^2} (p^{\ell-1})^T B p^{\ell-1} + \frac{\nu_\ell^2 - \left(\rho_\ell (g^\ell)^T p^{\ell-1}\right)^2}{(g^\ell)^T B^{-1} g^\ell},
\end{aligned}$$

and (42) follows. In terms of the ellipsoidal norms (2), (42) reads

$$\frac{\|p^\ell\|^2}{\nu_\ell^2} - \frac{\|p^{\ell-1}\|^2}{\nu_{\ell-1}^2} \le \frac{1}{\|g^\ell\|_*^2}.$$

Since $s^\ell = \alpha_\ell p^\ell$ and $\sigma_\ell = \alpha_\ell \nu_\ell$, we find

$$\|g^\ell\|_*^2 \Big(\frac{\|s^\ell\|^2}{\sigma_\ell^2} - \frac{\|s^{\ell-1}\|^2}{\sigma_{\ell-1}^2} \Big) \le 1.$$

Therefore (18) holds. (17) holds since $\nu_\ell > 0$, and (19) is guaranteed by the line search. Hence Theorem 4.1 applies and proves the claim. □

7.3 Theorem. *Under the conditions of Theorem 7.2, suppose that an efficient line search is used and there are positive constants κ_1 and κ_2 such that, for all sufficiently large ℓ, either p^ℓ is parallel to the simplified Newton direction $-B^{-1}g^\ell$ or the conditions*

$$(g^\ell)^T B^{-1} g^\ell \le \kappa_1 (y^{\ell-1})^T B^{-1} y^{\ell-1}, \tag{43}$$

$$(y^{\ell-1})^T p^{\ell-1} \le \kappa_2 \nu_{\ell-1} \tag{44}$$

hold (where $y^{\ell-1} := g^\ell - g^{\ell-1}$). Then the bounded angle condition (31) *holds. In particular, if the x^ℓ converge to a strong local minimizer, convergence is locally linear.*

Proof. Under the asssumption of strong convergence to $\widehat{x}$, relations (24) and (26) from the proof of Theorem 4.2 apply for x, x' sufficiently close to $\widehat{x}$, and give

$$\underline{\gamma} \|g(x') - g(x)\|_* \, \|x' - x\| \le \overline{\gamma} (g(x') - g(x))^T (x' - x).$$

Substituting $x' = x^\ell$ and $x = x^{\ell-1}$ and using (43), we find after division by $\alpha_{\ell-1}$ that

$$\underline{\gamma} \|y^{\ell-1}\| \, \|p^{\ell-1}\| \le \overline{\gamma} (y^{\ell-1})^T p^{\ell-1} \le \overline{\gamma} \kappa_2 \nu_{\ell-1}$$

for all sufficiently large ℓ for which (43) and (44) hold. For these ℓ,

$$\begin{aligned} (g^\ell)^T B^{-1} g^\ell \cdot (p^{\ell-1})^T B p^{\ell-1} &\le \kappa_1 (y^{\ell-1})^T B^{-1} y^{\ell-1} \cdot (p^{\ell-1})^T B p^{\ell-1} \\ &\le \kappa_1 \|y^{\ell-1}\|^2 \|p^{\ell-1}\|^2 \\ &\le \kappa_1 \Big(\frac{\overline{\gamma} \kappa_2 \nu_{\ell-1}}{\underline{\gamma}} \Big)^2 = c \nu_{\ell-1}^2 \end{aligned}$$

for some constant $c > 0$. Now (42) implies

$$\frac{(p^\ell)^T B p^\ell}{\nu_\ell^2} \le \frac{(p^{\ell-1})^T B p^{\ell-1}}{\nu_{\ell-1}^2} + \frac{1}{(g^\ell)^T B^{-1} g^\ell} \le \frac{c+1}{(g^\ell)^T B^{-1} g^\ell}.$$

Thus

$$\frac{\nu_\ell^2}{(p^\ell)^T B p^\ell \cdot (g^\ell)^T B^{-1} g^\ell} \ge \frac{1}{c+1} \tag{45}$$

for sufficiently large ℓ satisfying (43) and (44). But if (43) or (44) are violated, p^ℓ is the simplified Newton direction, for which (45) holds trivially. Since $0 < \nu_\ell = -(g^\ell)^T p^\ell$, this shows that the left hand side of (31) is bounded away from zero. Hence Theorem 5.1 implies local linear convergence. □

7.4 Algorithm. Nonlinear conjugate gradient method (NCG)

Purpose: Finds local minimizer of $f(x)$ (or a stationary point only)
Input: x^0 (starting point), B (preconditioner)
Requirements: B symmetric and positive definite
Parameters: $\kappa_1 > 1$, $\kappa_2 > 1$

```
for ℓ = 0, 1, ...,
    g^ℓ = g(x^ℓ);
    ω_ℓ = (g^ℓ)^T B^{-1} g^ℓ;
    if ω_ℓ ≤ 0, stop; end; % x^ℓ stationary
    if ℓ = 0,
        restart=1;
    else
        ω' = (g^ℓ)^T B^{-1} g^{ℓ-1};
        restart1=( ω_ℓ > κ_1(ω_ℓ − 2ω' + ω_{ℓ-1}) );
        restart2=( |(g^ℓ)^T p^{ℓ-1} + ν| > κ_2 ν );
        restart = restart1 or restart2;
    end;
    if restart,
        ν = ω_ℓ;  p^ℓ = −B^{-1} g^ℓ;
    else
        λ_ℓ = (ν + (g^ℓ)^T p^{ℓ-1}) / ω_ℓ;  p^ℓ = p^{ℓ-1} − λ_ℓ B^{-1} g^ℓ;
    end;
    determine α_ℓ by Algorithm 3.3 with x(α) = x^ℓ + α p^ℓ;
    x^{ℓ+1} = x^ℓ + α_ℓ p^ℓ;
end;
```

Since we expect that the new search direction is not too different from the old one, f is expected to behave along the new search path like along the old one. The initial step size for each but the first line search may therefore be chosen as the accepted step size of the previous line search. To start the iteration we take $p_{\text{old}} = 0$. In order to guarantee local linear convergence, we may need to reset p_{old} to zero also at suitable later stages. We call this a **restart**; the precise restart conditions used come from Theorem 7.3. For $B \neq I$, i.e., if preconditioning is used, one should store $h^\ell := B^{-1} g^\ell$ in the computation of ω_ℓ, for later use in the computation of p^ℓ. Finally, note that, by Theorem 7.1, $\nu = -(G^\ell)^T p^\ell$

remains constant as long as no restart is made. The result is Algorithm 7.4. It is called a **nonlinear conjugate gradient method** since for a quadratic function f with positive definite Hessian, it is by Theorem 7.5 below equivalent to the preconditioned conjugate gradient method for solving positive definite linear systems of equations.

There are many other variants of nonlinear conjugate gradient methods. A thorough survey of nonlinear conjugate gradient methods was given by by Hager & Zhang [11]. In the literature, they are generally described in terms of search directions of the form

$$d^\ell = -g^\ell + \beta_{\ell-1} d^{\ell-1} \tag{46}$$

and corresponding updates

$$x^{\ell+1} = x^\ell + \gamma_\ell d^\ell.$$

Many formulas for the βs are in use; the step sizes γ_ℓ are typically determined by a Wolfe line search. Our formulas can be cast into this form if no preconditioning is used ($B = I$), by considering the scaled vectors

$$d^\ell := \lambda_\ell^{-1} p^\ell = \lambda_\ell^{-1} p^{\ell-1} - g^\ell = -g^\ell + \frac{\lambda_{\ell-1}}{\lambda_\ell} d^{\ell-1}$$

and the correspondence

$$\beta_{\ell-1} := \frac{\lambda_{\ell-1}}{\lambda_\ell}, \quad \gamma_\ell := \frac{\alpha_\ell}{\lambda_\ell}.$$

Thus as long as all λ_ℓ (or β_ℓ) are positive, the two choices of search directions are equivalent apart from the choice of the initial step sizes for the line search. The first nonlinear conjugate gradient method, introduced by Fletcher & Reeves [7], uses no preconditioning ($B = 1$) and (46) with

$$\beta_{\ell-1} := \frac{\omega_\ell}{\omega_{\ell-1}}. \tag{47}$$

For a quadratic function

$$f(x) = \gamma + c^T x + \frac{1}{2} x^T G x \tag{48}$$

with positive definite Hessian G, they showed the equivalence with the conjugate gradient method of Hestenes & Stiefel [12] for solving the linear system of equations $g(x) = c + Gx = 0$. The latter showed that their algorithm stops after at most n iterations with a solution of the linear system, hence with the minimizer of $f(x)$. If G is not positive definite, the algebra remains the same, except that it is now possible that a line search ends with a direction of infinite descent. Thus the method of Fletcher and Reeves stops for quadratic functions after at most n steps with a minimizer or with a direction of infinite descent. The case with a preconditioner is easily reduced to the case $B = I$ by means of a linear transformation of the vector x of variables; hence the same properties hold for any symmetric and positive definite B.

7.5 Theorem. *Applied to quadratic functions f, Algorithm 7.4 produces the same sequence of x^ℓ as the nonlinear conjugate gradient method by Fletcher and Reeves. In particular, Algorithm 7.4 stops for quadratic functions after at most n steps with a minimizer or with a direction of infinite descent.*

Proof. We have

$$p^\ell = p^{\ell-1} - \lambda_\ell B^{-1} g^\ell, \quad x^{\ell+1} = x^\ell + \alpha_\ell p^\ell.$$

For a quadratic function (48) we have $g^\ell = c + Gx^\ell$, hence

$$g^\ell - g^{\ell-1} = G(x^\ell - x^{\ell-1}) = \alpha_{\ell-1} G p^{\ell-1}.$$

For quadratic functions, the line search of Algorithm 3.3 becomes exact, hence

$$\alpha_\ell = \frac{-(g^\ell)^T p^\ell}{(p^\ell)^T G p^\ell} = \frac{\nu}{(p^\ell)^T G p^\ell}$$

when no restarts are made for $\ell > 0$. Now $\nu = -(g^{\ell-1})^T p^{\ell-1}$, hence

$$\lambda_\ell = \frac{(g^\ell - g^{\ell-1})^T p^{\ell-1}}{\omega_\ell} = \frac{\nu}{\omega_\ell} > 0, \quad \beta_{\ell-1} = \frac{\lambda_{\ell-1}}{\lambda_\ell} = \frac{\omega_\ell}{\omega_{\ell-1}}.$$

Since due to exact line search, the result of the algorithm is the same for an arbitrary rescaling of the search direction, we may rewrite the iteration in terms of the d^ℓ (discussed below) and get for $B = I$ equivalence with the Fletcher-Reeves conjugate gradient method.

The well-known conjugacy properties

$$(g^\ell)^T p^{\ell-1} = (g^\ell)^T B^{-1} g^{\ell-1} = 0$$

of the linear conjugate gradient method (Hestenes & Stiefel [12]) imply that given the restrictions $\kappa_1 > 1$ and $\kappa_2 > 1$, indeed no restarts will be made. □

Since locally all twice continuously differentiable functions are well approximated by a quadratic, the final remark in the proof also holds locally for general C^2-functions with the line search from Algorithm 3.3. In particular, Algorithm 7.4 shares close to a strong local minimizer the excellent local convergence behavior (see, e.g., Axelsson & Lindskog [1]) of the preconditioned linear conjugate gradient method.

Finally, we have the following global convergence result.

7.6 Theorem. *The points x^ℓ produced by the nonlinear conjugate gradient method of Algorithm 7.4 satisfy*

$$\inf_\ell \|g(x^\ell)\|_* = 0 \quad \textit{or} \quad \lim_{\ell\to\infty} f(x^\ell) = -\infty,$$

and in case of convergence to a strong local minimizer, the convergence is locally linear.

Proof. Using Theorem 7.2, it is easy to see that the assumptions of Theorem 4.1 and Theorem 7.3 are satisfied. □

We have seen that Algorithm 7.4 enjoys many good theoretical properties and is guaranteed to perform well on general smooth functions. When a good starting point is available, no

restarts are made. Far away from a minimizer, however, a strong deviation from quadratic behavior may cause a restart. In particular, whenever very little progress is made while the gradient is still large, $g^\ell \approx g^{\ell-1}$, hence $y^{\ell-1} \approx 0$, and a restart is made. Thus jamming, a problem for the standard implementation of the nonlinear conjugate gradient method by Fletcher & Reeves [7] is not possible.

The algorithm can be implemented using very little storage only: Apart from what is needed for a Cholesky factor of the preconditioner, 4 vectors of storage (for $x, g, h = B^{-1}g, p$), and without preconditioning ($B = I$) even 3 vectors, suffice.

Compared to other nonlinear conjugate gradient methods, Algorithm 7.4 has two advantages:
(i) Because of the no zigzagging property, there is a simple rule for a good initial step size.
(ii) Since the λ_ℓ need not be positive, the restrictions on the line search is just the minimal requirement of efficiency.
On the other hand, the traditional nonlinear conjugate gradient methods need stronger assumptions on the line search for good performance; given only an efficient line search, many of them do not even always lead to descent directions. As a consequence, the convergence analysis is usually more complicated. They also do not have an optimality property with respect to zigzagging.

8 Optimality conditions for bound constraints

Given a feasible point x and an index i, we call the bound $\underline{x}_i$ or $\overline{x}_i$ **active** if $x_i = \underline{x}_i$ or $x_i = \overline{x}_i$, respectively. In both cases, we also call the index i and the component x_i **active**. Otherwise, i.e., if $x_i \in]\underline{x}_i, \overline{x}_i[$, the index i, the component x_i, and the bounds $\underline{x}_i$ and $\overline{x}_i$ are called **nonactive** or **free**. A **corner** of the box $\mathbf{x}$ is a point all of whose components are active. If the gradient $g = g(x)$ has a nonzero component g_i at a nonactive index i, we may change x_i slightly without leaving the feasible region. The value of the objective function is reduced by moving slightly to smaller or larger values depending on whether $g_i > 0$ or $g_i < 0$, respectively. However, if x_i is active, only changes of x_i in one direction are possible without losing feasibility. The value of the objective function can then possibly be reduced by moving slightly in the feasible direction only when

$$\begin{cases} g_i \le 0 & \text{if } x_i = \underline{x}_i, \\ g_i \ge 0 & \text{if } x_i = \overline{x}_i. \end{cases} \tag{49}$$

But a decrease is guaranteed only if the slightly stronger condition

$$\begin{cases} g_i < 0 & \text{if } x_i = \underline{x}_i, \\ g_i > 0 & \text{if } x_i = \overline{x}_i \end{cases} \tag{50}$$

holds.

8.1 Theorem. *(**optimality conditions for bound-constrained optimization**)*
(i) ***First order necessary conditions.*** *At any local minimizer x of* (4)*, the* ***reduced***

gradient $g_{\rm red}(x)$ *at* x*, with components*

$$g_{\rm red}(x)_i := \begin{cases} 0 & \text{if } x_i = \underline{x}_i = \overline{x}_i, \\ \min(0, g_i(x)) & \text{if } x_i = \underline{x}_i < \overline{x}_i, \\ \max(0, g_i(x)) & \text{if } x_i = \overline{x}_i > \underline{x}_i, \\ g_i(x) & \text{otherwise,} \end{cases} \tag{51}$$

vanishes.

(ii) ***First order sufficient conditions.*** *Every corner* x *of* $\mathbf{x}$ *such that* $g_i(x) > 0$ *at all active lower bounds and* $g_i(x) < 0$ *at all active upper bounds is a local minimizer of* (4)*.*

Proof. (i) Combining the various cases discussed above, we see that a decrease is always possible if the reduced gradient has a nonzero component.

(ii) In this case, any feasible point $x + \alpha p \neq x$ $(\alpha > 0)$ must have $p_i \geq 0$ if $\underline{x}_i$ is active, $p_i \leq 0$ if $\overline{x}_i$ is active, and at least one p_i is nonzero. Therefore

$$g(x)^T p = \sum_i g_i p_i > 0.$$

This implies that $f(x + \alpha p) - f(x) = \alpha g(x)^T p + o(\alpha) > 0$ for small $\alpha > 0$, hence $f(x)$ is locally minimal. □

If no bound is active, $g_{\rm red}(x) = g(x)$ and (i) reduces to the condition that x is a stationary point of the function f. In generalization of this, we call a feasible point x with $g_{\rm red}(x) = 0$ a **stationary point** of the optimization problem (4). By the above, a local minimizer x of (4) must be a stationary point of this problem. This statement is a concise expression of the first order optimality conditions.

Note that the reduced gradient need not be continuous – it may change abruptly when a bound becomes active. A simple example is

$$f(x) = x, \quad \mathbf{x} = [0, \infty], \tag{52}$$

where $g_{\rm red}(x) = 1$ for $x > 0$ but $g_{\rm red}(0) = 0$. It is therefore important that a weaker continuity statement still holds, expressed in the first part of the following theorem.

8.2 Theorem. *If the sequence* x^ℓ *converges to* $\widehat{x}$ *and* $\lim g_{\rm red}(x^\ell) = 0$ *then* $g_{\rm red}(\widehat{x}) = 0$*. Moreover, for every index* $i = 1, \ldots, n$*,*

$$g_i(\widehat{x}) > 0 \quad \Rightarrow \quad x_i^\ell = \widehat{x}_i = \underline{x}_i \text{ for sufficiently large } \ell, \tag{53}$$

$$g_i(\widehat{x}) < 0 \quad \Rightarrow \quad x_i^\ell = \widehat{x}_i = \overline{x}_i \text{ for sufficiently large } \ell. \tag{54}$$

Proof. Every free index i of $\widehat{x}$ is also free for x^ℓ with sufficiently large ℓ. Since f is continously differentiable, we conclude that $g_i(\widehat{x}) = \lim g_i(x^\ell) = 0$ for all free i. If $\widehat{x}_i = \underline{x}_i$

then the x_i^ℓ converge to $\underline{x}_i$, hence satisfy $x_i^\ell < \overline{x}_i$; thus $g_i(\widehat{x}) = \lim g_i(x^\ell) \geq 0$ for these i. Similarly, one sees that $g_i(\widehat{x}) \leq 0$ if $\widehat{x}_i = \overline{x}_i$. Together, this implies $g_{\rm red}(\widehat{x}) = 0$.

Now let i be an index i for which $g_i(\widehat{x}) > 0$. From (51) and $g_{\rm red}(\widehat{x}) = 0$ we conclude that $\widehat{x}_i = \underline{x}_i < \overline{x}_i$. The definition (51) of the reduced gradient implies that for sufficiently large ℓ,

$$g_{\rm red}(x^\ell)_i = \begin{cases} 0 & \text{if } x_i^\ell = \underline{x}_i, \\ g_i(x^\ell) & \text{otherwise.} \end{cases}$$

Now $g_{\rm red}(x^\ell)$ converges to zero, but by continuity of the gradient, $g_i(x^\ell) \to g_i(\widehat{x}) > 0$. Hence the second case is impossible for large ℓ. Therefore $x_i^\ell = \underline{x}_i$ for all large ℓ, and (53) holds for sufficiently large k.

Similarly, if i is an index for which $g_i(\widehat{x}) < 0$ then (54) holds for sufficiently large l. □

We say that the active variable x_i is **strongly active** if

$$\begin{cases} g_i > 0 & \text{if } x_i = \underline{x}_i, \\ g_i < 0 & \text{if } x_i = \overline{x}_i. \end{cases} \tag{55}$$

Thus slightly changing a single strongly active variable only cannot lead to a better feasible point. A stationary point is called **degenerate** if $g_i(x) = 0$ for some active index i, and **nondegenerate** otherwise, i.e., if all its active bounds are strongly active. This allows us to rephrase Theorem 8.2 as saying that *all strongly active variables are ultimately fixed* when the sequence x^ℓ converges and $\lim g_{\rm red}(x^\ell) = 0$.

In particular, in case of convergence to a nondegenerate stationary point, zigzagging through changes of the active set (as in the examples of Section 6) cannot occur infinitely often.

8.3 Corollary. *If the $x^\ell \in \mathbf{x}$ form a bounded sequence such that $\inf_\ell g_{\rm red}(x^\ell) = 0$ then some subsequence converges to a point $\widehat{x} \in \mathbf{x}$ satisfying $g_{\rm red}(\widehat{x}) = 0$.*

Proof. By assumption, there is a subsequence on which $g_{\rm red}(x^\ell) \to 0$. Boundedness implies that this subsequence has a convergent subsequence, and by Theorem 8.2, its limit $\widehat{x}$ satisfies the claim. □

The corollary justifies to accept a numerical approximation x to a stationary point $\widehat{x}$ as soon as a stopping test of the form

$$\|g_{\rm red}(x)\|_* \leq \varepsilon \tag{56}$$

holds for some fixed ε. In this stopping test, one traditionally uses for $\|\cdot\|_*$ the maximum norm, with $\varepsilon = 10^{-5}$ or $\varepsilon = 10^{-6}$. In a conceptual analysis of algorithms, however, one has no stopping test and investigates the behavior of an infinite number of approximations x^ℓ, with the goal of showing that the $g_{\rm red}(x^\ell)$, or at least a subsequence of them, converge to zero. This implies (at least in exact arithmetic) finite termination if the stopping test (56) is added to the algorithm.

We define the **feasible projection** $\pi[x]$ of an arbitrary point $x \in \mathbb{R}^n$ to the (fixed) box $\mathbf{x}$ by

$$\pi[x]_i := \max(\underline{x}_i, \min(x_i, \overline{x}_i)) = \begin{cases} \underline{x}_i & \text{if } x_i \leq \underline{x}_i, \\ \overline{x}_i & \text{if } x_i \geq \overline{x}_i, \\ x_i & \text{otherwise.} \end{cases} \tag{57}$$

Clearly $\pi[x] \in \mathbf{x}$, and $\pi[x] = x$ iff $x \in \mathbf{x}$. It is easy to see that (for any fixed $\alpha > 0$) the first order optimality conditions may also be written in the equivalent form

$$g^{(\alpha)}(x) := \pi[x - \alpha g(x)] - x = 0.$$

$g^{(\alpha)}(x)$ is continuous in x for any α. In contrast, Example (ii) of Section 6 showed that convergence to a stationary point $\widehat{x}$ is possible even when $\inf_\ell g_{\rm red}(x^\ell) > 0$, reflecting the lack of continuity of the reduced gradient. Such a counterintuitive situation means that infinitely many x^ℓ have activities different from the limiting $\widehat{x}$.

Traditional bound constrained solvers therefore aim only at a slightly weaker convergence statement that the **projected gradient** $g^{(1)}(x^\ell)$ has a subsequence converging to zero. The resulting simpler convergence analysis [2, 4, 6] is probably the reason why usually, e.g., in LBFGS-B [3] and in Hager & Zhang [10], a different stopping criterion of the form $\|g^{(1)}(x)\|_\infty \leq \delta$ is used in place of (56). For example, for (52), $x = \delta$ satisfies this criterion although $g_{\rm red}(x) = 1$. However, in finite precision arithmetic, this stopping criterion may accept very poor points as sufficiently stationary. For example, for (52), $x = 10^{17}$ also satisfies this criterion although x is extremely far from a stationary point! In this particular case, the reason is that, in double precision arithmetic, $g^{(1)}(x)$ numerically becomes identically zero due to severe cancellation of digits in the subtraction.

9 The bent search path

For solving the bound constrained optimization problem (4), a line search along a ray may lead to infeasible points. The most natural remedy, first suggested by Bertsekas [2], is to project the ray into the box. Thus we do each line search along a **bent search path**

$$x(\alpha) := \pi[x + \alpha q], \tag{58}$$

obtained by taking the ray $x + \alpha q$ ($\alpha \geq 0$) from the current point x into a direction $q \neq 0$ and projecting it into the feasible set using the projection (57). The bent search path is piecewise linear, with breakpoints at the $\alpha > 0$ with $x_i + \alpha q_i \in \{\underline{x}_i, \overline{x}_i\}$. Thus the breakpoints are the elements of the set

$$S := \left\{ \frac{\overline{x}_i - x_i}{q_i} \mid q_i > 0 \right\} \cup \left\{ \frac{\underline{x}_i - x_i}{q_i} \mid q_i < 0 \right\} \setminus \{0, \infty\}.$$

If the breakpoints $\alpha_1, \ldots, \alpha_m$ are ordered such that

$$0 = \alpha_0 < \alpha_1 < \ldots < \alpha_m < \alpha_{m+1} = \infty,$$

the bent search path is linear on each interval $[\alpha_{i-1}, \alpha_i]$ $(i = 1, ..., m+1)$. Note that when for some $\alpha > 0$, $x(\alpha)$ is a corner of the box then this corner is $x(\alpha_m)$ for some $m > 0$, and $x(\alpha)$ stays constant for all $\alpha \geq \alpha_m$.

In an **active set algorithm** for bound-constrained optimization, each iteration changes only a subset of the variables. To account for this we use a **working set** $I \subseteq \{1, \ldots, n\}$ satisfying

$$q_i = 0 \quad \text{for } i \notin I, \tag{59}$$

and denote by q_I the subvector of q indexed by I. We write

$$g = g(x), \quad g_{\rm red} = g_{\rm red}(x).$$

In order to ensure local linear convergence when the working set I stays constant we require the angle condition

$$\frac{g_I^T q_I}{\|g_I\|_* \|q_I\|} \leq -\delta < 0. \tag{60}$$

Sensible choices for the working set I include the set

$$I_-(x) := \{i \mid \underline{x}_i < x_i < \overline{x}_i\} \tag{61}$$

of free indices of x, or the set

$$\begin{aligned} I_+(x) &:= I_-(x) \cup \{i \mid (g_{\rm red})_i \neq 0\} \\ &= \{i \mid \underline{x}_i < x_i < \overline{x}_i \text{ or } \underline{x}_i = x_i < \overline{x}_i, g_i < 0 \text{ or } \underline{x}_i < x_i = \overline{x}_i, g_i > 0\}. \end{aligned} \tag{62}$$

By definition of the reduced gradient,

$$\|g_I(x)\|_* = \|g_{\rm red}(x)\|_* \quad \text{if } I = I_+(x). \tag{63}$$

To prove (in Section 11 below) the global convergence of a descent algorithm for bound constrained optimization we need at one critical place the condition

$$g_i(x) q_i \leq 0 \quad \text{for all } i \quad \text{if } I = I_+(x) \neq I_-(x). \tag{64}$$

Note that (59), (60), and (64) are satisfied with arbitrary I by directions of the form

$$q_i = \begin{cases} -g_i/d_i & \text{if } i \in I, \\ 0 & \text{otherwise,} \end{cases} \tag{65}$$

with positive elements d_i in a fixed interval $[\underline{d}, \overline{d}]$, where $0 < \underline{d} < \overline{d} < \infty$.

To find conditions that eliminate most of a major cause of inefficiency, namely the zigzagging behaviour, we reconsider the examples of Section 6. The first example does not depend on constraints and must be handled by the choice of the search direction in iterations where the working set remains fixed. By Theorem 7.1, a nonlinear conjugate gradient method reduces the zigzagging effect and eliminates all difficulties in this kind of examples. Indeed, by Theorem 7.6, the nonlinear conjugate gradient method terminates on n-dimensional quadratic problems in at most n iterations, and hence is fast near a strong minimizer where

the objective function is almost quadratic. One only needs to adapt the method to work on the subspace determined by the working set.

The second and third example of Section 6 show that algorithms unable to quickly identify the set of optimal active bound constraints may free and fix the same subset of variables alternatingly in a large number of successive iterations. To handle these example we control the conditions under which variables enter the working set I.

Using always $I = I_+(x)$ seems to be a good choice since it most quickly corrects a poor active set. However, in the second example of Section 6, $I_\ell = I_+(x^\ell) = \{1, 2\}$ and (64) holds; so the choice $I = I_+(x)$ is not always adequate. In this example, the alternative choice $I = I_-(x)$ is adequate; it forbids the zigzagging directions since $I(x^\ell)$ has size one. In the third example of Section 6, $I_\ell = \{1, 2\}$ while $I(x^\ell) = I_+(x^\ell)$ has size one. Therefore both (in this example identical) choices $I = I_-(x)$ or $I = I_+(x)$ forbid the zigzagging directions. However, we cannot always choose $I = I_-(x)$ since this might even be the empty set! Closer inspection reveals that we need to ensure that shrinking the gradient in the components indexed by I shrinks the reduced gradient at least asymptotically. We therefore require (in an arbitrary monotone norm) the condition

$$\|g_I\|_*^2 \geq \rho \|g_{\text{red}}\|_*^2 \tag{66}$$

for some $\rho \in]0, 1]$. This condition says that the components of the reduced gradients missed by restricting to I are bounded by a multiple of $\|g_I\|_*$.

Our examples indicate that an appropriate alternation between the choices $I = I_+(x)$ and $I = I_-(x)$ could eliminate zigzagging. (63) implies that (66) always holds when $I = I_+(x)$; thus (66) only restricts the situations in which $I = I_-(x)$ is allowed. We must ensure that this is the case in the second example of Section 6. Indeed, there this choice is possible without violating (66) if the 2-norm is used and

$$\rho < 1/n.$$

The factor n is needed in order to also eliminate related n-dimensional examples of zigzagging with $n-1$ freeable bounds. For the maximum norm $\|\cdot\|_*$, corresponding to the 1-norm $\|\cdot\|$, any positive $\rho < 1$ would suffice to eliminate these examples.

Taking into account these insights we propose the following algorithmic scheme, for which convergence and strong limitations on the possible forms of zigzagging will be proved.

9.1 Algorithm. (**BOPT**, **bound constrained optimization**)
Purpose: Minimizes a smooth $f(x)$ subject to $x \in \mathbf{x} = [\underline{x}, \overline{x}]$
Input: $x^0 \in \mathbb{R}^n$ (starting point)
Parameters: $\beta \in]0, \frac{1}{4}[$, $q > 1$ (line search parameters)
$0 < \delta < 1$ (reduced angle parameters)
$0 < \rho < 1/n$ (factor safeguarding (66))
and parameters specifying a pair of monotone dual norms

$x = x^0$; $I = I_+(x)$; `freeing`=0;
while $g_{\rm red}(x) \neq 0$,
  choose q with (59), (60), and (64);
  do a line search (Algorithm 3.3) along the bent search path (58)
  update x and $I = I_-(x)$;
  `freeing`=((66) fails);
  if `freeing`, update $I = I_+(x)$; end;
end;

By (63) and the stopping test, we have

$$\|g_I(x)\|_* = \|g_{\rm red}(x)\|_* > 0 \quad \text{if } I = I_+(x). \tag{67}$$

In particular if (66) fails for $I = I_-(x)$, the resetting of I ensures that (66) holds for the I used in the next iteration. Therefore (66) holds at every iteration except possibly the first.

Note that $i \in I_+(x) \setminus I_-(x)$ iff i is active and (50) holds. In this case we call the index i **freeable** and say that the variable x_i can be **freed** from its bound. Indeed, (50) says for an active index i that the ith components of the reduced gradient is nonzero, so that the function value decreases when moving the corresponding components x_i into the interior. We call any iteration where

$$I = I_+(x) \neq I_-(x)$$

a **freeing iteration**, since this is the condition that at least one bound is freed. In a freeing iteration one typically uses a search direction of the form (65), which guarantees the conditions required in the algorithm. In a non-freeing iteration, (64) is not a restriction, and one typically uses a search direction appropriate for an unconstrained method in the subspace defined by I, which, once the optimal activities are identified, leads to faster local convergence. For example, we may use a conjugate gradient method in the subspace; after any change of I, the subspace changes, hence the conjugate gradient method must be restarted.

10 Some auxiliary results

We now prove a few technical results that are needed for our convergence proof in the next section.

10.1 Proposition. *For nonzero q and $\alpha > 0$,*

$$p_q(\alpha) := \frac{\pi[x + \alpha q] - x}{\alpha}$$

satisfies (in any monotone norm)

$$|p_q(\alpha)| \leq |q|, \quad \|p_q(\alpha)\| \leq \|q\|,$$

and with $p \in \mathbb{R}^n$ defined by

$$p_i := \begin{cases} 0 & \text{if } \underline{x}_i = x_i = \overline{x}_i, \\ \max(0, q_i) & \text{if } \underline{x}_i = x_i < \overline{x}_i, \\ \min(q_i, 0) & \text{if } \underline{x}_i < x_i = \overline{x}_i, \\ q_i & \text{if } \underline{x}_i < x_i < \overline{x}_i, \end{cases} \tag{68}$$

we have

$$p_q(\alpha) = p \quad \text{for sufficiently small } \alpha > 0, \tag{69}$$

Proof. Rescaling α and q if necessary, we may assume that $\|q\| = 1$. Then

$$\begin{aligned} \alpha p_q &= \pi[x + \alpha q] - x = \sup\Big(\underline{x}, \inf(x + \alpha q, \overline{x})\Big) - x \\ &= \sup\Big(\underline{x} - x, \inf(\alpha q, \overline{x} - x)\Big), \end{aligned}$$

hence

$$|p_q| = \Big|\sup\Big(\frac{\underline{x} - x}{\alpha}, \inf\Big(q, \frac{\overline{x} - x}{\alpha}\Big)\Big)\Big| \le |q|,$$
$$\|p_q\| \le \|q\| = 1.$$

The term $\dfrac{\underline{x}_i - x_i}{\alpha}$ vanishes if $x_i = \underline{x}_i$, and it becomes arbitrarily large negative if $x_i > \underline{x}_i$ and α is sufficiently small. Similarly, $\dfrac{\overline{x}_i - x_i}{\alpha}$ vanishes if $x_i = \overline{x}_i$, and becomes arbitrarily large positive if $x_i > \underline{x}_i$ and α is sufficiently small. Evaluating componentwise the sup and inf therefore results in (68). □

10.2 Proposition. *If the index set $I \subseteq I_+(x)$ satisfies* (59) *and* (60) *then*

$$g^T q = g_I^T q_I < 0, \quad \|q\| = \|q_I\|. \tag{70}$$

If, in addition,

$$g_i(x) q_i \le 0 \quad \text{for all } i \tag{71}$$

then

$$p = q, \tag{72}$$

and we have $f(x(\alpha)) < f(x)$ for sufficiently small $\alpha > 0$.

Proof. By (59) and (60),

$$g^T q = \sum_i g_i q_i = \sum_{i \in I} g_i q_i = g_I^T q_I < 0,$$

giving (70). Since $I \subseteq I_+(x)$, any active i satisfies one of the conditions in (71). Thus if (71) holds then (68) gives (72). The piecewise linear structure of the search path now gives $x(\alpha) = x + \alpha p_q(\alpha) = x + \alpha q$ for all sufficiently small $\alpha > 0$, and therefore

$$f(x(\alpha)) = f(x + \alpha q) = f(x) + \alpha g^T q + o(\alpha) = f(x) + \alpha(g^T q + o(1)) < f(x)$$

for sufficiently small $\alpha > 0$. □

10.3 Proposition. *Suppose that*

$$g_i(x^\ell)q_i^\ell \le 0 \quad \text{for } i \in I_+(x^\ell), \tag{73}$$

$$q_i^\ell = 0 \quad \text{for } i \notin I_+(x^\ell).$$

If

$$\lim_{\ell\to\infty} x^\ell = x, \quad \lim_{\ell\to\infty} \alpha_\ell = 0, \quad \lim_{\ell\to\infty} q^\ell = q$$

then

$$r^\ell := \frac{\pi[x^\ell + \alpha_\ell q^\ell] - x^\ell}{\alpha_\ell \|q^\ell\|}$$

satisfies

$$\lim_{\ell\to\infty} r^\ell = q.$$

Proof. We first simplify the assumptions by replacing q^ℓ with $q^\ell/\|q^\ell\|$ and α_ℓ with $\alpha_\ell\|q^\ell\|$. Then the assumptions on the q^ℓ and α_ℓ take the form

$$\|q^\ell\| = 1, \quad \alpha_\ell > 0 \quad \text{for all } \ell,$$

$$\lim_{\ell\to\infty} q^\ell = q, \quad \lim_{\ell\to\infty} \alpha_\ell = 0,$$

$$r^\ell = \frac{\pi[x^\ell + \alpha_\ell q^\ell] - x^\ell}{\alpha_\ell}.$$

By Proposition 10.1, $|r^\ell| \le |q^\ell|$, and by assumption,

$$r_i^\ell = q_i^\ell = 0 \quad \text{for } i \notin I_+(x^\ell).$$

Since the q^ℓ are bounded and $\alpha_\ell \to 0$, Proposition 10.1 also implies that for sufficiently large ℓ,

$$r_i^\ell = \begin{cases} 0 & \text{if } \underline{x}_i = x_i^\ell = \overline{x}_i, \\ \max(0, q_i^\ell) & \text{if } \underline{x}_i = x_i^\ell < \underline{x}_i, \\ \min(q_i^\ell, 0) & \text{if } \underline{x}_i < x_i^\ell = \overline{x}_i, \\ q_i^\ell & \text{if } \underline{x}_i < x_i^\ell < \overline{x}_i. \end{cases}$$

In view of (73), this implies $r_i^\ell = q_i^\ell$ for $i \in I_+(x^\ell)$ and sufficiently large ℓ. Taking the limit, we find $r^\ell \to q$, as claimed. □

11 Convergence – bound constrained case

11.1 Theorem. *Let f be continuously differentiable, with Lipschitz continuous gradient g. Let x^ℓ denote the value of x in Algorithm 9.1 after its ℓth update. Then one of the following three cases holds:*

(i) The iteration stops after finitely many steps at a stationary point.

(ii) We have

$$\lim_{\ell\to\infty} f(x^\ell) = \widehat{f} \in \mathbb{R}, \quad \inf_{\ell\geq 0} \|g_{\mathrm{red}}(x^\ell)\|_* = 0.$$

Some limit point $\widehat{x}$ of the x^ℓ satisfies $f(\widehat{x}) = \widehat{f} \leq f(x^0)$ and $g_{\mathrm{red}}(\widehat{x}) = 0$.

(iii) $\sup_{\ell\geq 0} \|x^\ell\| = \infty$.

Proof. If the algorithm stops after finitely many steps, the stopping condition implies that we have a stationary point; hence (i) holds. Thus we may assume that infinitely many iterations.

For the point x, the working set I, the direction q, and the tangent direction p given by (68) at iteration ℓ before updating I, we write x^ℓ, I_ℓ, q^ℓ, and p^ℓ, respectively. Since function values decrease monotonically by construction, the infimum $\widehat{f}$ of the $f(x^\ell)$ is finite, and we have

$$\lim_{\ell\to\infty} f(x^\ell) = \widehat{f}. \tag{74}$$

For any index set I, we consider the set L_I of indices ℓ satisfying

$$I = I_\ell = I_+(x^\ell) \neq I(x^\ell)$$

and distinguish two cases, depending on the amount of zigzagging.

Case 1 (limited zigzagging): All L_I are finite. Since every ℓ for which the ℓth iteration is freeing belongs to some L_I and there are only finitely many possibilities for I, the number of freeing iterations is finite. Thus there is a number N_f such that no iteration with index $\ell > N_f$ is freeing. Algorithm 9.1 and (66) imply that $I_\ell = I(x^\ell)$ for $\ell > N_f$. Therefore a line search in iteration $\ell > N_f$ never frees an already active bound; hence bounds can only be fixed. This can happen only finitely many times; so there is an N such that $I(x^\ell)$ remains fixed for all $\ell > N$,

$$I_\ell = I(x^\ell) = I \quad \text{for } \ell > N, \tag{75}$$

and no bound is fixed for $\ell > N$. Therefore the line search accepts a point on the initial ray of the bent search path, where $p(\alpha) = p$ is given by (68). By (69), Theorem 3.1 implies that there is a number $\delta' > 0$ such that

$$\frac{(f(x^\ell) - f(x^{\ell+1}))\|p^\ell\|^2}{(g(x^\ell)^T p^\ell)^2} \geq \delta' \quad \text{for all } \ell > N.$$

(59) and (60) hold by the specification of Algorithm 9.1, and (70) follows by Proposition 10.2. Using (60), (70), and (66) (the latter holds by the remark after Algorithm 9.1), we find that for all $\ell > N$,

$$\begin{aligned} f(x^\ell) - f(x^{\ell+1}) &\geq \delta'\Big(\frac{g(x^\ell)^T p^\ell_{I_\ell}}{\|p^\ell_{I_\ell}\|}\Big)^2 = \delta'\Big(\frac{g_{I_\ell}(x^\ell)^T p^\ell_{I_\ell}}{\|p^\ell_{I_\ell}\|}\Big)^2 \\ &\geq \delta'\Big(\delta\|g_{I_\ell}(x^\ell)\|_*\Big)^2 \geq \delta'\Big(\delta\rho\|g_{\mathrm{red}}(x^\ell)\|_*\Big)^2 \geq \Delta := \delta'(\delta\rho\gamma^*)^2, \end{aligned}$$

where

$$\gamma^* := \inf_{\ell \geq 0} \|g_{\text{red}}(x^\ell)\|_*. \tag{76}$$

For $\ell \to \infty$, (74) implies that the left hand side tends to zero, hence $\Delta = 0$ and therefore $\gamma^* = 0$. Thus there is a subsequence x^{ℓ_k} with $\|g_{\text{red}}(x^\ell)\|_* \to \gamma^* = 0$, and since the x^ℓ are bounded, we may assume (by deleting part of the subsequence) that the subsequence converges, $x^{\ell_k} \to \widehat{x}$. Now Theorem 8.2 implies that $g_{\text{red}}(\widehat{x}) = 0$. Thus (ii) holds.

Case 2 (unlimited zigzagging): Some L_I is an infinite set. Handling this case requires a detailed look at what happens at the bounds. Since all conditions used in Algorithm 9.1 and Algorithm 3.3 are invariant under appropriate scaling we may assume w.l.o.g. that all directions q^ℓ are scaled such that

$$\|q^\ell\| = 1. \tag{77}$$

According to Algorithm 9.1, (60) and (64) hold. (64) and Proposition 10.2 imply (70) and

$$p^\ell = q^\ell \quad \text{for } \ell \in L_I. \tag{78}$$

If the x^ℓ are unbounded, (iii) holds and we are done. Otherwise the set of tuples $[x^\ell, q^\ell]$ is bounded. Thus there is an infinite sequence $\ell_k \in L_I$ $(k = 1, 2, \ldots)$ such that, for $k \to \infty$,

$$x^{\ell_k} \to \widehat{x}, \quad q^{\ell_k} \to q.$$

Using (74), we find $f(\widehat{x}) = \widehat{f}$, and we have

$$\|q\| = 1, \quad q_i = 0 \quad \text{for } i \notin I. \tag{79}$$

Taking limits in (59), and in (70), (60) gives $\|q_I\| = \|q\| = 1$ and $g^T q = g_I^T q_I \leq -\delta \|g_I\|_*$.

Assume for the moment that

$$g_I \neq 0. \tag{80}$$

Then we conclude that

$$g^T q < 0. \tag{81}$$

We write α_ℓ, and μ_ℓ for the step size α chosen by the line search and the Goldstein quotient at iteration ℓ,

$$\mu_\ell := \mu(\alpha_\ell) = \frac{f(x^{\ell+1}) - f(x^\ell)}{\alpha_\ell g(x^\ell)^T p^\ell}. \tag{82}$$

μ_ℓ is bounded away from zero by the line search since the accepted step size satisfies the descent condition (9).

Since $g(x^{\ell_k})^T p^{\ell_k} \to g^T q \neq 0$ by (81), we find from (78) that

$$\alpha_{\ell_k} = \frac{f(x^{\ell_k+1}) - f(x^{\ell_k})}{\mu_{\ell_k} g(x^{\ell_k})^T q^{\ell_k}} \to 0 \quad \text{for } k \to \infty. \tag{83}$$

Now Proposition 10.3 applies since by (59), $q_i^\ell = 0$ for $i \notin I$. Using (77), we therefore find that, by definition of r^ℓ,

$$x^{\ell_k+1} = \pi[x^{\ell_k} + \alpha_{\ell_k} q^{\ell_k}] = x^{\ell_k} + \alpha_{\ell_k} r^{\ell_k},$$

$$r^{\ell_k} \to q.$$

Taylor expansion gives

$$f(x^{\ell_k+1}) = f(x^{\ell_k} + \alpha_{\ell_k} r^{\ell_k}) = f(x^{\ell_k}) + \alpha_{\ell_k} g(x^{\ell_k})^T r^{\ell_k} + O(\alpha_{\ell_k}^2) \quad \text{for } \ell_k \in L_I.$$

Comparing with (82), we find

$$\mu_{\ell_k} = \frac{f(x^{\ell_k+1}) - f(x^{\ell_k})}{\alpha_{\ell_k} g(x^{\ell_k})^T q^{\ell_k}} = \frac{g(x^{\ell_k})^T r^{\ell_k} + O(\alpha_{\ell_k})}{g(x^{\ell_k})^T q^{\ell_k}} \to 1.$$

But this contradicts the fact that the line search of Algorithm 3.3 guarantees $\mu_\ell|\mu_\ell - 1| \geq \beta$. Thus our assumption (80) cannot hold, and we have $g_I = 0$. By (67),

$$g_I(x^\ell) = g_{I_\ell}(x^\ell) = g_{\rm red}(x^\ell) \quad \text{for } \ell \in L_I.$$

Since L_I is infinite, this implies that $\inf_\ell \|g_I(x^\ell)\|_* = 0$. As in case 1, we now conclude that $g_{\rm red}(\widehat{x}) = 0$ and (ii) holds. □

The assumption that $\mathbf{x}$ is bounded, or the weaker assumption that for a given initial iterate x^0, the set $\{x \in \mathbf{x} : f(x) \leq f(x^0)\}$ is compact, implies the boundedness of the sequence x^ℓ, so that (i) or (ii) holds. (We conjecture that when neither (i) or (ii) holds then $f_\ell \to -\infty$.)

The typical situation. is that there is only one limit point $\widehat{x}$, so that $x^\ell \to \widehat{x}$. In exact arithmetic, the stationary points found are usually local minimizers as convergence of a subsequence to a nonminimizing stationary point is unstable under arbitrarily small generic perturbations. Thus one usually converges to a single local minimizer. In finite precision, one typically ends up anywhere in a region where the reduced gradient is dominated by noise due to rounding errors, so that the theory (which assumes exact arithmetic) no longer gives a reliable description of the finite precision behavior. This may in particular happen in very flat regions of the feasible domain where there is no nearby stationary point; numerical misconvergence is then possible. However, all optimization methods using only function values and gradients necessarily face this kind of difficulties.

Theorem 8.2 says that in case of convergence to a nondegenerate stationary point, all strongly active variables are ultimately fixed. Thus zigzagging through changes of the active set (as in the examples of Section 6) cannot occur infinitely often.

References

[1] O. Axelsson and G. Lindskog. On the rate of convergence of the preconditioned conjugate gradient method. *Numerische Mathematik* **48** (1986), 499–523. [23]

[2] D.P. Bertsekas. Projected Newton methods for optimization problems with simple constraints. *SIAM J. Control Opimization* **20** (1982), 221–246. [3, 27]

[3] R.H. Byrd, P. Lu, J. Nocedal, and C. Zhu. A limited memory algorithm for bound constrained optimization. *SIAM J. Sci. Comput.* **16** (1995), 1190. [27]

[4] Paul Calamai and Jorge Moré. Projected gradient methods for linearly constrained problems. *Mathematical Programming* **39** (sep 1987), 93–116. [27]

[5] A.R. Conn, N.I.M. Gould, and Ph.L. Toint. Global convergence of a class of trust region algorithms for optimization with simple bounds. *SIAM J. Numer. Anal.* **25** (1988), 433. [3]

[6] J. C. Dunn. On the convergence of projected gradient processes to singular critical points. *J. Optim. Theory Appl.* **55** (1987), 203–216. [27]

[7] R. Fletcher and C. M. Reeves. Function minimization by conjugate gradients. *Computer J.* **7** (1964), 149–154. [22, 24]

[8] A. Goldstein and J. Price. An effective algorithm for minimization. *Numer. Math.* **10** (1967), 184–189. [3]

[9] A.A. Goldstein. On steepest descent. *J. SIAM, Ser. A: Control* **3** (1965), 147–151. [6]

[10] W.W. Hager and H. Zhang. A new active set algorithm for box constrained optimization. *SIAM J. Optimization* **17** (2006), 526–557. [3, 27]

[11] W.W. Hager and H. Zhang. A survey of nonlinear conjugate gradient methods. *Pacific J. Optimization* **2** (2006), 35–58. [22]

[12] M. R. Hestenes and E. Stiefel. Methods of conjugate gradients for solving linear systems. *J. Res. Nat. Bur. Stand.* **49** (1952), 409–436. [22, 23]

[13] M. Kimiaei, A. Neumaier, and B. Azmi. LMBOPT – A limited memory method for bound-constrained optimization. Technical report, University of Vienna (2019). [3, 11]

[14] J. Nocedal and S. Wright. *Numerical optimization*. Springer Science & Business Media (2006). [2]

[15] P.M. Pardalos and N. Kovoor. An algorithm for a singly constrained class of quadratic programs subject to upper and lower bounds. *Math. Programming* **46** (1990), 321–328. [19]

[16] W. Warth and J. Werner. Effiziente Schrittweitenfunktionen bei unrestringierten Optimierungsaufgaben. *Computing* **19** (1977), 59–72. [3, 9, 11]

[17] P. Wolfe. Convergence conditions for ascent methods. *SIAM Rev.* **11** (1969), 226–235. [3]