On the Existence of Affine Invariant Descent Directions

Preprint, Aug. 10, 2018

Yu-Hong Dai*), Florian Jarre†), and Felix Lieder†)

*) The Chinese Academy of Sciences, Beijing, China,

†) Heinrich-Heine-Universität Düsseldorf, Germany.

Abstract

This paper explores the existence of affine invariant descent directions for unconstrained minimization. While there may exist several affine invariant descent directions for smooth functions at a given point, it is shown that for quadratic functions there exists exactly one invariant descent direction in the strictly convex case and generally none in the case where the Hessian is singular or indefinite. These results can be generalized to smooth nonlinear functions and have implications regarding the initialization of minimization algorithms. They stand in contrast to recent works on constrained convex and nonconvex optimization for which there may exist an affine invariant "frame" that depends on the feasible set and that can be used to define an affine invariant descent direction.

Key words: Affine invariance, descent direction, Newton direction.

1 Introduction

A prominent example of a polynomial time algorithmic scheme are interior-point methods for convex optimization. In this scheme, affine invariance is crucial in order to allow a statement about the rate of convergence of Newton's method with fixed universal constants that do not depend on affine transformations of the problem, see [11]. Due to its relevance for polynomial time algorithms, in this paper the concept of affine invariance is investigated for unconstrained minimization problems. It turns out that affine invariance may be in conflict with the descent property of a search step. This observation stands in a certain contrast to the affine invariance of Newton's method that has already been established in [5, 6], for example. In the context of minimization algorithms, affine invariance has been established for properly initialized DFP or BFGS algorithms in [9], but as shown here, the initialization may not be possible in many cases.

For constrained convex minimization a very intriguing approach to define affine invariant descent directions is presented in the recent paper [2]. Also the recent presentation [7] of a Frank-Wolfe algorithm for non-convex objective functions relies on an affine invariant analysis.

While the above list of references is far from complete, in the context of unconstrained minimization algorithms, the concept of affine invariance generally has not received due attention, and several simple but fundamental properties seem not to be widely known.

Let the first and second derivative of a function f at a point $x^{(1)}$ be given and let the first derivative be nonzero. The following facts are shown: Based on this information there does not exist any affine invariant algorithm that generates a descent direction for f at $x^{(1)}$ if zero is a multiple eigenvalue of the Hessian of f at $x^{(1)}$, or if it is nonsingular and indefinite. In shorthand notation, based on the given information "there does not exist any affine invariant descent direction". While the non-existence of affine invariant descent directions does not imply the non-existence of polynomial-time minimization algorithms, it does eliminate one of the main principles used to prove existence of such algorithms. On the other hand it is also shown, when the Hessian at $x^{(1)}$ is positive definite, then, based on this information, the only affine invariant directions are multiples of the Newton direction.

In Section 2, a formal definition of affine invariance is given along with an example highlighting the difference between an algorithm being affine invariant and the result of this algorithm for certain input data. In Section 3 the concept of affine invariance is related to the descent property and the three main results are proved. A brief conclusion is made in Section 4.

1.1 Notation

The orthogonal complement of a vector $a \in \mathbb{R}^n$ is denoted by $a^\perp = \{x \in \mathbb{R}^n \mid a^T x = 0\}$, where $\mathbb{R}^n$ is the n-dimensional real space. For a nonsingular matrix A, the inverse of its transpose is denoted by the shorthand $A^{-T} := (A^T)^{-1} = (A^{-1})^T$. For a real positive definite $n \times n$-matrix $B \succ 0$, the B-norm of a vector $y \in \mathbb{R}^n$ is given by $\|y\|_B := (y^T B y)^{1/2}$. The identity matrix is denoted by I with dimensions evident from the context. The Hadamard product (componentwise product) of two vectors x, y of same dimension is denoted by $x \circ y$, the diagonal of a matrix A is denoted by $\mathrm{diag}(A)$ and the diagonal matrix with diagonal entries x_i is denoted by $\mathrm{Diag}(x)$. For smooth functions $f : \mathbb{R}^n \to \mathbb{R}$ the k-th derivative at a point $x \in \mathbb{R}^n$ is a linear form represented by a row vector $Df(x)$ for $k = 1$, a bilinear form represented by the Hessian matrix $D^2 f(x)$ for $k = 2$, and a multilinear form represented by a tensor $D^k f(x)$ for $k > 2$. The gradient is denoted by $\nabla f(x) = Df(x)^T$.

2 Definition of affine invariance and examples

We consider an algorithmic "rule" $\mathcal{R}$ that defines a point $y \in \mathbb{R}^n$ or a (possibly ordered) set of points $Y \subset \mathbb{R}^n$ where y, Y depend on the set of one or more selected points $x^{(k)} \in \mathbb{R}^n$, $1 \le k \le m$ and on certain values of a smooth function $f : \mathbb{R}^n \to \mathbb{R}$ at the set of points $x^{(k)}$.

More precisely, y, Y may depend on the first p derivatives of f at the set of points $x^{(k)}$. If $\mathcal{R}$ defines a point y for example, then we write

$$y = \mathcal{R}\left(x^{(1)}, f(x^{(1)}), Df(x^{(1)}), D^2 f(x^{(1)}), \dots D^p f(x^{(1)}), \dots, x^{(m)}, f(x^{(m)}), \dots, D^p f(x^{(m)})\right),$$

or shortly

$$y = \mathcal{R}\left(x^{(k)}, D^l f(x^{(k)}), \quad 1 \le k \le m, \;\; 0 \le l \le p\right), \tag{1}$$

where D^l stands for the l-th derivative of f assuming the derivatives exist.

Note that each step of an algorithm such as the steepest descent algorithm with exact line search is a rule $\mathcal{R}$ defining the next iterate. Thus, the following also applies to algorithms $\mathcal{R}$.

The following discussions reduce to trivial cases when f depends on $n = 1$ variable. To avoid the separate treatment of trivial cases, the following assumption is made throughout this paper:

Assumption 2.1 *The number n of variables satisfies $n \ge 2$.*

Another elementary assumption that is made throughout this paper is that the definition of y in (1) is invariant under translations:

Assumption 2.2 *[Invariance under translations] Let $a \in \mathbb{R}^n$ be given and $\alpha \in \mathbb{R}$ and let $\hat{f}(x) := f(x+a)+\alpha$. Setting $\hat{x}^{(k)} := x^{(k)} - a$ for $1 \le k \le m$, the vector y in (1) also satisfies*

$$y = \mathcal{R}\left(\hat{x}^{(k)}, D^l \hat{f}(\hat{x}^{(k)}), \quad 1 \le k \le m, \;\; 0 \le l \le p\right).$$

It is well known and easy to see that the Newton direction and the steepest descent direction are invariant under translations. The invariance property that is of main concern in this paper is defined next:

Definition 2.1 *A rule $\mathcal{R}$ is called* affine invariant *if for any f from a specified class of functions and any nonsingular $n \times n$ matrix A and for $\hat{f}(x) := f(Ax)$, the equality*

$$\begin{aligned} & \mathcal{R}\left(\quad x^{(k)}, D^l f(x^{(k)}), \qquad 1 \le k \le m, \;\; 0 \le l \le p\right) \\ = \;\; & A\, \mathcal{R}\left(A^{-1}x^{(k)}, D^l \hat{f}(\hat{x})|_{\hat{x}=A^{-1}x^{(k)}}, \quad 1 \le k \le m, \;\; 0 \le l \le p\right) \end{aligned} \tag{2}$$

is always true.
Here $\hat{x} = A^{-1}x$ is a linear transformation of x, and $x = A\hat{x}$ its inverse transformation.

$\mathcal{R}$ is called scaling invariant *if for any positive scalar λ and for $\hat{f}(x) := \lambda f(x)$, equation (2) is satisfied when $A := I$ is the identity matrix.*

$\mathcal{R}$ is called fully affine invariant *if it is affine invariant and scaling invariant. In this case, relation (2) holds for $\hat{f}(\hat{x}) := \lambda f(A\hat{x})$ not only when $\lambda = 1$ but for any fixed $\lambda > 0$.*

The above definition assumes that the function $\hat{f}$ belongs to the specified class of functions whenever f does.

Remark 2.1 *Consider the case where $\mathcal{R}$ generates a single output y. Given an invertible matrix A, on the one side one may apply $\mathcal{R}$ to the original data. On the other side the variable x may first be linearly transformed to $\hat{x} = A^{-1}x$. Then, $\mathcal{R}$ can be applied to the transformed data returning in an intermediate result $\hat{y}$, and then, the result $\hat{y}$ can be mapped back to $y = A\hat{y}$. The definition of affine invariance states that one obtains the same point $y \in \mathbb{R}^n$ in both cases.*

The following proposition follows directly from Definition 2.1:

Proposition 2.1 *If two rules $\mathcal{R}_1$ and $\mathcal{R}_2$ defining sets $Y^{(1)}$ and $Y^{(2)}$ are (fully) affine invariant, then so are $\alpha Y^{(1)} \cup \beta Y^{(2)}$, $\alpha Y^{(1)} \cap \beta Y^{(2)}$, and $\alpha Y^{(1)} + \beta Y^{(2)}$ for any fixed real numbers α, β.*

In this paper the concept of affine invariance is related to *descent directions*. As descent directions are not always defined in the same way, a formal definition of the notation used in this paper is given next.

Definition 2.2 *A vector s is called a descent direction for a differentiable function f at a point x if $\nabla f(x)^T s < 0$.*

Clearly, given a differentiable function f and a point x, a descent direction for f at x can be defined if, and only if, $\nabla f(x) \neq 0$. Thus, $\nabla f(x) \neq 0$ will always be assumed for the rules defining a descent direction at a point x.

For algorithms that compute a normalized descent direction y with $\|y\| = 1$ affine invariance in the sense of Definition 2.1 never holds (simply choose $A = 2I$). Therefore the following slightly weaker requirement for affine invariance is also used below:

Definition 2.3 *Let $\mathcal{R}$ be a rule defining a normalized descent direction y with $\|y\| = 1$ for a function f from a specified class of functions at a given point $x^{(1)}$ with $\nabla f(x^{(1)})^T s \neq 0$. Then $\mathcal{R}$ is called an* affine invariant normalized descent direction *if for any f and any nonsingular $n \times n$ matrix A and for $\hat{f}(x) := f(Ax)$, and*

$$\begin{aligned} y &:= \mathcal{R}\big(x^{(k)}, D^l f(x^{(k)}), \quad 1 \le k \le m, \quad 0 \le l \le p \big) \\ \bar{y} &:= A\, \mathcal{R}\Big(A^{-1}x^{(k)}, D^l \hat{f}(\hat{x})|_{\hat{x}=A^{-1}x^{(k)}}, \quad 1 \le k \le m, \quad 0 \le l \le p \Big) \end{aligned}$$

the equality $y/\|y\| = \bar{y}/\|\bar{y}\|$ holds true.

Above, the requirement $y = \bar{y}$ of Definition 2.1 is replaced with $y/\|y\| = \bar{y}/\|\bar{y}\|$ since generally $\|\bar{y}\| \neq 1$ while by definition $\|y\| = 1$. Evidently, if y is an affine invariant descent direction in the sense of Definition 2.1, then $y/\|y\|$ is a normalized affine invariant descent direction in the sense of Definition 2.3.

Given an affine invariant descent direction, most commonly used line search algorithms for defining a step length along this search direction are affine invariant, so that the resulting overall algorithm is affine invariant as well. Unfortunately, as shown below, defining an

affine invariant descent direction generally is not possible in the context of unconstrained minimization, while – as shown in [2] – it is possible for example, when certain inequality constraints are given. Intuitively, this is due to the fact that the constraints provide further information that is transformed as well when changing from f to $\hat{f}$ in Definition 2.1. A similar argument holds for search directions in interior point methods for convex conic programs.

2.1 Examples

The most well known search directions in the context of strictly convex minimization (see e.g. [5, 9]) are probably the steepest descent direction and the Newton direction. For later reference, these basic examples are briefly repeated:

E1. The rule for the steepest descent direction uses $m = 1$ point and $p = 1$ derivative, and defines

$$\mathcal{R}(x, Df(x)) := -\nabla f(x) := -Df(x)^T$$

omitting the superscript " $^{(1)}$" in $x^{(1)}$ and just using x for convenience. Using $\lambda = 1$ and the definition of $\hat{f}$ in Definition 2.1, we obtain with $x = A\hat{x}$ and $\hat{f}(\hat{x}) := f(A\hat{x})$ that

$$A\,\mathcal{R}(A^{-1}x, D\hat{f}(\hat{x})|_{\hat{x}=A^{-1}x}) = -A(Df(x)A)^T = -AA^T\nabla f(x) \neq -\nabla f(x) = \mathcal{R}(x, Df(x))$$

unless A is an orthogonal matrix or $\nabla f(x) = 0$. In particular, the steepest descent step is orthogonal invariant (in the obvious sense) but not affine invariant. Nor is it scaling invariant.

Note that the normalized steepest descent step $\hat{\mathcal{R}}(x, Df(x)) := -\nabla f(x)/\|\nabla f(x)\|$ is invariant when A is a nonzero multiple of an orthogonal matrix, exemplifying the claim that Definition 2.3 is slightly weaker than Definition 2.1.

E2. On the other hand, the Newton step

$$\mathcal{R}(x, Df(x), D^2f(x)) = -(D^2f(x))^{-1}\nabla f(x)$$

satisfies

$$\begin{array}{rcl} A\,\mathcal{R}(A^{-1}x, D\hat{f}(\hat{x})|_{\hat{x}=A^{-1}x}, D^2\hat{f}(\hat{x})|_{\hat{x}=A^{-1}x}) & = & -A(A^TD^2f(x)A)^{-1}A^T\nabla f(x) \\ = \; -(D^2f(x))^{-1}\nabla f(x) & = & \mathcal{R}(x, Df(x), D^2f(x)) \end{array}$$

and is thus affine invariant in the sense of Definition 2.1. In fact it is fully affine invariant.

Remark 2.2 *The rate of convergence of an algorithm that is not affine invariant changes in general, when the problem data is given in a linearly transformed space. This offers the chance to possibly accelerate the algorithm by identifying a suitable transformation that allows a faster rate of convergence, and might thus seem as an advantage of this algorithm.*

However, it is generally just as hard to identify such transformation as to solve the problem in its original setting. For example, when minimizing $f(x) := \frac{1}{2}\|Ax - b\|_2^2$ *for a given nonsingular matrix* A *and right hand side* b*, the steepest descent method is known to be very unsuitable. Identifying the transformation such that the steepest descent method always converges in one step amounts to computing* $(A^TA)^{-1}$ *and corresponds to the affine invariant Newton method. Moreover, the affine setting in which a problem is given, typically is far away from the best possible affine setting. To condense the argument, designing an affine invariant algorithm is a very desirable feature whenever affine invariance is in reach.*

The following simple further examples may help to familiarize the concept of affine invariance.
Further Examples:

E3. While, as recalled above, the gradient $\nabla f(x)$ at a given point x is not affine invariant, the set

$$Y := \nabla f(x)^\perp := \{z \mid \nabla f(x)^T z = 0\}$$

is fully affine invariant. This, of course, is only possible since apart from $\nabla f(x)$ also the scalar product defining $\nabla f(x)^\perp$ is not affine invariant. (To establish affine invariance of Y, simply note that $\nabla f(x)^T z = 0 \Longleftrightarrow (A^T \nabla f(x))^T (A^{-1} z) = 0$.)

Also note that the open half space $\{z \mid \nabla f(x)^T z < 0\}$ is the set of all descent directions and is fully affine invariant. Thus, while the set of all descent directions is affine invariant it is shown below that identifying a single affine invariant descent direction is not possible, in general.

E4. Likewise, the sets $\{z \mid z^T D^2 f(x)^T z = 0\}$ or $\{z \mid z^T D^2 f(x)^T z \leq 0\}$ are fully affine invariant, and the set

$$Y := \{z \mid z^T D^2 f(x)^T z \leq 1\}$$

is also affine invariant (also in the nonconvex case), but it is not invariant with respect to scaling.

For strictly convex functions f, the above Y may be interpreted as an ellipsoidal trust region, and minimizing $f(x) + \nabla f(x)^T s$ over this trust region returns a scalar multiple of the Newton step, allowing the well known interpretation of Newton's method as a trust region algorithm with an affine invariant trust region. Moreover, the deviation of $f(x+s)$ from the first order approximation $f(x) + \nabla f(x)^T s$ is approximately constant

$$f(x+s) - (f(x) + \nabla f(x)^T s) \approx \frac{1}{2} s^T D^2 f(x) s \equiv \frac{1}{2}$$

on the boundary of Y, so that minimizing the linear approximation $f(x) + \nabla f(x)^T s$ over Y coincides with minimizing the quadratic approximation $f(x) + \nabla f(x)^T s + \frac{1}{2} s^T D^2 f(x) s$ over Y whenever the trust region constraint is active.

E5. For a strictly convex function $f : \mathbb{R}^n \to \mathbb{R}$ and a given $x \in \mathbb{R}^n$, let $y^{(1)}$ be the step to the global minimizer of the quadratic approximation to f, i.e. the Newton step at x and let $y^{(2)}$ be the global minimizer of the fourth order Taylor approximation to f if it exists; if not let $y^{(2)} := 0$. When $y^{(2)} \neq 0$, both $y^{(1)}$ and $y^{(2)}$ are fully affine invariant descent steps showing that there may be more than one fully affine invariant descent step for convex functions. (On the other hand, the first part of the proof of Proposition 3.1 below shows that there might not be any affine invariant descent step when f is convex but not strictly convex. In particular, using, in whatever form, the pseudo inverse in place of the inverse of the Hessian to determine a search direction generally does not lead to an affine invariant rule.)

To illustrate the example of two affine invariant descent steps, let $n = 2$ and $f(z) := z^T z + \frac{1}{4} z_2^4$ for $z \in \mathbb{R}^2$, and $x := (1, 1)^T$.

Then $-\nabla f(x) = -(2, 3)^T$, and the directions $y^{(1)} = (-1, -3/5)^T$ and $y^{(2)} = (-1, -1)^T$ are both affine invariant. Here, by Proposition 2.1, the linear combination

$$\frac{-5}{2} y^{(1)} + \frac{9}{2} y^{(2)} = \begin{bmatrix} \frac{5}{2} \\ \frac{3}{2} \end{bmatrix} + \begin{bmatrix} \frac{-9}{2} \\ \frac{-9}{2} \end{bmatrix} = \begin{bmatrix} -2 \\ -3 \end{bmatrix}$$

is fully affine invariant. It coincides with the steepest descent direction, but the steepest descent direction is not affine invariant. Thus, affine invariance does not hinge on the result y but on the rule $\mathcal{R}$ that generates y. The first rule $y = \frac{-5}{2} y^{(1)} + \frac{9}{2} y^{(2)}$ is affine invariant, and the second one $y = -\nabla f(x)$ is not. In a transformed space, $\hat{x} = A^{-1} x$ and $\hat{y}$ generated in the transformed space, the equation $\frac{-5}{2} \hat{y}^{(1)} + \frac{9}{2} \hat{y}^{(2)} = -\nabla \hat{f}(\hat{x})$ generally no longer holds true.

3 Nonexistence of fully affine invariant descent directions

In this section, rules $\mathcal{R}$ are considered that generate a descent direction y for a differentiable function f at a given point $x^{(1)}$ with $\nabla f(x^{(1)}) \neq 0$.

As was shown in Example E3 in Section 2.1, the set of descent directions is affine invariant. Nevertheless, as shown below, in general, it may not be possible to define a single "meaningful" affine invariant descent direction for a smooth function f at a given point $x^{(1)}$ unless f is strictly convex, see the discussion following the proof of Theorem 3.1 below.

We begin with an example for which the definition of an affine invariant descent direction is always possible, as long as $\nabla f(x^{(1)}) \neq 0$.

3.1 Invariant frames

Example 3.1 *[Simplex descent direction] Let $m := n+1$ affinely independent[1] points $x^{(k)}$ be given along with the function values at these points and with the first derivative at the point $x^{(1)}$ satisfying $\nabla f(x^{(1)}) \neq 0$. Then*

$$\begin{aligned} y &:= \mathcal{R}(x^{(1)},\ f(x^{(1)}),\ \nabla f(x^{(1)}),\quad x^{(k)},\ f(x^{(k)})\ \ (2 \leq k \leq n+1)) \\ &:= -\sum_{i=2}^{n+1} sign\left(\nabla f(x^{(1)})^T (x^{(i)} - x^{(1)})\right) \cdot (x^{(i)} - x^{(1)}) \end{aligned}$$

defines a descent direction at $x^{(1)}$ (since the points $x^{(k)}$ are affinely independent).

In Example 3.1 the points $x^{(k)}$ are not used to gain information about the function f, but only to define a "frame" that is used for the definition of y. And since the frame (i.e. the points $x^{(k)}$) are also transformed in Definition 2.1 of affine invariance, y is an affine invariant descent direction in the sense of Definition 2.1.

We note that the Nelder-Mead simplex method [10] also generates affine invariant descent *steps*. To generate a descent *direction*, the Nelder Mead approach was modified for the above example.

It is well known that – while easy to implement – if it converges, then the Nelder Mead algorithm often converges rather slowly in spite of being affine invariant. Hence, affine invariance is not the only aspect when aiming for rapid convergence. There are other aspects why the definition of a search direction based on such a "frame" as in the above example is inadequate:

1. Given $x^{(1)}$, it is unclear, how to generate the initial points $x^{(k)}$ for $2 \leq k \leq m$ in an affine invariant fashion. And if the points $x^{(k)}$ for $2 \leq k \leq m$ are *not* generated by an affine invariant procedure, then the overall algorithm of first generating the points $x^{(k)}$ for $2 \leq k \leq m$ and then applying the simplex descent direction is not affine invariant either.

2. Let $\mathcal{R}$ be a rule of the form (1) generating a descent direction y at the point $x^{(1)}$. Let f be locally analytic near $x^{(1)}$ and denote by $y(\epsilon)$ the result of the rule $\mathcal{R}$ when the points $x^{(k)}$ for $2 \leq k \leq m$ are replaced with $x^{(1)} + \epsilon(x^{(k)} - x^{(1)})$ for some $\epsilon \in (0,1)$. Then, in the limit as $\epsilon \to 0$, the only information about f that is used by the rule $\mathcal{R}$ are the values of f and its derivatives at the point $x^{(1)}$. Let us call the rule $\mathcal{R}$ for generating a descent direction *locally consistent in the limit* if $\bar{y} := \lim_{\epsilon \to 0} y(\epsilon)/\|y(\epsilon)\|$ exists and

   $$\bar{y} \text{ is a descent direction that only depends on } f \text{ and its derivatives at } x^{(1)}. \quad (3)$$

[1] The notion "affinely independent" is not to be confused with "affine invariant"; here, some points $x^{(k)}$ are affinely independent for $1 \leq k \leq m$ if the vectors $x^{(k)} - x^{(1)}$ are linearly independent for $2 \leq k \leq m$.

Then the above simplex descent direction lacks such local consistency in the limit: for a linear function f, the simplex descent direction $y(\epsilon)/\|y(\epsilon)\|$ is independent of $\epsilon \in (0,1)$ but does depend on the possibly arbitrary location of the support points $x^{(k)}$ for $2 \leq k \leq m$.

The condition of local consistency in the limit does not exclude the concept of using the points $x^{(k)}$ to form a model of the function f, for example a linear interpolation based on $n+1$ affinely independent support points or a quadratic interpolation as in UOBYQA [13] with a set of up to $1 + n(n+3)/2$ support points, but it does exclude the above simplex descent direction where the points are used as some form of "frame".

In the following we focus on rules $\mathcal{R}$ for which there is no such "frame dependency" of y on the location of the points $x^{(k)}$. There are several possibilities to restrict to rules that are not frame dependent. The strongest restriction would be to consider only rules that are based on only one point ($m = 1$). In order not to exclude successful approaches such as UOBYQA we use the following slightly weaker assumption:

Assumption 3.1 *The rule $\mathcal{R}$ for generating a descent direction y for f at the point $x^{(1)}$ is locally consistent in the limit in the sense of (3).*

It is assumed for the remainder of this paper that Assumption 3.1 is satisfied. The following simple observation will be used below:

Note 3.1 *Let $\mathcal{R}$ be an affine invariant descent direction in the sense of Definition 2.1 and let Assumption 3.1 be satisfied. Define a new "limiting rule" $\overline{\mathcal{R}}$ by setting $\bar{y} := \lim_{\epsilon \to 0} y(\epsilon)/\|y(\epsilon)\|$. Then $\overline{\mathcal{R}}$ is an affine invariant normalized descent direction in the sense of Definition 2.3.*

Proof. Assume that $\overline{\mathcal{R}}$ is not an affine invariant normalized descent direction, i.e. there exists a nonsingular matrix A and a function f violating Definition 2.3 for the case $m = 1$. Using the fact that the descent directions $y(\epsilon)$ in (3) satisfy the stronger condition of Definition 2.1 for this particular matrix A leads to a contradiction. □

Next, it is shown that in the degenerate case – and under Assumption 3.1 – there does not exist any algorithm generating an affine invariant descent direction:

3.2 A degenerate example

Proposition 3.1 *Let f be a function from the class of polynomials with $\lim_{\|x\| \to \infty} f(x) = \infty$. Let $\mathcal{R}$ be a rule defining a normalized descent direction for f at a point $x^{(1)}$ with $\nabla f(x^{(1)}) \neq 0$ based on the data (1). If $\mathcal{R}$ satisfies Assumption 3.1, then $\mathcal{R}$ is not affine invariant in the sense of Definition 2.3.*

Proof. First, consider the class of linear functions f with $f(x) := c^T x$ with some constant vector $c \neq 0$. While it does not make much sense to consider affine invariant descent

directions for linear functions, Definition 2.3 is still applicable to such f. In this case, due to Note 3.1, we may assume without loss of generality that the rule $\mathcal{R}$ actually is a function

$$\mathcal{R}\left(x^{(k)}, D^l f(x^{(k)}), \quad 1 \le k \le m, \;\; 0 \le l \le p\right) \;=\; \breve{\mathcal{R}}(c)$$

of just the vector $c = \nabla f(x^{(1)})$. Let A be an invertible matrix and let

$$\begin{aligned} \breve{\mathcal{R}}(c) \;&:=\; \mathcal{R}\left(\quad x^{(k)}, D^l f(x^{(k)}), \qquad\qquad 1 \le k \le m, \;\; 0 \le l \le p\right) \qquad \text{and} \\ A\,\breve{\mathcal{R}}(A^T c) \;&:=\; A\,\mathcal{R}\left(A^{-1}x^{(k)}, D^l \hat{f}(\hat{x})|_{\hat{x}=A^{-1}x^{(k)}}, \quad 1 \le k \le m, \;\; 0 \le l \le p\right). \end{aligned}$$

Then, in order for $\breve{\mathcal{R}}$ to be affine invariant in the sense of Definition 2.3, the relation

$$\breve{\mathcal{R}}(c) = A\,\breve{\mathcal{R}}(A^T c)/\|A\,\breve{\mathcal{R}}(A^T c)\| \tag{4}$$

must be satisfied for all nonzero c and all nonsingular A. Now, fix some nonzero vector $\bar{c}$ and let $\bar{y} := \breve{\mathcal{R}}(\bar{c})$. Then, by the descent property, $\bar{c}^T\bar{y} < 0$, and thus, $\bar{A} := (I - \frac{\bar{c}\bar{c}^T}{\bar{c}^T\bar{c}} - \frac{\bar{y}\bar{y}^T}{\bar{c}^T\bar{y}})^{1/2}$ is well defined, positive definite, and satisfies $\bar{y} = -\bar{A}^2\bar{c}$. Now, for any $\widetilde{c}$ with $\widetilde{c}^T\bar{A}\bar{c} \neq 0$ let $\widetilde{A}^T := \bar{A}^{-1} + uv^T$ where $u := \bar{c} - \bar{A}^{-1}\widetilde{c}$ and $v := \widetilde{c}/\|\widetilde{c}\|_2^2$. Then, by the Sherman-Morrison update formula for inverse matrices, $\widetilde{A}^T$ is nonsingular, and by definition of $\widetilde{A}^T$ it follows that $\widetilde{A}^T\widetilde{c} = \bar{c}$. Let $\beta = \beta(\widetilde{c}) := 1/\|\widetilde{A}\,\breve{\mathcal{R}}(\widetilde{A}^T\widetilde{c})\|$. By (4) and the above definitions it follows

$$\breve{\mathcal{R}}(\widetilde{c}) = \beta\widetilde{A}\,\breve{\mathcal{R}}(\widetilde{A}^T\widetilde{c}) = \beta\widetilde{A}\,\breve{\mathcal{R}}(\bar{c}) = \beta\widetilde{A}\bar{y} = -\beta\widetilde{A}\bar{A}^2\bar{c} = -\beta\widetilde{A}\bar{A}^2\widetilde{A}^T\widetilde{c}.$$

By definition, $\bar{A}\widetilde{A}^T = \bar{A}(\bar{A}^{-1} + uv^T) = I + \bar{u}v^T$ where $\bar{u} := \bar{A}u$, and thus,

$$\widetilde{A}\bar{A}^2\widetilde{A}^T = (I + v\bar{u}^T)(I + \bar{u}v^T).$$

Using $v^T\widetilde{c} = 1$, $\bar{u} = \bar{A}\bar{c} - \widetilde{c}$, and $v = \widetilde{c}/\|\widetilde{c}\|_2$ we obtain

$$\breve{\mathcal{R}}(\widetilde{c}) = -\beta(I + v\bar{u}^T)(\widetilde{c} + \bar{u}) = -\beta(I + v\bar{u}^T)(\bar{A}\bar{c}) = -\beta(\rho\widetilde{c} + \bar{A}\bar{c}) \tag{5}$$

where $\rho = \rho(\widetilde{c}) = \bar{u}^T\bar{A}\bar{c}/\|\widetilde{c}\|_2^2$ is some scalar, $-\widetilde{c}$ is the steepest descent direction, and $\bar{A}\bar{c}$ is a fixed vector. To show that this rule is not affine invariant, relation (4) is applied again: For a given invertible matrix A let $\hat{\beta} := 1/\|A\,\breve{\mathcal{R}}(A^T\widetilde{c})\|$. Then (5) and (4) imply that

$$-\beta(\widetilde{c})(\rho(\widetilde{c})\widetilde{c} + \bar{A}\bar{c}) = \breve{\mathcal{R}}(\widetilde{c}) = \hat{\beta}A\,\breve{\mathcal{R}}(A^T\widetilde{c}) = -\hat{\beta}A\left(\rho(A^T\widetilde{c})A^T\widetilde{c} + \bar{A}\bar{c}\right) \tag{6}$$

must be satisfied for all invertible matrices A and any $\widetilde{c}$ with

$$(A\widetilde{c})^T\bar{A}\bar{c} \neq 0. \tag{7}$$

Choose $\widetilde{c} := -\bar{A}\bar{c} + w$ where $w \in (\bar{A}\bar{c})^\perp$ is some nonzero vector orthogonal to $\bar{A}\bar{c}$. Then choose $A = A^T$ such that $\bar{A}\bar{c}$ is an eigenvector of A to the eigenvalue $\beta(\widetilde{c})/\hat{\beta} > 0$ and such

that all other eigenvalues of A are greater than $\beta(\widetilde{c})/\hat{\beta}$. For this choice it follows (from the orthogonality of the eigenvectors of A) that

$$(A\widetilde{c})^T \bar{A}\bar{c} = (\beta(\widetilde{c})/\hat{\beta})(-\bar{A}\bar{c})^T \bar{A}\bar{c} < 0$$

i.e. condition (7) is satisfied. Moreover, relation (6) simplifies to

$$\beta(\widetilde{c})\rho(\widetilde{c})\widetilde{c} = \hat{\beta}\rho(A\widetilde{c})A^2\widetilde{c}. \tag{8}$$

Since $w \neq 0$, the vector $\widetilde{c}$ is not an eigenvector of A^2. Therefore, (8) implies $\rho(\widetilde{c}) = \rho(A\widetilde{c}) = 0$. Since $\widetilde{c}^T \bar{A}\bar{c} < 0$ this is in contradiction to the descent property of $-(\rho(\widetilde{c})\widetilde{c} + \bar{A}\bar{c})$.

Above, the case of $p = 1$ derivative at the points $x^{(k)}$ was considered. Now, assume $p \geq 1$ and let

$$r_k(x) := \|x - x^{(k)}\|_2^{2p} \geq 0 \quad \text{for} \quad 1 \leq k \leq m$$

and set $q(x) := \prod_{k=1}^m r_k(x)$. Let $c \in \mathbb{R}^n \backslash \{0\}$ and $f : \mathbb{R}^n \to \mathbb{R}$ be defined by $f(x) := c^T x + q(x)$.

Then, f satisfies the assumptions of Proposition 3.1, and f and all derivatives of f up to order p coincide at all $x^{(k)}$ with the values of the linear function ℓ with $\ell(x) := c^T x$ for $x \in \mathbb{R}^n$. Thus, all the information that is input for $\mathcal{R}$ coincides with the information that would be provided by the function ℓ. As seen above, this input does not allow the definition of an affine invariant descent direction. □

Remark 3.1 *The above proof holds for functions of at least 2 variables. It also applies to convex quadratic functions if zero is an eigenvalue of the Hessian of multiplicity at least two. The case where the Hessian is nonsingular is discussed in Proposition 3.2 and Theorem 3.1 below. The case of a convex quadratic function where zero is an eigenvalue of the Hessian of multiplicity one is left open. Note, however, that the generalized Newton step (using the pseudo-inverse H^+ in place of the inverse H^{-1} of the Hessian) does not provide an affine invariant descent step in this case. (First of all, this step may be zero even when the gradient is nonzero, and even if it is nonzero it is not affine invariant since the necessary relation $A(A^T H A)^+ A^T = H^+$ for all invertible A does not hold).*

By Note 3.1 also the following statement is true:

Corollary 3.1 *Let f be a function from the class of p-times continuously differentiable convex functions with a unique minimizer and with $\lim_{\|x\|\to\infty} f(x) = \infty$. Let $\mathcal{R}$ be a rule defining a normalized descent direction for f at a point $x^{(1)}$ with $\nabla f(x^{(1)}) \neq 0$ based on the data (1). If $\mathcal{R}$ satisfies Assumption 3.1, then $\mathcal{R}$ is not affine invariant in the sense of Definition 2.3.*

Proof. The argument of the preceding proof also applies when fixing some positive definite $n \times n$-matrix $B \succ 0$ and some sufficiently large constant M, and replacing $q(x)$ with $\max\{0, \|x\|_B - M\}^{2p}$. □

Proposition 3.1 and Corollary 3.1 consider very limited classes of functions. If it is not possible to define affine invariant descent directions for such small classes, then of course, it is not possible for larger classes of functions either. Nevertheless, the example in the proof of Proposition 3.1 is not fully convincing as an argument to abandon the search for affine invariant descent steps in the nonconvex case: After all, Newton's method is not defined in the above setting either, but nevertheless, in the convex case, Newton's method generally is the minimization method of choice if the derivatives are available at acceptable computational costs. The next example therefore considers the nondegenerate case.

3.3 A nondegenerate example

As seen in Example 3 in Section 2 there may be more than one linearly independent affine invariant descent direction for strictly convex functions f. On the other hand, given a point x, $g := \nabla f(x) \neq 0$, and $H := D^2 f(x) \succ 0$, there are also numerous ways to define a descent direction s for a strictly convex function f at x, such as $s := (H + \mathrm{Diag}(g \circ g))^{-1} g$, for example. However, apart from the Newton direction, none of them is affine invariant:

Proposition 3.2 *When f is strictly convex and quadratic, then, up to scalar multiples, the Newton direction for minimizing f is the only affine invariant direction that satisfies Assumption 3.1.*

Proof. Let $f(x) := \frac{1}{2} x^T H x + c^T x$ be a strictly convex quadratic function. Let some support points $x^{(k)}$ $(1 \leq k \leq m)$ and the values of f and its derivatives at the support points be given. By invariance with respect to translations we assume without loss of generality that $x^{(1)} = 0$. Then, similar to the proof of Proposition 3.1, the rule $\mathcal{R}$ only depends on c and H. By $\breve{\mathcal{R}}$ we denote again the restriction of $\mathcal{R}$ to changes in c and H. Let $y^{(1)} := -H^{-1}c = \breve{\mathcal{R}}_1(c, H)$ be the Newton direction. Assume that there are c, H for which there is a second, linearly independent, affine invariant direction $y^{(2)} = \breve{\mathcal{R}}_2(c, H)$. Further assume that $c \neq 0$. (The case that $c = 0$ implies with analogous arguments that apart from $y = 0$ there is no affine invariant direction at all.) Then, there is a congruence transformation $H \mapsto A^T H A$ such that $A^T H A = I$ is the identity matrix. Further there exists an orthogonal matrix U such that $U^T A^T c =: \bar{c}$ where $\bar{c}$ is a multiple of the first canonical unit vector. Let $\bar{A} := AU$ and $\hat{f}(\hat{x}) := f(\bar{A}\hat{x})$. Then

$$\nabla \hat{f}(0) = \bar{c} \qquad \text{and} \qquad D^2 \hat{f}(0) = \bar{H} := I.$$

By assumption, f and also $\hat{f}$ has two linearly independent affine invariant directions $\hat{y}^{(1)}$ and $\hat{y}^{(2)}$. Here, $\hat{y}^{(1)} := -\bar{H}^{-1}\bar{c} = -\bar{c}$ is the Newton direction and $\hat{y}^{(2)} = \breve{\mathcal{R}}_2(\bar{c}, \bar{H})$. Note that the first component of $\hat{y}^{(1)}$ satisfies $\rho_1 := \hat{y}_1^{(1)} \neq 0$. If $\rho_2 := \hat{y}_1^{(2)} \neq 0$ then replace $\breve{\mathcal{R}}_2$ with $\widetilde{\mathcal{R}_2} := \breve{\mathcal{R}}_2 - \frac{\rho_2}{\rho_1}\breve{\mathcal{R}}_1$. By Proposition 2.1, $\widetilde{\mathcal{R}_2}$ is affine invariant and $0 = \widetilde{y_1}^{(2)} = \widetilde{\mathcal{R}_2}(\bar{c}, \bar{H})_1$. Thus, we may assume without loss of generality that the first component of $\hat{y}^{(2)}$ satisfies $\hat{y}_1^{(2)} = 0$. Now, let $\tilde{U}$ be the orthogonal transformation that leaves the first canonical unit vector invariant and that changes the sign of the remaining components – including the sign

of all nonzero components of $\hat{y}^{(2)}$. Then, $\bar{c} = \tilde{U}^T \bar{c}$ and $\bar{H} = \tilde{U}^T \bar{H} \tilde{U}$ are unchanged, and affine invariance

$$-\hat{y}^{(2)} = \tilde{U}\hat{y}^{(2)} = \tilde{U}\breve{\mathcal{R}}_2(\bar{c}, \bar{H}) = \breve{\mathcal{R}}_2(\tilde{U}^T\bar{c}, \tilde{U}^T\bar{H}\tilde{U}) = \breve{\mathcal{R}}_2(\bar{c}, \bar{H}) = \hat{y}^{(2)}$$

implies that $\hat{y}^{(2)} = 0$. Thus, the assumption that there is a linearly independent affine invariant direction apart from the Newton direction is wrong. □

We now consider the existence of affine invariant descent directions in the nondegenerate nonconvex case:

Theorem 3.1 *Let $\mathcal{R}$ be a rule defining a descent direction at a point $x^{(1)}$ based on the data (1) with m points and $p \geq 2$ derivatives. Let Assumption 3.1 be satisfied, and let f be from the class of polynomials with gradient $\nabla f(x^{(1)}) \neq 0$ and determinant of the Hessian $\det(D^2 f(x^{(1)})) \neq 0$. Then $\mathcal{R}$ is not fully affine invariant, in general, even when $\lim_{\|x\|\to\infty} f(x) = \infty$.*

Proof. As in the proof of Proposition 3.1 we first consider an unbounded function: For a fixed vector $c = (c_1, c_2)^T \in \mathbb{R}^2$ let $f(x) := \frac{1}{2}(x_1^2 - x_2^2) + c^T x$. The Hessian of f is constant and equals $H = \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix}$. Following the proof of Proposition 3.1, let again some support points $x^{(k)}$ $(1 \leq k \leq m)$ and the values of f and its derivatives at the support points be given. By invariance with respect to translations we assume again that $x^{(1)} = 0$. Then, the rule $\mathcal{R}$ only depends on c and H, and by the assumptions of Theorem 3.1, $c \neq 0$. By $\breve{\mathcal{R}}$ we denote again the restriction of $\mathcal{R}$ to changes in c and H. Let $y^{(1)} := -H^{-1}c = \breve{\mathcal{R}}_1(c, H)$ be the Newton direction for solving $\nabla f(x) = 0$ starting at $x^{(1)} = 0$. Then, the step s with

$$s := y^{(1)} \;\text{ if }\; c^T y^{(1)} < 0 \qquad \text{and} \qquad s := -y^{(1)} \;\text{ if }\; c^T y^{(1)} > 0 \tag{9}$$

provides an affine invariant descent direction for f at $x^{(1)} = 0$ whenever $|c_1| \neq |c_2|$. (For $|c_1| = |c_2|$ neither of the two cases in (9) applies; in this case the Newton step does not supply any descent direction.) Assume that for $c = (1, 1)^T$ (the case where the Newton direction does not provide a descent step) there is a second, linearly independent, affine invariant direction $y^{(2)} = \breve{\mathcal{R}}_2(c, H)$.

Let

$$B := \begin{bmatrix} \sqrt{b^2+1} & b \\ b & \sqrt{b^2+1} \end{bmatrix} \tag{10}$$

for some $b \in \mathbb{R}$, $b > 0$. Then $\det(B) = 1$ and $B^T H B = H$.

For $\lambda > 0$ now consider the function $\hat{f}$ with $\hat{f}(\hat{x}) := \lambda f(\frac{1}{\sqrt{\lambda}} B\hat{x})$. Then, the Hessian of $\hat{f}$ is given by $\hat{H} = H$ (independently of the choice of $b > 0$ and $\lambda > 0$).

Full affine invariance with respect to $A := \frac{1}{\sqrt{\lambda}} B$ and $\lambda > 0$ implies

$$\begin{aligned} & \mathcal{R}\big(\quad x^{(k)}, D^l f(x^{(k)}), \qquad\qquad 1 \leq k \leq m, \;\; 0 \leq l \leq p \big) \\ = \;\; & A\, \mathcal{R}\Big(A^{-1}x^{(k)}, D^l \hat{f}(\hat{x})|_{\hat{x}=A^{-1}x^{(k)}}, \quad 1 \leq k \leq m, \;\; 0 \leq l \leq p \Big) \end{aligned}$$

where the gradient of $\hat{f}$ at $\hat{x} = 0$ is given by

$$\hat{c} := \nabla \hat{f}(0) = \lambda \frac{1}{\sqrt{\lambda}} B^T \nabla f(0) =: \sqrt{\lambda} B^T c.$$

Full affine invariance of the descent step $\breve{\mathcal{R}}_2$ therefore implies

$$y^{(2)} = \breve{\mathcal{R}}_2(c, H) = A\ \breve{\mathcal{R}}_2(\hat{c}, \hat{H}) = \frac{1}{\sqrt{\lambda}} B\ \breve{\mathcal{R}}_2(\hat{c}, H).$$

Fixing $\lambda := (\sqrt{b^2+1} + b)^{-2}$ it follows that $\hat{c} = c$, and thus,

$$y^{(2)} = \frac{1}{\sqrt{\lambda}} B\ y^{(2)}$$

i.e. $y^{(2)}$ is an eigenvector of $\frac{1}{\sqrt{\lambda}} B$ to the eigenvalue 1. Hence, $y^{(2)}$ is a multiple of $(1, -1)^T$, i.e. a multiple of the Newton direction. In particular, the Newton direction is the only affine invariant direction (up to multiples) and thus, for $c = (1, 1)^T$ there does not exist any affine invariant descent direction at $x^{(1)} = 0$.

Note that the example of this proof can be modified as in the proof of Proposition 3.1 to guarantee $\lim_{\|x\| \to \infty} f(x) = \infty$.

Also note that when $\lim_{\|x\| \to \infty} f(x) = \infty$ and f has a unique global minimizer, the step towards this minimizer is affine invariant, but given some data format as in Definition 2.1 one can construct a polynomial f with and $\lim_{\|x\| \to \infty} f(x) = \infty$ such that this minimizer cannot be computed based on this data. □

To be consistent with Definition 2.1, the above proof considers changes of c fixing $x^{(1)} = 0$. Instead, one may fix $c = 0$ and consider variations of $x = x^{(1)}$. In this case, the step s of (9) provides an affine invariant descent direction for f whenever $|x_1| \neq |x_2|$. Thus, nonexistence of an affine invariant descent direction is established only on a set of measure zero. However, when x approaches this set, the above step s violates the standard assumption for descent methods (see e.g. [12])

$$s^T \nabla f(x) \leq -\sigma \|s\|_2 \|\nabla f(x)\|_2 \qquad \text{for some fixed } \ \sigma \in (0, 1] \tag{11}$$

and it is thus not a suitable search direction for minimization algorithms also in a neighborhood of $x = (t, \pm t)^T$ for any $t \in \mathbb{R}$. This observation clarifies the statement at the beginning of Section 3 saying that there is no "meaningful" affine invariant descent step: While affine invariance is a desirable property, one would generally not consider the above search step s even if $x \neq (t, \pm t)^T$ with $t \in \mathbb{R}$.

We close this section with the remark that in the generic nonconvex case of an indefinite but nonsingular Hessian matrix, there always is a (non-unique) linear transformation as in Definition 2.1 such that $D^2 \hat{f}(x) = \begin{bmatrix} I & 0 \\ 0 & -I \end{bmatrix}$ where, the dimensions of I and of $-I$ may

differ. Moreover, as in the proof of Proposition 3.2, one may assume that c has at most two nonzero components, one in the block associated with I the other in the block associated with $-I$. Restricting the considerations to the subspace spanned by these two components leads to the example in the proof of Theorem 3.1. Hence, the counterexample in the proof considers the generic case for $n = 2$.

As is well known, see e.g. Theorems 7.23 and 8.10 in [9], the DFP algorithm or the BFGS algorithm with exact line search are affine invariant if they are initialized with the Newton direction and the Hessian or its inverse at $x^{(1)}$ – and if the Hessian at $x^{(1)}$ is positive definite.

The results of this section indicate, that when $D^2f(x^{(1)})$ is not positive definite, then it is impossible, in general, to define an affine invariant initialization for the DFP/BFGS algorithm based on $Df(x^{(1)})$ and $D^2f(x^{(1)})$, and it is also impossible, in general, to generate an affine invariant initialization based on a finite number of initial points unless these initial points are generated by an affine invariant algorithm.

4 Conclusion

The concept of affine invariance for unconstrained minimization algorithms implies several interesting facts that are not dealt with in the more general framework focusing on nonlinear equations as discussed, for example, in [5, 6]. In particular several new results regarding the existence and the uniqueness of affine invariant descent directions of a function f at a given starting point $x^{(1)}$ could be established in Section 3. The results of this work indicate that, in general, it is not possible to define an affine invariant descent direction based on a finite number of function and derivative values of f even when the Hessian of f at $x^{(1)}$ is nonsingular.

5 Acknowledgment

Part of this work was completed while the second author was visiting CAS. Generous support from CAS is gratefully acknowledged. The authors are thankful to two anonymous referees. Their criticism was of great help and lead to significant changes in the revision of the first version of this paper.

References

[1] X.M. An, D.H. Li, and Y.H. Xiao: Sufficient descent directions in unconstrained optimization, Comp. Opt. and Appl. Vol. 48, No. 3, (2011) 515–532.

[2] A. d'Aspremont, C. Guzman, and M. Jaggi: An Optimal Affine Invariant Smooth Minimization Algorithm ArXiv PREPRINT: 1301.0465 (2016).

[3] A.R. Conn, N.I.M. Gould, and P.L. Toint: Convergence of quasi-Newton matrices generated by the symmetric rank one update, Math. Prog. 50, (1991) 177–195.

[4] J.E. Dennis and J.J. Moré: Quasi-Newton methods, motivation and theory, SIAM Review 19, (1977) 46–89.

[5] P. Deuflhard: Newton Methods for Nonlinear Problems Affine Invariance and Adaptive Algorithms, Springer Series in Computational Mathematics (2004) Springer Verlag.

[6] P. Deuflhard and G. Heindl: Affine Invariant Convergence Theorems for Newton's Method and Extensions to Related Methods, SIAM J. Numer. Anal. 16, 1–10 (1979).

[7] S. Lacoste-Julien: Convergence Rate of Frank-Wolfe for Non-Convex Objectives, Preprint, arXiv:1607.00345v1, (2016).

[8] M. Lazar and F. Jarre: Calibration by optimization without using derivatives, Optimization and Engineering Vol. 17, No. 4 (2016) 833-860.

[9] J. N. Lyness: The Affine Scale Invariance of Minimization Algorithms, Mathematics of Computation, Vol. 33, No. 145 (1979) 265–287.

[10] Nelder, J.A. and Mead, R. (1965): A simplex method for function minimization. Computer J., **7**, 308-313

[11] Y. Nesterov and A. Nemirovskii: Interior-Point Polynomial Algorithms in Convex Programming, Studies in Applied and Numerical Mathematics SIAM (1994).

[12] J. Nocedal and S. Wright: Numerical Optimization, Springer Series in Operations Research and Financial Engineering, Springer-Verlag New York (2006).

[13] M.J.D. Powell: UOBYQA: unconstrained optimization by quadratic approximation, Math. Prog., Ser. B. Vol 92 (2002) 555–582.