

LOCAL RESULTS FOR THE GAUSS-NEWTON METHOD ON CONSTRAINED RANK-DEFICIENT NONLINEAR LEAST SQUARES

JERRY ERIKSSON AND MÅRTEN E. GULLIKSSON

ABSTRACT. A nonlinear least squares problem with nonlinear constraints may be ill posed or even rank-deficient in two ways. Considering the problem formulated as $\min_x 1/2\|f_2(x)\|_2^2$ subject to the constraints $f_1(x) = 0$, the Jacobian $J_1 = \partial f_1/\partial x$ and/or the Jacobian $J = \partial f/\partial x$, $f = [f_1; f_2]$, may be ill conditioned at the solution.

We analyze the important special case when J_1 and/or J do not have full rank at the solution. In order to solve such a problem, we formulate a nonlinear least norm problem. Next we describe a truncated Gauss-Newton method. We show that the local convergence rate is determined by the maximum of three independent Rayleigh quotients related to three different spaces in \mathbb{R}^p .

Another way of solving an ill-posed nonlinear least squares problem is to regularize the problem with some parameter that is reduced as the iterates converge to the minimum. Our approach is a Tikhonov based local linear regularization that converges to a minimum norm problem. This approach may be used both for almost and rank-deficient Jacobians.

Finally we present computational tests on constructed problems verifying the local analysis.

1. INTRODUCTION

A difficult problem when solving nonlinear least squares problems with nonlinear constraints is when the Jacobians involved become ill conditioned. This may be the case at an iteration point when the Gauss-Newton method is used. Different stability strategies have been developed in order to get a well-defined search direction at the same time achieving global convergence with a fast local convergence rate. Two important ways to stabilize a Gauss-Newton method is subspace minimization and Levenberg-Marquardt techniques; see [13] and [14]. However, these kinds of stabilization require quite a lot of technical details both in theory and implementation. Moreover, these techniques are not directly applicable to problems where the Jacobians are ill-conditioned or rank-deficient at the solution point.

In this paper we want to initialize another approach aiming to regularize the original problem and develop Gauss-Newton based methods that can solve ill-conditioned constrained problems. For the unconstrained case see [5]. It is natural to start with the case where the Jacobians are rank-deficient in a neighbourhood

Received by the editor March 29, 2002 and, in revised form, February 12, 2003.

2000 *Mathematics Subject Classification*. Primary 65F22, 65K05.

Key words and phrases. Nonlinear least squares, nonlinear constraints, optimization, regularization, Gauss-Newton method.

of the solution. Thus, the analysis will be local. The methods we will consider are a truncated Gauss-Newton method and a locally defined Tikhonov method. However, our final goal is to construct a Gauss-Newton method on a suitable regularized problem that can solve almost any kind of ill-conditioned problem.

Examples of rank-deficient problems are underdetermined problems [16], nonlinear regression problems [1], nonlinear total least squares problems [12], and artificial neural networks [6]. Note that all these problems may have nonlinear (rank-deficient) constraints. Another equally important reason for looking at rank-deficient problems is the connection with regularization [10].

Our analysis, in the linear case, can partly be found in [18], [19], [11], [3], [4], [10] but is treated here in a way that fits a nonlinear setting. The local results for the rank-deficient constrained nonlinear least squares problem are, to our best knowledge, new but build on earlier work in [13], [20], [9], [5], [8].

1.1. Outline of the paper. The paper is structured as follows. First we briefly motivate and formulate the least norm problems that are relevant for solving rank-deficient problems.

In Section 3 we linearize the minimization problem and make the local convergence analysis. This analysis is divided in two parts where we first derive the asymptotic convergence rate and then perform a more complex local analysis. The results from these two approaches reveal different aspects of the local behavior of the truncated method.

The Tikhonov regularization is introduced in Section 4. In this section we begin by describing the unconstrained regularization to show that the constrained case is quite different. Then we conclude that a straightforward use of penalty techniques together with Tikhonov regularization of the Jacobian is not adequate when the constraints are rank-deficient at the solution. Therefore, we consider a more sophisticated use of the linearized problem attaining a Tikhonov regularization method that gives well-defined estimates of the Lagrange parameters.

We have chosen to use artificial test problems when performing the computational experiments as described in Section 5. Thus, we are able to verify the results from the local convergence analysis.

Finally, we make some conclusions and describe the problems to be solved in order to attain a complete globally convergent optimization method.

2. REFORMULATING THE PROBLEMS

2.1. The need for a reformulation. We will formulate a problem that can be used to solve constrained problems that are rank-deficient at the solution. However, let us for the sake of clarity first consider the unconstrained least squares problem

$$(1) \quad \min_x \frac{1}{2} \|f(x)\|_2^2 = F(x),$$

where $f : \mathbb{R}^n \rightarrow \mathbb{R}^m$ is at least twice continuously differentiable and $\|\cdot\|_2$ is the 2-norm. The first order KKT-condition for (1) is

$$(2) \quad \nabla_x F = J^T f = 0,$$

where $J = \partial f / \partial x$ is the Jacobian of f . A solution \hat{x} to (2) will be called a critical point. The following theorem characterizes a problem that has a rank-deficient Jacobian in a neighborhood of a critical point. The proof of the theorem can be found in [5].

Theorem 2.1. *Let \hat{x} be a critical point and let the rank of J be equal to $r < n$ in a neighborhood of \hat{x} . Then $\nabla_{xx}^2 F(\hat{x})$ is a matrix of rank $r < n$ with its nullspace containing the nullspace of $J(\hat{x})$.*

We may conclude that having J rank-deficient makes (1) an ill-posed problem in the sense that (2) does not have a unique solution (but a local minimum to (1) may exist though). Therefore, a reformulation of the problem is needed.

Consider now the nonlinear least squares problem with nonlinear constraints. We formulate this problem as

$$(3) \quad \min_x \frac{1}{2} \|f_2(x)\|_2^2$$

$$(4) \quad \text{s.t.} \quad f_1(x) = 0,$$

where $f_1 : \mathbb{R}^n \rightarrow \mathbb{R}^{m_1}$, $f_2 : \mathbb{R}^n \rightarrow \mathbb{R}^{m_2}$ with, for the sake of simplicity, $n \leq m = m_1 + m_2$. For notational convenience we define $f = [f_1; f_2]$, $J = [J_1; J_2] = \partial f / \partial x$ with $J_i = \partial f_i / \partial x$, $i = 1, 2$. The first order KKT-conditions for this problem read

$$(5) \quad J_2^T f_2 + J_1^T \lambda_1 = 0, \quad f_1 = 0.$$

We will call a solution to (5) a critical point. We assume that $\text{rank}(J) = r \leq n$ and $\text{rank}(J_1) = s \leq m_1$ in a neighborhood of the critical point of interest.

It is easy to state the KKT-conditions when J, J_1 both have full rank in a neighborhood of the solution. If either J or J_1 is not of full rank at a critical point, we say that the problem is rank-deficient (or ill posed). We will motivate this statement further before going into the different problem reformulations. It is natural to consider the constrained problem (3–4) ill posed if (5) does not have a locally unique solution. This will be the case if the matrix

$$(6) \quad K = \begin{bmatrix} \nabla_{xx}^2 \mathcal{L} & J_1^T \\ J_1 & 0 \end{bmatrix}, \quad \nabla_{xx}^2 \mathcal{L} = J_2^T J_2 + \lambda_1^T \odot f_1'' + f_2^T \odot f_2''$$

is singular. Here we have introduced the operator \odot defined as

$$y^T \odot g'' = \sum_{j=1}^m y_j g_j''$$

for $y \in \mathbb{R}^m$ and $g : \mathbb{R}^n \rightarrow \mathbb{R}^m$ a twice continuously differentiable function. We have the following lemma from [8].

Lemma 2.1. *Define $\mathcal{P}_{\mathcal{N}(J_1)}$ as the projection on the nullspace of the Jacobian of f_1, J_1 . The matrix K in (6) is singular if and only if J_1^T or $\mathcal{P}_{\mathcal{N}(J_1)} \nabla_{xx}^2 \mathcal{L} \mathcal{P}_{\mathcal{N}(J_1)}$ is rank-deficient.*

We will assume that $m_1 < r$. This assumption may be regarded as a constraint qualification when J is rank-deficient and seems not to be a severe restriction in practice. The assumption is implicitly used in the following theorem also from [8].

Theorem 2.2. *Assume that J is rank-deficient in a neighborhood of a critical point. Then $\nabla_{xx}^2 \mathcal{L}$ in (6) is singular with $\mathcal{N}(J) \subset \mathcal{N}(\nabla_{xx}^2 \mathcal{L})$ and $\mathcal{R}(\nabla_{xx}^2 \mathcal{L}) = \mathcal{R}(J^T)$. Moreover, $\mathcal{P}_{\mathcal{N}(J_1)} \nabla_{xx}^2 \mathcal{L} \mathcal{P}_{\mathcal{N}(J_1)}$ (and thus K) is singular with a nullspace in $\mathcal{N}(J_1) \cap \mathcal{N}(J)$.*

Theorem 2.2 makes it clear that J or J_1 rank-deficient in a neighborhood of a critical point gives an ill-posed problem.

2.2. Reformulation to minimum norm problems. Now we turn to the question of reformulating our problems and start with the unconstrained problem. In the unconstrained case it is natural to find the minimum norm solution when J is rank-deficient since it is of interest that the solution is of reasonable size with a residual as small as possible. Therefore, we may use the minimum norm problem

$$(7) \quad \min_x \frac{1}{2} \|x - x_c\|_2^2$$

$$(8) \quad \text{s.t. } \min_x \frac{1}{2} \|f(x)\|_2^2$$

as a regularized version of (1). The center x_c is chosen from a priori information and should ideally be an approximation of the solution.

One possible extension of (7–8) to the constrained problem when only J is rank-deficient is to consider

$$(9) \quad \min_x \frac{1}{2} \|x - x_c\|_2^2$$

$$(10) \quad \text{s.t. } \min_x \frac{1}{2} \|f_2(x)\|_2^2$$

$$(11) \quad \text{s.t. } f_1(x) = 0.$$

Problem (9–11) is to be understood as minimizing $\|x - x_c\|_2$ where x is in the solution set of problem (3). If in addition the constraints are ill posed in the sense that J_1 is rank-deficient in a neighborhood of a critical point, we formulate the problem as

$$(12) \quad \min_x \frac{1}{2} \|x - x_c\|_2^2$$

$$(13) \quad \text{s.t. } \min_x \frac{1}{2} \|f_2(x)\|_2^2$$

$$(14) \quad \text{s.t. } \min_x \frac{1}{2} \|f_1(x)\|_2^2.$$

Again these three minimization problems are to be thought of as finding the minimum distance to x_c subject to x being in the solution set of the two inner minimization problems.

3. A TRUNCATED GAUSS-NEWTON METHOD

A locally defined truncated Gauss-Newton method for (12–14) is attained if we linearize f_1 , f_2 , and x around the current iterate x_k ; i.e., at iteration k we solve the linearized problem

$$(15) \quad \min_p \frac{1}{2} \|p + x_k - x_c\|_2^2$$

$$(16) \quad \text{s.t. } \min_p \frac{1}{2} \|J_2(x_k)p + f_2(x_k)\|_2^2$$

$$(17) \quad \text{s.t. } \min_p \frac{1}{2} \|J_1(x_k)p + f_1(x_k)\|_2^2,$$

where x_c is chosen by the user and $x_c = 0$ if no a priori information is known.

In this section we will analyze the local convergence when (15–17) is used for attaining the new approximation $x_{k+1} = x_k + p_k$ where p_k is the solution of (15–17).

We will do this analysis in two different ways since the two approaches will show different aspects of the method. First we use the formulation used in Section 3.1 and derive expressions for the local asymptotic convergence rate by differentiation. Secondly, we use the perturbation theory in [19] and attain estimates for the local convergence with a remainder term.

3.1. A solution based on projections and pseudoinverses. In this section we derive a solution of (15–17) by using pseudo inverses of J and J_1 . The solution will then be used in an asymptotic convergence analysis.

Let us simplify (15–17) by skipping the arguments and indices giving

$$(18) \quad \min_p \frac{1}{2} \|p - p_c\|_2^2$$

$$(19) \quad \text{s.t. } \min_p \frac{1}{2} \|J_2 p + f_2\|_2^2$$

$$(20) \quad \text{s.t. } \min_p \frac{1}{2} \|J_1 p + f_1\|_2^2,$$

where $p_c = 0$ is one possible choice.

The solution to (20) is given by $p = -J_1^+ f_1 + \mathcal{P}_{\mathcal{N}(J_1)} p_1$, where $\mathcal{P}_{\mathcal{N}(J_1)}$ is the orthogonal projection onto the null space of J_1 . When we substitute this into (19), the second minimization problem in (18) gives

$$p_1 = (J_2 \mathcal{P}_{\mathcal{N}(J_1)})^+ (J_2 J_1^+ f_1 - f_2) + \mathcal{P}_{\mathcal{N}(J_2)} p_2.$$

Since $p_c = \mathcal{P}_{\mathcal{N}(J)} p_c + \mathcal{P}_{\mathcal{R}(J^T)} p_c$ and $\mathcal{N}(J) = \mathcal{N}(J_1) \cap \mathcal{N}(J_2)$, we have

$$(21) \quad p = -J_1^+ f_1 + (J_2 \mathcal{P}_{\mathcal{N}(J_1)})^+ (J_2 J_1^+ f_1 - f_2) + \mathcal{P}_{\mathcal{N}(J)} p_c.$$

The three terms of p are contained in three orthogonal subspaces $\mathcal{R}(J_1^T)$, $\mathcal{N}(J_1) \cap \mathcal{R}(J^T)$, and $\mathcal{N}(J)$, respectively. In Figure 1 we have these three spaces together with two other important subspaces.

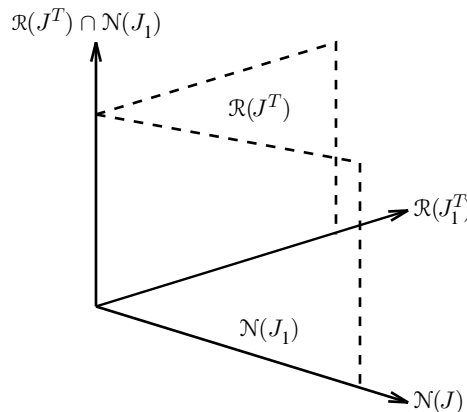


FIGURE 1. Decomposition of \mathbb{R}^n .

3.2. The asymptotic convergence rate. The asymptotic linear convergence rate is determined by the spectral radius of $\nabla(\hat{x} + p(\hat{x}))$; see [15]. However, we will start by using differentials to obtain a simple form of $d(x + p(x))$. Then this result is easily formulated with differentials and we are able to state and prove our main theorem in this section.

From (21) we have that the search direction p is composed of three mutually orthogonal directions as $p = p_1 + p_2 + p_3$ where

$$(22) \quad p_1 = -J_1^+ f_1,$$

$$(23) \quad p_2 = -M^+(f_2 - J_2 J_1^+ f_1),$$

$$(24) \quad p_3 = -\mathcal{P}_{\mathcal{N}(J)}(x - x_c),$$

and we have defined $M = J_2 \mathcal{P}_{\mathcal{N}(J_1)}$. We will use the following theorem to derive the differentials at \hat{x} .

Theorem 3.1. *If J is differentiable at x and of constant rank in a neighbourhood of x , then J^+ and $\mathcal{P}_{\mathcal{N}(J)}$ are differentiable at x and the differentials can be written*

$$d(J^+) = -J^+(dJ)J^+ + \mathcal{P}_{\mathcal{N}(J)}(dJ)^T (JJ^T)^+ + (J^T J)^+(dJ)^T \mathcal{P}_{\mathcal{N}(J^T)}$$

and

$$d(\mathcal{P}_{\mathcal{N}(J)}) = J^+ dJ \mathcal{P}_{\mathcal{N}(J)} - \mathcal{P}_{\mathcal{N}(J)}(dJ)^T J^{+T}.$$

Proof. See [17] and [7]. □

We will study the three last terms of the right-hand side of

$$(25) \quad d(x + p(x)) = dx + d(p_1) + d(p_2) + d(p_3)$$

separately.

The first of these becomes

$$(26) \quad d(p_1) = -J_1^+ J_1 dx - (J_1^T J_1)^+(dJ_1)^T \mathcal{P}_{\mathcal{R}(J_1^T)} f_1,$$

since $J_1^+ f_1 = 0$ at the solution. The differential of p_2 is more complicated. We have

$$d(p_2) = d(p_{2a}) + d(p_{2b}) = -M^+ d(f_2 - J_2 J_1^+ f_1) - d(M^+)(f_2 - J_2 J_1^+ f_1).$$

Furthermore,

$$\begin{aligned} d(p_{2a}) &= -M^+(J_2 dx - d(J_2)J_1^+ f_1 - J_2 d(J_1^+)f_1 - J_2 J_1^+ J_1 dx) \\ &= -M^+(M dx - J_2 d(J_1^+)f_1) \\ (27) \quad &= -\mathcal{P}_{\mathcal{R}(M^T)} dx + M^+ J_2 d(J_1^+)f_1, \end{aligned}$$

where we utilize that $J_1^+ f_1 = 0$. Moreover

$$\begin{aligned} d(p_{2b}) &= d(M^+)(f_2 - J_2 J_1^+ f_1) \\ (28) \quad &= -M^+ d(M)M^+ f_2 \end{aligned}$$

$$(29) \quad + \mathcal{P}_{\mathcal{N}(M)} d(M)(MM^T)^+ f_2$$

$$(30) \quad + (M^T M)^+ d(M)^T \mathcal{P}_{\mathcal{N}(M^T)} f_2.$$

The terms (28) and (29) are zero due to the first order KKT-condition. The term (30) becomes $(M^T M)^+ u$, where

$$(31) \quad u = (d(J_2) \mathcal{P}_{\mathcal{N}(J_1)})^T \mathcal{P}_{\mathcal{N}(M^T)} f_2$$

$$(32) \quad + J_1^+ (d(J_1)) \mathcal{P}_{\mathcal{N}(J_1)} J_2^T f_2$$

$$(33) \quad + \mathcal{P}_{\mathcal{N}(J_1)} d(J_1)^T J_1^{+T} J_2^T f_2.$$

The term (32) equals zero since $(J_2 \mathcal{P}_{\mathcal{N}(J_1)})^T f_2 = 0$ at the solution.

The final term in (25) becomes

$$(34) \quad d((\mathcal{P}_{\mathcal{N}(J)})(x - x_c)) = J^+(dJ)(\mathcal{P}_{\mathcal{N}(J)})(x - x_c)$$

$$(35) \quad - (\mathcal{P}_{\mathcal{N}(J)})(dJ)^T J^{+T}(x - x_c)$$

$$(36) \quad + \mathcal{P}_{\mathcal{N}(J)} dx.$$

Due to the first order conditions, (34) is zero.

To derive the local convergence rates, we will go from differentials to derivatives by $d(g(x)) = \nabla g(x) dx$, where $g(x) = x + p(x)$ and $d(J)^T f = (f \odot f'') dx$. The following theorem describes fully the asymptotic behaviour of the truncated method.

Theorem 3.2. *Assume that the $\{p_k\}$ are generated by solving (18) and that $x_{k+1} = x_k + p_k$ converges to \hat{x} . Then*

$$\limsup_{k \rightarrow \infty} \frac{\|x_{k+1} - \hat{x}\|}{\|x_k - \hat{x}\|} \leq \mathcal{K}$$

with

$$(37) \quad \mathcal{K} = \max\{\mathcal{K}_{f_1}, \mathcal{K}_{f_2}, \mathcal{K}_x\},$$

where

$$(38) \quad \mathcal{K}_{f_1} = \max_{v \in \mathcal{R}(J_1^T)} \frac{|v^T (f_1 \odot f_1'') v|}{v^T J_1^T J_1 v},$$

$$(39) \quad \mathcal{K}_{f_2} = \max_{v \in \mathcal{R}(M^T)} \frac{|v^T (f_2 \odot f_2'' + \lambda_1 \odot f_1'') v|}{v^T J_2^T J_2 v},$$

and

$$(40) \quad \mathcal{K}_x = \max_{v \in \mathcal{N}(J)} \frac{|v^T (\gamma \odot f'') v|}{v^T v}.$$

Proof. First observe that from (25), (26), (27), and (36) it follows that

$$dx - J_1^+ J_1 dx - \mathcal{P}_{\mathcal{R}(M^T)} dx - \mathcal{P}_{\mathcal{N}(J)} dx = 0.$$

The three directions p_1, p_2 , and p_3 are mutually orthogonal and can be treated independently of each other.

From (26) we get $\nabla(p_1) = -J_1^+ J_1 - (J_1^T J_1)^+(f_1 \odot f_1'')$, and the largest eigenvalue is given by (38).

The next component $\nabla(p_2)$ consists of three terms, (31), (33), and (27). The term (31) implies that $\nabla(p_{2b}) = (M^T M)^+(\mathcal{P}_{\mathcal{N}(J_1)}(f_2 \odot f_2'') + \mathcal{P}_{\mathcal{N}(J_1)}(\lambda_1 \odot f_1''))$, where $\lambda_1 = J_1^{+T} J_2^T f_2$.

□

3.3. A local convergence analysis. The asymptotic analysis above does not give any information about the actual relation between $x_{k+1} - \hat{x}$ and $x_k - \hat{x}$ or the influence of any second order information. In this section we derive other local results based on perturbation analysis of the linearized problem. We start by introducing a convenient formulation of the linearized problem and then use a perturbation analysis in order to state the local convergence results.

3.3.1. *The augmented system.* We begin by considering the linearization of (20) (skipping indices and arguments)

$$\begin{aligned} \min_{p \in \mathbb{R}^n} & \frac{1}{2} \|J_2 p + f_2\|_2^2 \\ \text{s.t.} & \quad J_1 p + f_1 = 0, \end{aligned}$$

where we initially do not assume rank-deficiency. In order to attain a suitable form of this problem, we introduce the Lagrange function

$$\mathcal{L}(p, \lambda_1) = \frac{1}{2} \|J_2 p + f_2\|_2^2 + \lambda_1^T (J_1 p + f_1).$$

For a critical point we will have

$$\begin{aligned} \frac{\partial \mathcal{L}}{\partial p} &= J_2^T (J_2 p + f_2) + J_1^T \lambda_1 = 0, \\ \frac{\partial \mathcal{L}}{\partial \lambda_1} &= J_1 p + f_1 = 0, \end{aligned}$$

or if we define $r_2 = J_2 p + f_2$,

$$(41) \quad \begin{bmatrix} 0 & 0 & J_1 \\ 0 & I & J_2 \\ J_1^T & J_2^T & 0 \end{bmatrix} \begin{bmatrix} \lambda_1 \\ r_2 \\ p \end{bmatrix} = \begin{bmatrix} f_1 \\ f_2 \\ 0 \end{bmatrix}.$$

Equation (41) is the augmented system for linear least squares (see [2]), and we formulate this as

$$(42) \quad S y = d,$$

where S and d are given in (41). We call the matrix S the system matrix. It is easily seen that S is rank-deficient if and only if J_1 or J is rank-deficient further motivating the approach taken.

We now turn to the connection between (42) and the minimization problem (18) in the case where S may be rank-deficient.

Theorem 3.3. *The solution p in the solution of the least norm problem*

$$(43) \quad \min_y \frac{1}{2} \|y - y_c\|_2^2$$

$$(44) \quad \text{s.t.} \min_y \frac{1}{2} \|S y - d\|_2^2,$$

where $y_c = [0; 0; p_c]$, is equivalent to the solution of the minimization problem (15–17). Moreover, the solution of (43)–(44) is given by $y = S^+ d + \mathcal{P}_{\mathcal{N}(S)} y_c$, where $\mathcal{P}_{\mathcal{N}(S)}$ is the projection on the null space of S .

Proof. First we prove the equivalence between the solution of (15)–(17) and p in (42). The inner-most minimization problem (17) can be solved by doing a complete orthogonal transformation of J_1 [2]; i.e.,

$$Q_1^T J_1 P_1 = \begin{bmatrix} L_{11} & 0 \\ 0 & 0 \end{bmatrix},$$

where L_{11} is lower triangular. If $p_1 = P_1^T p$ is partitioned into $[p_{11}; p_{12}]$ and $Q_1^T f_1 = [f_{11}; f_{12}]$, we get $p_{11} = -L_{11}^{-1} f_{11}$ and $p = P_{11} p_{11} + P_{12} p_{12}$. Note that $\mathcal{R}(P_{11}) = \mathcal{R}(J_1^T)$ and $\mathcal{R}(P_{12}) = \mathcal{N}(J_1)$. Our second minimization problem (16) becomes a problem in p_{12} and looks like

$$\min_{p_{12}} \frac{1}{2} \|J_{21} p_{11} + J_{22} p_{12} + f_2\|_2^2,$$

where we have defined

$$J_2 P_1 = \begin{bmatrix} J_{21} & J_{22} \end{bmatrix}.$$

Again we can use a complete orthogonal transformation

$$Q_2^T J_{22} P_2 = \begin{bmatrix} L_{22} & 0 \\ 0 & 0 \end{bmatrix},$$

where L_{22} is lower triangular. By defining

$$Q_2^T J_{21} P_2 = \begin{bmatrix} L_{21} \\ L_{31} \end{bmatrix}, \quad Q_2^T f_2 = \begin{bmatrix} f_{21} \\ f_{31} \end{bmatrix}, \quad p_{12} = P_2 \begin{bmatrix} \bar{p}_{21} \\ \bar{p}_{22} \end{bmatrix} = P_{21} \bar{p}_{21} + P_{22} \bar{p}_{22},$$

we get the problem

$$\min_{\bar{p}_{21}} \frac{1}{2} \left\| \begin{bmatrix} L_{21} \\ L_{31} \end{bmatrix} p_{11} + \begin{bmatrix} L_{22} \\ 0 \end{bmatrix} \bar{p}_{21} + \begin{bmatrix} f_{21} \\ f_{31} \end{bmatrix} \right\|_2^2$$

with the solution $\bar{p}_{21} = -L_{22}^{-1}(f_{21} - L_{21}p_{11})$. We rewrite this in a more compact form as

$$(45) \quad \begin{bmatrix} p_{11} \\ \bar{p}_{21} \end{bmatrix} = \begin{bmatrix} L_{11} & 0 \\ L_{21} & L_{22} \end{bmatrix}^{-1} \begin{bmatrix} -f_{11} \\ -f_{21} \end{bmatrix}.$$

Finally we have the outer-most minimization problem that must have a solution in $\mathcal{N}(J)$. Since $\mathcal{N}(J) = \mathcal{R}(P_{12}P_{22})$, we get

$$p = P_{11}p_{11} + P_{12}P_{21}\bar{p}_{21} + P_{12}P_{22}P_{22}^T P_{12}^T p_c$$

corresponding to p in (21).

Consider now the minimization problem (44) and perform the complete orthogonal transformations on J ; i.e.,

$$\begin{bmatrix} Q_1^T & 0 \\ 0 & Q_2^T \end{bmatrix} \begin{bmatrix} J_1 \\ J_2 \end{bmatrix} P_1 \begin{bmatrix} I_s & 0 \\ 0 & P_2 \end{bmatrix} = \begin{bmatrix} L_{11} & 0 & 0 \\ 0 & 0 & 0 \\ L_{21} & L_{22} & 0 \\ L_{31} & 0 & 0 \end{bmatrix}$$

and

$$\begin{bmatrix} Q_1^T & 0 \\ 0 & Q_2^T \end{bmatrix} \begin{bmatrix} f_1 \\ f_2 \end{bmatrix} = \begin{bmatrix} f_{11} \\ f_{12} \\ f_{21} \\ f_{22} \end{bmatrix}.$$

The problem (44) is transformed into

$$\frac{1}{2} \left\| \begin{bmatrix} 0 & & & L_{11} & 0 & 0 \\ & 0 & & 0 & 0 & 0 \\ & & I & L_{21} & L_{22} & 0 \\ & & & L_{31} & 0 & 0 \\ L_{11}^T & 0 & L_{21}^T & L_{31}^T & & \\ 0 & 0 & L_{22}^T & 0 & & \\ 0 & 0 & 0 & 0 & & \end{bmatrix} \begin{bmatrix} \lambda_{11} \\ \lambda_{12} \\ r_{11} \\ r_{12} \\ p_{11} \\ \bar{p}_{21} \\ p_3 \end{bmatrix} - \begin{bmatrix} f_{11} \\ f_{12} \\ f_{21} \\ f_{22} \\ 0 \\ 0 \\ 0 \end{bmatrix} \right\|_2^2.$$

We immediately get $r_{11} = 0$ and (45) is attained, again proving the first part of the theorem.

By substituting $z = y - y_c$, we want to find the minimum norm solution of $\min \|Sz - d + Sy_c\|_2$ that is given by $z = S^+(d - Sy_c)$. Therefore, we have $y = S^+d + (I - S^+S)y_c$ and $I - S^+S = \mathcal{P}_{\mathcal{N}(S)}$ gives the second result in the theorem. \square

3.4. A local convergence analysis using perturbation analysis. Having established the connection between the minimization problem and the minimum norm solution of the augmented system (42), we can use the perturbation analysis in [20] on (42) to analyze the local convergence behaviour more closely.

Assume that \hat{x} is a local minimum to our nonlinear problem (12). Write the pseudoinverse of the system matrix

$$S = \begin{bmatrix} 0 & 0 & J_1 \\ 0 & I & J_2 \\ J_1^T & J_2^T & 0 \end{bmatrix}$$

as (see [20] for details)

$$S^+ = \begin{bmatrix} H & B^T \\ B & -B_2 B_2^T \end{bmatrix},$$

where $B = [B_1, B_2]$. The special form $-B_2 B_2^T$ of the lower right block in S^+ is attained by looking more closely on $S^+ S$; see [20]. By expressing S^+ in the normal form (or any other rank revealing form) (see the proof of Theorem 3.3), it is easy to show that

$$\begin{bmatrix} \lambda_k \\ r_k \\ p_k \end{bmatrix} = -S^+ \begin{bmatrix} -f_k \\ 0 \end{bmatrix} + \begin{bmatrix} 0 \\ \mathcal{P}_{\mathcal{N}(J)}(x_k - x_c) \end{bmatrix}$$

is the solution to (43) and

$$p_k = -B_k f_k + \mathcal{P}_{\mathcal{N}(J_k)}(x_k - x_c) = [-B_k, \mathcal{P}_{\mathcal{N}(J_k)}] \begin{bmatrix} f_k \\ x_k - x_c \end{bmatrix} = [-B_k, \mathcal{P}_{\mathcal{N}(J_k)}] y_k$$

with an obvious definition for y_k .

The following lemma will be very useful.

Lemma 3.1. *Partition the matrix B in S^+ as $B = [B_1, B_2]$. Then*

$$[0, B_2] = B_2 B_2^T J^T, \quad B_1 = B_1 (J_1^+)^T J_1^T.$$

Moreover, we have that

$$[B, \mathcal{P}_{\mathcal{N}(J)}] \begin{bmatrix} J \\ I \end{bmatrix} = I;$$

i.e., $B J = \mathcal{P}_{\mathcal{R}(J^T)}$.

Proof. Since $S^+ S = I - \mathcal{P}_{\mathcal{N}(S)}$, we have

$$\begin{bmatrix} H_{11} & H_{12} & B_1^T \\ H_{12}^T & H_{22} & B_2^T \\ B_1 & B_2 & -B_2 B_2^T \end{bmatrix} \begin{bmatrix} 0 & 0 & J_1 \\ 0 & I & J_2 \\ J_1^T & J_2^T & 0 \end{bmatrix} = \begin{bmatrix} I - \mathcal{P}_{\mathcal{N}(J_1^T)} & 0 & 0 \\ 0 & I & 0 \\ 0 & 0 & I - \mathcal{P}_{\mathcal{N}(J)} \end{bmatrix}.$$

Pure identification gives the first and last part of the lemma. From the derivation of the solution in the linear case we know that $B_1 = J_1^+ - B_2 J_2 J_1^+$ and since $(J_1 J_1^+)^T$ is an orthogonal projection onto $\mathcal{R}(J_1)$, the second part of the lemma is true. \square

Define $q_k = x_k - \hat{x}$ as the quantity we are interested in. Using the Taylor expansion

$$y(\hat{x}) = y(x_k - q_k) = y_k - \begin{bmatrix} J_k \\ I \end{bmatrix} q_k + \begin{bmatrix} r_k \\ 0 \end{bmatrix}$$

where

$$r_k = \int_0^1 (1 - \tau) \begin{bmatrix} q_k^T f_1''(x_k - \tau q_k) q_k \\ \vdots \\ q_k^T f_m''(x_k - \tau q_k) q_k \end{bmatrix} d\tau,$$

we get

$$[B_k, \mathcal{P}_{\mathcal{N}(J_k)}] \hat{y} = [B_k, \mathcal{P}_{\mathcal{N}(J_k)}] y_k - q_k + B_k r_k.$$

Rearranging terms, we find that

$$(46) \quad q_{k+1} = q_k - B_k \hat{f} + \mathcal{P}_{\mathcal{N}(J_k)}(\hat{x} - x_c) + B_k r_k$$

and we want to relate this to q_k . The term $\mathcal{P}_{\mathcal{N}(J_k)}(\hat{x} - x_c)$ has been analyzed in [5] and we have

$$\mathcal{P}_{\mathcal{N}(J_k)}(\hat{x} - x_c) = \mathcal{P}_{\mathcal{N}(J)}(J_k - \hat{J})^T ((\hat{J}^+)^T (\hat{x} - x_c)).$$

The remaining part is then $B_k \hat{f}$ and in order to simplify the notation, we skip the index k .

Lemma 3.2. *If $\delta J = J - \hat{J}$, then*

$$B \hat{f} = B_1 (J_1^+)^T (\delta J)^T \hat{f}_1 + B_2 B_2^T [(\delta J_1)^T \hat{\lambda}_1 + (\delta J_2)^T \hat{f}_2].$$

Proof. We have

$$B \hat{f} = [B_1, B_2] \hat{f} = B_1 \hat{f}_1 + [0, B_2] \begin{bmatrix} \hat{\lambda}_1 \\ \hat{f}_2 \end{bmatrix}.$$

From Lemma 3.1 and $\hat{J}_1^T \hat{f}_1 = 0$ we have

$$B_1 \hat{f}_1 = B_1 (J_1^+)^T J_1^T \hat{f}_1 = B_1 (J_1^+)^T (\delta J_1)^T \hat{f}_1.$$

In the same way, from Lemma 3.1 and $\hat{J}_1^T \hat{\lambda}_1 + \hat{J}_2^T \hat{f}_2 = 0$ we find the second term of $B \hat{f}$ in the lemma. □

We immediately get the following theorem describing the local convergence of our truncated Gauss-Newton method.

Theorem 3.4. *Assume that the $\{p_k\}$ are generated by solving (43) or (18) and that $x_{k+1} = x_k + p_k$ converges to \hat{x} . If \hat{x} is the solution of (12) and $\hat{\lambda}_1$ is the vector λ_1 from (43) at \hat{x} , then*

$$q_{k+1} = K_{GN} q_k + B_k r_k,$$

where $q_k = x_k - \hat{x}$,

$$K_{GN} = -B_1 (J_1^+)^T (\bar{H}_1^T \odot \hat{f}_1) - B_2 B_2^T (\bar{H}_1^T \odot \hat{\lambda}_1 + \bar{H}_2^T \odot \hat{f}_2) + \mathcal{P}_{\mathcal{N}(J)}(\bar{H} \odot \hat{\gamma}),$$

$\hat{\gamma} = (\hat{J}^+)^T (\hat{x} - x_c)$, and

$$\bar{H}_1 = \begin{bmatrix} \int_0^1 f_1''(\hat{x} + \tau q_k) d\tau \\ \vdots \\ \int_0^1 f_p''(\hat{x} + \tau q_k) d\tau \end{bmatrix}, \quad \bar{H}_2 = \begin{bmatrix} \int_0^1 f_{p+1}''(\hat{x} + \tau q_k) d\tau \\ \vdots \\ \int_0^1 f_m''(\hat{x} + \tau q_k) d\tau \end{bmatrix}$$

with $\bar{H} = [\bar{H}_1; \bar{H}_2]$.

Proof. The theorem is evident from the earlier discussion if we use the fact that

$$(J_k - \hat{J})^T = \int_0^1 [f_1''(\hat{x} + \tau q_k)q_k, \dots, f_m''(\hat{x} + \tau q_k)q_k] d\tau.$$

□

4. TIKHONOV REGULARIZATION

For an unconstrained ill-posed nonlinear least squares problem it is possible to use the regularized problem

$$\min_x \frac{1}{2} \|f(x)\|_2^2 + \frac{1}{2} \mu^2 \|x - x_c\|_2^2.$$

If this problem is linearized, a Gauss-Newton method is attained where the search direction p can be found by solving the linear problem

$$(47) \quad \min_x \frac{1}{2} \|Jp + f\|_2^2 + \frac{1}{2} \mu^2 \|p - p_c\|_2^2.$$

The augmented system corresponding to (47) is

$$(48) \quad \begin{bmatrix} I & J \\ -J^T & \mu^2 I \end{bmatrix} \begin{bmatrix} r \\ p \end{bmatrix} = \begin{bmatrix} -f \\ \mu^2 p_c \end{bmatrix}.$$

Note the skew symmetric structure of the system matrix. By using the SVD of J , it is easy to show that if $p(\mu)$ solves (47), then $p(0) = \lim_{\mu \rightarrow 0} p(\mu)$ solves the least norm problem

$$\begin{aligned} & \min_p \frac{1}{2} \|p - p_c\|_2^2 \\ \text{s.t. } & \min_p \frac{1}{2} \|Jp + f\|_2^2. \end{aligned}$$

In other words, for an *exactly* rank-deficient matrix J we get the solution to the corresponding truncated problem and locally a Tikhonov method has exactly the same properties as a truncated Gauss-Newton method.

Unfortunately, it is not that easy in the constrained case. One possible generalization of the Tikhonov regularization in (47) is to consider the penalty problem

$$\min_x \frac{1}{2} \frac{1}{\mu^2} \|f_1(x)\|_2^2 + \frac{1}{2} \|f_2(x)\|_2^2 + \frac{1}{2} \mu^2 \|x - x_c\|_2^2.$$

A Gauss-Newton method based on this formulation gives the weighted linear least squares problem

$$(49) \quad \min_x \frac{1}{2} \frac{1}{\mu^2} \|J_1 p + f_1\|_2^2 + \frac{1}{2} \|J_2 p + f_2\|_2^2 + \frac{1}{2} \mu^2 \|p - p_c\|_2^2.$$

The augmented system corresponding to this regularized problem looks like

$$(50) \quad \begin{bmatrix} \frac{1}{\mu^2} I & 0 & J_1 \\ 0 & I & J_2 \\ -J_1^T & -J_2^T & \mu^2 I \end{bmatrix} \begin{bmatrix} \lambda_1 \\ r_2 \\ p \end{bmatrix} = \begin{bmatrix} f_1 \\ f_2 \\ \mu^2 p_c \end{bmatrix}.$$

We have the following theorem.

Theorem 4.1. *Let $p(\mu)$ be the solution of (49). Then $\lim_{\mu \rightarrow 0} p(\mu) = p(0)$ where $p(0)$ is the solution of (18).*

Proof. Consider the augmented system in (50) and make a transformation of J to the normal form just as in the proof of Theorem 3.3. Then we get

$$\begin{bmatrix} \frac{1}{\mu^2}I_{p-s} & & & & L_{11} & 0 & 0 \\ & \frac{1}{\mu^2}I_{p-s} & & & 0 & 0 & 0 \\ & & I & & L_{21} & L_{22} & 0 \\ -L_{11}^T & 0 & -L_{21}^T & -L_{31}^T & L_{31} & 0 & 0 \\ 0 & 0 & -L_{22}^T & 0 & \mu^2 I & & \\ 0 & 0 & 0 & 0 & & \mu^2 I & \end{bmatrix} \begin{bmatrix} \lambda_{11} \\ \lambda_{12} \\ r_{11} \\ r_{12} \\ p_{11} \\ \bar{p}_{21} \\ p_3 \end{bmatrix} = \begin{bmatrix} -f_{11} \\ -f_{12} \\ -f_{21} \\ -f_{22} \\ \mu^2 c_1 \\ \mu^2 c_2 \\ \mu^2 c_3 \end{bmatrix}.$$

If we solve for

$$(51) \quad \lambda_{12} = \frac{1}{\mu^2}f_{12},$$

then we are left with a problem whose solution has a well-defined limit. We see that λ_{11} tends to the Lagrange parameters for the constraints $L_{11}p_1 = -f_{11}$, $\lim_{\mu \rightarrow 0} r_{12} = 0$, $\lim_{\mu \rightarrow 0} [p_{11}; \bar{p}_{21}]$ is the solution to (45), and $\lim_{\mu \rightarrow 0} p_3$ will be such that $\|p - p_c\|_2$ is minimized just as for p_3 in (18) (compare to the unconstrained Tikhonov regularized case). \square

This looks very promising but the approach seems inappropriate in a Gauss-Newton method for *exactly* rank-deficient problems (in the almost rank-deficient case this form of regularization is quite possible). The reason is very simple. Assume that we use the weighted problem (49) with a very small μ on a rank-deficient problem. Then the local convergence rate (see [9]) will be determined by

$$q_{k+1} = -B_k M B_k^T (\bar{H} \odot \hat{\lambda}) q_k + \mathcal{P}_{N(J_k)}(\bar{H} \odot \hat{\gamma}) q_k + 1/2 B_k r_k$$

where $q_k = x_k - \hat{x}$ and the other quantities are defined in Theorem 3.4. Moreover, $[\lambda_1; r_2]$ from (50) will not be a good estimate of the Lagrange parameter $\hat{\lambda}$. The vector λ_1 will be very large because λ_{12} is large in (51) (unless f_{12} is very small which is unlikely). Therefore, we cannot generally get convergence with this kind of method and we will show this in the computational experiments.

There is another way to regularize the linearized problem that will make the Gauss-Newton method locally convergent. Consider the augmented system (43). This is no more than the least norm problem to an underdetermined linear system of equations and it is perfectly adequate, at least from a theoretical point of view, to use the regularized problem

$$(52) \quad \min_y \frac{1}{2} \|S y - d\|_2^2 + \frac{1}{2} \mu^2 \|y - y_c\|_2^2,$$

where $y_c = [0; 0; p_c]$ in order to correspond to (18). A more interesting formulation of problem (52) is attained by using the augmented system as a linear least squares problem, i.e.

$$\min_{\lambda_1, r_2, p} \frac{1}{2} \left\| \begin{bmatrix} 0 & 0 & J_1 \\ 0 & I & J_2 \\ J_1^T & J_2^T & 0 \\ \mu I & 0 & 0 \\ 0 & \mu I & 0 \\ 0 & 0 & \mu I \end{bmatrix} \begin{bmatrix} \lambda_1 \\ r_2 \\ p \end{bmatrix} + \begin{bmatrix} f_1 \\ f_2 \\ 0 \\ 0 \\ 0 \\ -\mu p_c \end{bmatrix} \right\|_2^2.$$

It is clearly seen how J_1^T and J are regularized quite similarly to the unconstrained case in (48). We have the following theorem showing that (52) gives the correct least norm solution.

Theorem 4.2. *If $y(\mu)$ is the solution to (52), then $\lim_{\mu \rightarrow 0} y(\mu) = y(0)$ where $y(0)$ is the solution to (43). Moreover, $p(0)$ in $y(0)$ is the solution to (15)–(17).*

Proof. It is well known that

$$\lim_{\mu \rightarrow 0} (S^T S + \mu^2 I)^{-1} = S^+$$

and the normal equations for (52) will, in the limit, give the solution $y(0) = S^+ d + \text{diag}(0, 0, \mathcal{P}_{N(J)} p_c)$. \square

This theorem gives a possibility to construct a Tikhonov method for the nonlinear problem (12). The local properties as $\mu \rightarrow 0^+$ of the method are the same as for the truncation method. It is possible to do a more detailed analysis of the local convergence properties for a small $\mu > 0$ but we will not do this here. Moreover, there are difficulties concerning efficiency and global convergence that are far from trivial.

5. COMPUTATIONAL EXPERIMENTS

In this section the theoretical results for the three different local approaches proposed will be verified by computational experiments. We will use artificial problems for which we can determine the local behaviour of the problems.

5.1. Generation of artificial problems. In [8] it is shown that (12) can be divided into three minimization problems which in turn can be analyzed separately. The following lemma describes the actual form of f and f_1 when we assume that J and J_1 have constant rank in a neighbourhood of the critical point.

Lemma 5.1. *Assume that $\text{rank}(J_1) = s \leq m_1$ and $\text{rank}(J) = r \leq n$ in a neighbourhood of a critical point to (12). Then there exist functions $h_1 : \mathbb{R}^r \rightarrow \mathbb{R}^{m_1}$, $h_2 : \mathbb{R}^r \rightarrow \mathbb{R}^{m_2}$, $z : \mathbb{R}^n \rightarrow \mathbb{R}^r$ such that $f_2(x) = h_2(z(x))$ and $f_1(x) = h_1(z(x))$. The Jacobians of $h = [h_1; h_2]$ and z are of full rank in a neighbourhood of a critical point to the constrained problem (12). Moreover, there exist functions $c : \mathbb{R}^s \rightarrow \mathbb{R}^p$, $d : \mathbb{R}^r \rightarrow \mathbb{R}^s$, whose Jacobians are of full rank, such that $h_1(z) = c(d(z))$.*

Using the lemma, we can formulate the constrained problem (12) as

$$(53) \quad \min_x \frac{1}{2} \|x - x_c\|_2^2$$

$$(54) \quad \text{s.t.} \quad \min_x \frac{1}{2} \|h_2(z(x))\|_2^2$$

$$(55) \quad \text{s.t.} \quad \min_x \frac{1}{2} \|c(d(z(x)))\|_2^2$$

around any critical point where J_1 or J are rank-deficient. Problem (53)–(55) can be solved at three levels. First we have \hat{d} as the solution of

$$(56) \quad \min_d \frac{1}{2} \|c(d)\|_2^2$$

and the inner minimization problem (54) becomes

$$(57) \quad \min_x \frac{1}{2} \|h_2(z(x))\|_2^2$$

$$(58) \quad \text{s.t. } d(z(x)) = \hat{d}.$$

This problem decouples into

$$(59) \quad \min_z \frac{1}{2} \|h_2(z)\|_2^2$$

$$(60) \quad \text{s.t. } d(z) = \hat{d}$$

with a solution \hat{z} and the final problem is

$$(61) \quad \min_x \frac{1}{2} \|x - x_c\|_2^2$$

$$(62) \quad \text{s.t. } z(x) = \hat{z}.$$

When constructing a problem with known local properties, it is sufficient to consider the second order Taylor expansions of h, z, c and d . The functions f_1 and f_2 are then attained from the chain rule. We have

$$(63) \quad z(\hat{x} + \Delta x) = z(\hat{x}) + E\Delta x + \frac{1}{2} \begin{pmatrix} \Delta x^T z_1'' \Delta x \\ \vdots \\ \Delta x^T z_r'' \Delta x \end{pmatrix} + o(\|\Delta x\|^2),$$

$$(64) \quad d(\hat{z} + \Delta z) = d(\hat{z}) + D\Delta z + \frac{1}{2} \begin{pmatrix} \Delta z^T d_1'' \Delta z \\ \vdots \\ \Delta z^T d_s'' \Delta z \end{pmatrix} + o(\|\Delta z\|^2),$$

and

$$(65) \quad c(\hat{d} + \Delta d) = c(\hat{d}) + C\Delta d + \frac{1}{2} \begin{pmatrix} \Delta d^T c_1'' \Delta d \\ \vdots \\ \Delta d^T c_p'' \Delta d \end{pmatrix} + o(\|\Delta d\|^2),$$

which implies that

$$\begin{aligned} d(z(\hat{x} + \Delta x)) &= d \left(z(\hat{x}) + E\Delta x + \frac{1}{2} \begin{pmatrix} \Delta x^T z_1'' \Delta x \\ \vdots \\ \Delta x^T z_r'' \Delta x \end{pmatrix} \right) \\ &= d(z(x)) + DE\Delta x + \frac{1}{2} D \begin{pmatrix} \Delta x^T z_1'' \Delta x \\ \vdots \\ \Delta x^T z_r'' \Delta x \end{pmatrix} \\ &\quad + \frac{1}{2} \begin{pmatrix} (E\Delta x)^T d_1'' (E\Delta x) \\ \vdots \\ (E\Delta x)^T d_s'' (E\Delta x) \end{pmatrix}, \end{aligned}$$

and

$$\begin{aligned}
 c(d(z(\hat{x} + \Delta x))) &= c(d(z(\hat{x}))) + CDE\Delta x + \frac{1}{2}CD \begin{pmatrix} \Delta x^T z_1'' \Delta x \\ \vdots \\ \Delta x^T z_r'' \Delta x \end{pmatrix} \\
 &+ \frac{1}{2}C \begin{pmatrix} (E\Delta x)^T d_1''(E\Delta x) \\ \vdots \\ (E\Delta x)^T d_s''(E\Delta x) \end{pmatrix} + \frac{1}{2} \begin{pmatrix} (E\Delta x)^T c_1''(E\Delta x) \\ \vdots \\ (E\Delta x)^T c_p''(E\Delta x) \end{pmatrix}.
 \end{aligned}$$

We define $f_1(x) = c(d(z(x)))$ and get $J_1 = CDE$. In a similar way we can define $f_2(x) = h_2(z(x))$ and $J_2 = A_2E$, with $h_2 \in \mathbb{R}^q$ and $A_2 \in \mathbb{R}^{q \times n}$.

The main advantage of generating artificial test problems is that we create problems with exactly known local properties. Thus, for the present constrained problem we can determine the curvatures, $\mathcal{K}_{f_1}, \mathcal{K}_{f_2}$ and \mathcal{K}_x for each problem $f_1(x) = c(d(z(x)))$ and $f_2(x) = h_2(z(x))$.

A new test problem is generated by the six steps below. The inputs to the generator are the problem sizes m_1, m_2, n , the ranks s, r , and the desired curvatures $\mathcal{K}_{f_1}, \mathcal{K}_{f_2}$, and \mathcal{K}_x . The outputs from the generator are the functions f_1, f_2 and their first and second derivatives. Since the solution \hat{x} is generated and known, it is easy to choose suitable starting points and to measure the convergence rate. The matrices and vectors are generated randomly in the interval $[-1, 1]$. The steps 2–4 are for the first order KKT-conditions and the steps 5–6 create second derivatives such that the second order conditions are fulfilled; see [8].

- (1) Generate the Jacobians $A \in \mathbb{R}^{m_2 \times r}, C \in \mathbb{R}^{m_1 \times s}, D \in \mathbb{R}^{s \times r}$ and $E \in \mathbb{R}^{r \times n}$ giving $J_1 = CDE$ and $J_2 = A_2E$.
- (2) To generate the first order condition $J_1^T f_1 = 0$, we choose a residual for the constraints c such that $C^T c = 0$. This means that $c(0) = c$ in (65) and that $d(0) = d$ in (64).
- (3) To fulfill the second first order condition $J_2^T f_2 + J_1^T \lambda_1 = 0$, we generate λ_1 and f_2 such that $\mathcal{P}_{\mathcal{R}(J_1^T)} J_2^T f_2 = -J_1^T \lambda_1$ and $\mathcal{P}_{\mathcal{N}(J_1)} J_2^T f_2 = 0$.
- (4) Generate the center x_c and the Lagrange parameter λ . Take $\hat{x} = x_c + S^T A^T \lambda$. Thus, $z(0) = 0$ in (63).
- (5) Generate the symmetric second derivatives of z, c, d and h .
- (6) Solve the generalized eigenvalue problems (38)–(40) and scale the second derivatives to get the wanted curvatures and the expected convergence rate.

5.2. Test results. We have made a test with $m_2 = 16, m_1 = 4, s = 2, r = 5, n = 7$ and $\mathcal{K}_{f_1} = 0.3, \mathcal{K}_{f_2} = 0.1$, and $\mathcal{K}_x = 0.2$. Thus, $\mathcal{K} = 0.3$ and we can expect a linear convergence rate equal to 0.3. The actual convergence rate is computed as

$$\rho = \frac{\|x_{k+1} - \hat{x}\|_2}{\|x_k - \hat{x}\|_2}$$

during the iterations.

The results for the truncated method are given in Table 1. It is apparent that the method behaves as predicted by our analysis. There is no problem with local convergence and all three separate directions, p_1, p_2, p_3 , decrease almost equally fast. For this test problem the actual convergence rate becomes almost identical to the predicted one (0.3).

TABLE 1. This table gives the results for the truncated method. The three leftmost columns show how the norm of the steps decreases towards zero. The fourth and fifth columns verify that the iterates go to the correct solution and that the actual convergence rate tends to the predicted rate, respectively.

$\ p_1\ $	$\ p_2\ $	$\ p_3\ $	$\ x - \hat{x}\ $	ϱ
3.3486e-05	8.1224e-05	7.4975e-06	8.3818e-05	2.7051e-05
1.0040e-05	1.1180e-04	6.1877e-06	2.9516e-05	3.5214e-01
3.0107e-06	3.7127e-05	8.4157e-06	8.6793e-06	2.9406e-01
9.0302e-07	1.0960e-05	2.6591e-06	2.6351e-06	3.0360e-01
2.7086e-07	3.3173e-06	7.9964e-07	7.8795e-07	2.9902e-01
8.1250e-08	9.9267e-07	2.4036e-07	2.3663e-07	3.0031e-01
2.4373e-08	2.9803e-07	7.2081e-08	7.0956e-08	2.9986e-01
7.3116e-09	8.9374e-08	2.1623e-08	2.1287e-08	3.0000e-01

TABLE 2. The results for the Tikhonov method based on (49). At the beginning of the iterations the method seems to converge but then it diverges.

$\ \lambda - \hat{\lambda}\ $	$\ f - \hat{f}\ $	$\frac{\ \mathcal{P}_{\mathcal{N}(J)}^T(x-x_c)\ }{\ (x-x_c)\ }$	ϱ	μ
2.3990e+00	3.2572e-03	1.3375e-01	1.5334e+00	2.5000e-02
4.3825e+00	2.2501e-01	7.1319e-02	7.8218e-01	6.2500e-03
1.6212e+00	3.7991e-02	1.2255e-01	1.2159e+00	1.5625e-03
4.4842e-01	2.6574e-03	9.5353e-02	8.1887e-01	3.9063e-04
2.4020e-01	1.6006e-04	4.2653e-02	9.6258e-01	9.7656e-05
4.5417e-03	4.8893e-06	5.2934e-02	9.7717e-01	2.4414e-05
1.0010e-02	4.7836e-08	8.9964e-02	9.9082e-01	6.1035e-06
1.9429e-02	2.6003e-06	1.6048e-01	9.4481e-01	1.5259e-06

The next test is performed using the Tikhonov method based on (49) where we choose $\mu_{k+1} = \mu_k/4$. The information from the iterations is shown in Table 2. As the theory reveals, it is not possible to get convergence, although there is some progress for a few iterations.

The next method to test is the modified Tikhonov method described in the end of Section 4; see (52). Now the convergence is much better, as shown in Table 3. The method converges with the actual convergence rate not far from the theoretical. However, the way of decreasing μ is not obvious. It is certainly possible to decrease μ so that an even greater agreement with the theory is achieved.

Finally, we have tested the truncated method for a larger problem using the dimensions $m_2 = 300, m_1 = 200, s = 20, r = 80$, and $n = 100$. The problem generated has the same curvatures as in the previous tests. Thus $\mathcal{K} = 0.3$, and we can expect a linear convergence rate equal to 0.3. As is shown in Table 4, the method converges with a rate rather close to the expected. Since this problem is larger, more iteration steps are required to get a rate very close to 0.3. This test shows that our proposed truncated method is able to solve medium-size problems efficiently and also large problems if sufficient computing resources are available.

TABLE 3. The results using (52). The method converges with a rate close to the predicted.

$\ \lambda - \hat{\lambda}\ $	$\ f - \hat{f}\ $	$\frac{\ \mathcal{P}_{\mathcal{N}(J)}(x-x_c)\ }{\ (x-x_c)\ }$	ϱ	μ
7.2191e-01	3.0416e-03	2.3702e-03	8.1085e-02	2.5000e-02
4.7359e-01	2.9494e-03	1.8117e-02	9.0462e-01	6.2500e-03
1.5348e-01	2.3147e-04	2.1066e-02	1.2731e-01	1.5625e-03
3.7197e-02	2.1237e-05	3.9038e-03	5.9200e-01	3.9063e-04
4.5656e-03	9.4829e-07	4.1990e-04	1.4707e-01	9.7656e-05
2.0869e-03	8.0415e-08	5.6042e-04	3.8261e-01	2.4414e-05
5.3742e-04	7.1754e-09	4.4660e-05	2.6717e-01	6.1035e-06
1.7114e-04	6.4529e-10	3.4224e-05	3.0841e-01	1.5259e-06

TABLE 4. The results using the truncated method for a larger problem.

$\ \lambda - \hat{\lambda}\ $	$\ f - \hat{f}\ $	$\frac{\ \mathcal{P}_{\mathcal{N}(J)}(x-x_c)\ }{\ (x-x_c)\ }$	ϱ	μ
5.6141e-06	1.1475e-05	9.6687e-06	6.9204e-06	3.1833e-01
1.7514e-06	4.1997e-06	2.4801e-06	1.8938e-06	2.7366e-01
4.9831e-07	9.9228e-07	7.2683e-07	5.9346e-07	3.1336e-01
1.5484e-07	3.4532e-07	1.9981e-07	1.6924e-07	2.8518e-01
4.5084e-08	9.0651e-08	5.9999e-08	5.2195e-08	3.0840e-01
1.3820e-08	2.9808e-08	1.7203e-08	1.5206e-08	2.9134e-01
4.0767e-09	8.2799e-09	5.1782e-09	4.6413e-09	3.0322e-01
1.2385e-09	2.6265e-09	1.5166e-09	1.3681e-09	2.9778e-01

6. CONCLUSIONS AND FUTURE WORK

We have considered local properties for the Gauss-Newton method on rank-deficient nonlinear least squares problems with rank-deficient nonlinear constraints.

The local convergence properties for a truncated Gauss-Newton method is well understood. It seems quite possible to construct a Gauss-Newton method that has global convergence as well as fast local convergence.

The Tikhonov regularization based on the least norm problem for the augmented system may be used for rank-deficient problems. Moreover, this approach seems suitable also in the case of an ill-posed problem where the Jacobians are almost (not exactly) rank-deficient. Exciting future work could be to explore this Tikhonov regularization. In the unconstrained case it is necessary to have a clear gap in the singular values in order to be able to analyze the problem properly. A similar assumption is most probably needed in the constrained case even if the matter is more complex. Other important and difficult questions to be answered are the choice of regularization parameter, merit function and an efficient solution of the linear least squares problem (52).

REFERENCES

1. D. Bates and D. Watts, *Nonlinear regression analysis and its applications*, John Wiley, 1988. MR **92f**:62002
2. Å. Björck, *Numerical methods for least squares problems*, SIAM, 1996. MR **97g**:65004

3. L. Elden, *Algorithms for the regularization of ill conditioned least squares problems*, BIT **17** (1977), 134–145. MR **57**:14541
4. ———, *A weighted pseudoinverse, generalized singular values and constrained least squares problems*, BIT **22** (1982), 487–502. MR **84g**:65048
5. J. Eriksson, *Optimization and regularization of nonlinear least squares problems*, Tech. Report UMINF 96.09 (Ph.D. Thesis), Dept. of Comp. Science, Umeå University, Umeå, Sweden, 1996.
6. J. Eriksson, M. E. Gulliksson, P. Lindström, and P.-Å. Wedin, *Regularization tools for training feed-forward neural networks part I: Theory and basic algorithms*, Tech. Report UMINF 96.05, Dept. of Comp. Science, 1996.
7. G. H. Golub and V. Pereyra, *The differentiation of pseudoinverses and nonlinear least squares problems whose variables separate*, SIAM J. Num. Anal. **10** (1973), 413–432. MR **49**:1753
8. M. E. Gulliksson, *KKT-conditions for exactly rank deficient nonlinear least squares with exactly rank deficient nonlinear constraints*, Journal of Optimization Theory and Application (JOTA) **100** (1999), no. 1, 145–160. MR **99m**:93057
9. M. E. Gulliksson, I. Söderkvist, and P.-Å. Wedin, *Algorithms for constrained and weighted nonlinear least squares*, SIAM J. Opt. **7** (1997), no. 1, 208–224. MR **97m**:90063
10. P.C. Hansen, *Rank-deficient and discrete ill-posed problems. Numerical aspects of linear inversion*, SIAM, Philadelphia, 1997. MR **99a**:65037
11. R. J. Hanson and C. L. Lawson, *Solving least squares problems*, Prentice Hall, Englewood Cliffs, N. J., 1974. MR **51**:2270
12. Sabine Van Huffel and Joos Vandewalle, *The total least squares problem - computational aspects and analysis*, SIAM, 1991. MR **93b**:65001
13. P. Lindström and P.-Å. Wedin, *Methods and software for nonlinear least squares problems*, Tech. Report UMINF-133.87, Inst. of Info. Proc., Univ. of Umeå, Umeå, Sweden, 1988.
14. J. J. More, *The Levenberg-Marquardt algorithm: implementation and theory*, Proceedings of the 1977 Dundee conference on numerical analysis (Berlin, Heidelberg, New York, Tokyo) (G. A. Watson, ed.), Lecture notes in mathematics 630, Springer Verlag, 1978, pp. 105–116. MR **58**:3446
15. J. M. Ortega and W. C. Rheinboldt, *Iterative solution of nonlinear equations in several variables*, Academic Press, New York, 1970. MR **42**:8686
16. H. F. Walker, *Newton-like methods for underdetermined systems*, Lectures in Applied Mathematics **26** (1990), 679–699. MR **91h**:65086
17. P.-Å. Wedin, *Perturbation theory for pseudo-inverses*, BIT **13** (1973), 217–232. MR **49**:1755
18. ———, *Notes on the constrained linear least squares problem. A new approach based on generalized inverses*, Technical Report UMINF 75.79, Inst. of Info. Proc., Univ. of Umeå, 1979.
19. ———, *Perturbation theory and condition numbers for generalized and constrained linear least squares problems*, Tech. Report UMINF 125.85, Inst. of Info. Proc., Univ. of Umeå, Umeå, Sweden, 1985.
20. ———, *On the use of a quadratic merit function for constrained nonlinear least squares*, Tech. report, Inst. of Info. Proc., Univ. of Umeå, Umeå, Sweden, 1987.

DEPARTMENT OF COMPUTING SCIENCE, UMEÅ, SWEDEN

E-mail address: `jerry@cs.umu.se`

DEPARTMENT OF ENGINEERING, PHYSICS, AND MATHEMATICS, MID-SWEDEN UNIVERSITY,
SUNDSVALL, SWEDEN

E-mail address: `marten@fmi.mh.se`