SIAM J. Numer. Anal.
Vol. 49, No. 3, pp. 1244–1266

RANK-DEFICIENT NONLINEAR LEAST SQUARES PROBLEMS AND SUBSET SELECTION*

I. C. F. IPSEN†, C. T. KELLEY†, and S. R. POPE‡

Abstract. We examine the local convergence of the Levenberg–Marquardt method for the solution of nonlinear least squares problems that are rank-deficient and have nonzero residual. We show that replacing the Jacobian by a truncated singular value decomposition can be numerically unstable. We recommend instead the use of subset selection. We corroborate our recommendations by perturbation analyses and numerical experiments.

Key words. nonlinear least squares, Levenberg–Marquardt algorithm, subset selection, singular value decomposition

AMS subject classifications. 15A18, 65F20, 65F22, 65F15, 65F35, 65H10, 65K10, 65L09

DOI. 10.1137/090780882

1. Introduction.

The purpose of this paper is to show how the accuracy of the Levenberg–Marquardt trust-region algorithm for nonlinear least squares problems can be compromised if the numerical rank of the Jacobians is ill-defined and if there are errors in the evaluations of residuals and Jacobians. We argue that the accepted practice of replacing a rank-deficient Jacobian by a truncated singular value decomposition (SVD) should be discontinued because the truncated SVD is numerically unstable, that is, it can produce a unnecessary loss of accuracy due to the computation of sensitive singular vectors. Instead we recommend that the list of variables be pruned by selecting a judiciously chosen set of columns of the Jacobian from a subset selection method.

Our paper is motivated by applications to cardiovascular modeling [7, 10, 17, 18]. In this application the model parameters were nonlinearly dependent, and the model was complicated enough so that dependency could not be eliminated analytically. This dependency was the cause of rank-deficiency. The resulting nonlinear least squares problem was rank-deficient and a truncated SVD failed to solve the problem. However, subset selection methods succeeded. We present perturbation analyses to explain these observations.

1.1. The problem.

The unconstrained nonlinear least squares problem for a function $R : \mathbb{R}^M \to \mathbb{R}^N$ with $M \geq N$ is

$$\min_p f(p), \qquad \text{where} \quad f(p) = \frac{1}{2} R(p)^T R(p) = \frac{1}{2}\|R(p)\|^2. \tag{1.1}$$

Here the superscript T denotes the transpose and $\|\cdot\|$ the Euclidean norm.

One way to solve nonlinear least squares problems is by a Gauss–Newton method.

*Received by the editors December 22, 2009; accepted for publication (in revised form) April 4, 2011; published electronically June 23, 2011. This work was partially supported by National Science Foundation grants CDI-0941253 and DMS-0707220, USACE grants W912HZ-10-P-0221 and W912HZ-10-P-0256, and Army Research Office grants W911NF-06-1-0412, and W911NF-07-1-0112.
http://www.siam.org/journals/sinum/49-3/78088.html

†Department of Mathematics, North Carolina State University, Box 8205, Raleigh, NC 27695-8205 (ipsen@ncsu.edu, Tim_Kelley@ncsu.edu).

‡SAS Institute, 100 SAS Campus Drive, Cary, NC 27513 (scott.pope@gmail.com).

Let J be the Jacobian of R and denote the gradient of f by

$$g(p) = \nabla f(p) = J(p)^T R(p). \tag{1.2}$$

Given a current approximation p_c to a solution p^*, a Gauss–Newton iteration computes a new approximation p_+ from

$$p_+ = p_c - \left(J(p_c)^T J(p_c)\right)^{-1} g(p_c).$$

Assume that $J(p^*)$ has full column rank. If the residual is zero, i.e., $R(p^*) = 0$, and p_c is close enough to p^*, then [13] the Gauss–Newton iteration converges locally q-quadratically to p^*,

$$\|p_+ - p^*\| = \mathcal{O}\left(\|p_c - p^*\|^2\right).$$

If the residual is nonzero but small, and p_c is close to p^*, then the Gauss–Newton iteration still converges fast for sufficiently small $f(p^*)$, because

$$\|p_+ - p^*\| = \mathcal{O}\left(\|p_c - p^*\|^2 + \|R(p^*)\|\|p_c - p^*\|\right).$$

However, if p_c is far from p^*, then the Gauss–Newton method may not converge. The Levenberg–Marquardt iteration is one way to obtain convergence from points that are far from p^*.

In this paper we are concerned with the local convergence of the Levenberg–Marquardt algorithm for small residual problems where, in addition, the Jacobian is rank-deficient. By "small residual" we mean that there exists a point p^*, which is not necessarily unique, such that $f(p^*)$ is small enough for the Gauss–Newton iteration to converge rapidly. By "local convergence" we mean that the initial iterate is near p^*.

There are classical results on convergence of Gauss–Newton methods and related iterations for nonlinear problems with rank-deficient Jacobians [1, 2, 6, 13, 19]. For instance, if J is Lipschitz continuous and has constant rank, and if the initial iterate is near the manifold of solutions, then the iteration

$$p_+ = p_c - J(p_c)^\dagger R(p_c) \tag{1.3}$$

converges, in exact arithmetic, to a solution of the nonlinear least squares problem. Here $J(p_c)^\dagger$ is the Moore–Penrose inverse [3] of $J(p_c)$. One can globalize the iteration with, for example, continuation methods [2, 23, 24] or a line search [19].

1.2. Overview.

In section 2 we present our assumptions and derive convergence results in exact arithmetic. We extend the results of [1, 2] to show how the standard [5, 13] approach to managing the Levenberg parameter is still effective in the rank-deficient, small-residual case in exact arithmetic. In section 3 we the discuss the finite precision computation of the Levenberg–Marquardt trial step, by means of truncated SVD and subset selection. A numerical example in section 4 concludes the paper.

2. Convergence of the Levenberg–Marquardt algorithm in exact arithmetic.

We start with a brief description of the Levenberg–Marquardt iteration in section 2.1, and in section 2.2 we present the assumptions for our analysis. Then we show in section 2.3 that if the initial iterate is in a certain set containing the solution manifold, then the Levenberg–Marquardt parameter is bounded and the iterates converge to the solution manifold. In section 2.4 we show that if the residual is sufficiently small, then the iterates converge to a point in the solution manifold.

2.1. A typical Levenberg–Marquardt iteration.

The Levenberg–Marquardt algorithm [14, 15] method updates a current approximation p_c by computing a trial step

$$s_t = -\left(\nu I + J(p_c)^T J(p_c)\right)^{-1} g(p_c)$$

and testing a trial solution

$$p_t = p_c + s_t$$

to determine how well the local quadratic model

$$m_c(p) = f(p_c) + g(p_c)^T (p - p_c) + \frac{1}{2}(p - p_c)^T \left(\nu I + J(p)^T J(p)\right)(p - p_c)$$

approximates f. One does this by comparing the actual reduction

$$ared = f(p_c) - f(p_t)$$

with the predicted reduction, i.e., the reduction in the quadratic model

$$pred = m_c(p_c) - m_c(p_t) = f(p_c) - m_c(p_t) = -\nabla f(p_c)^T s_t/2.$$

If the approximation is good, then p_t is accepted as the new point p_+, otherwise the Levenberg parameter ν is increased and p_t is recomputed. In the rank-deficient case, s_t is the minimum norm solution of the linearized problem if $\nu = 0$.

Algorithm `levmar_step` below represents a single iteration in a typical trust-region based Levenberg–Marquardt implementation; see [4, 5, 13]. The algorithm manages the transition from a current point p_c to the new point p_+ and controls the increase or decrease of the Levenberg parameter ν. The following parameters control the iteration:

$$0 < \omega_{down} < 1 < \omega_{up}, \quad \nu_0 \geq 0, \qquad \text{and} \qquad 0 \leq \mu_0 < \mu_{low} \leq \mu_{high} < 1.$$

Typical values, which we use in the numerical results in section 4, are

$$\mu_0 = 10^{-4}, \quad \mu_{low} = 1/4, \quad \mu_{high} = 3/4, \quad \omega_{down} = 1/2, \quad \text{and} \quad \omega_{up} = 2. \tag{2.1}$$

```
ALGORITHM. levmar_step($p_c, p_t, p_+, f, \nu$)
    1. $z = p_c$
    2. Do while $z = p_c$
        (a) $ared = f(p_c) - f(p_t)$, $s_t = p_t - p_c$, $pred = -\nabla f(p_c)^T s_t/2$
        (b) If $ared/pred < \mu_0$, then set $z = p_c$, $\nu = \max(\omega_{up}\nu, \nu_0)$, and recompute
            the trial point with the new value of $\nu$.
        (c) If $\mu_0 \leq ared/pred < \mu_{low}$, then set $z = p_t$ and $\nu = \max(\omega_{up}\nu, \nu_0)$.
        (d) If $\mu_{low} \leq ared/pred$, then set $z = p_t$.
            If $\mu_{high} < ared/pred$, then set $\nu = \omega_{down}\nu$.
            If $\nu < \nu_0$, then set $\nu = 0$.
    3. $p_+ = z$.
```

The ratio $\rho = ared/pred$ controls the Levenberg parameter ν. As long as either $\nu > 0$, or else $\nu = 0$ and $J(p_c)$ has full column rank, we have

$$pred = \frac{1}{2}(J(p_c)^T R(p_c))^T (\nu I + J(p_c)^T J(p_c))^{-1} J(p_c)^T R(p_c) > 0.$$

Hence $\rho < 0$ means that the step produced a decrease in f, which is a clear indicator that the model is poor. Even if $\rho > 0$, a small value implies that the model cannot be trusted to give good results, so the standard response is to reject the new point and increase the Levenberg parameter. Increasing the Levenberg parameter ν increases the singular values of the model Hessian and reduces the size of s_t, both of which will improve the quality of the model. Decreasing the Levenberg parameter ν if ρ is close to 1 is important for fast convergence. In the small-residual case of interest in this paper, one would like a method that approaches the fast convergence of the Gauss–Newton iteration for the full-rank case. In the Algorithm `levmar_step`, for example, ν is set to zero when it is sufficiently small.

2.2. Assumptions. The rank-deficient problems in which we are interested have a special structure: We seek to fit N model parameters, but the effects of these parameters are not independent. By this we mean that there is a map $B : \mathbb{R}^M \to \mathbb{R}^k$, with $k \leq N$ so that the model depends on the output of B. In [7, 17, 18] we treated the models as if they had this property and the numerical results were consistent with our examples in this paper.

To formalize this we assume that R can be decomposed as

$$R(p) = \tilde{R}(B(p)), \tag{2.2}$$

where the Jacobian $\tilde{R}' : \mathbb{R}^k \to \mathbb{R}^M$ has full column rank k and B' has full row rank k. The decomposition will be needed only in one part of the analysis and is never used in computation. We also assume, for simplicity, that

$$\tilde{f} = \frac{1}{2}\tilde{R}^T\tilde{R}$$

has a unique minimizer $b^* \in \mathbb{R}^k$. Hence the minimum value f^* of f is the same as $\tilde{f}(b^*)$. We will analyze the convergence of the Levenberg–Marquardt iteration near the solution manifold

$$\mathcal{Z} = \{p \,|\, f(p) = f^*\} = \{p \,|\, B(p) = b^*\}.$$

If the residual is zero and J has full rank $k = N$, then $\mathcal{Z}$ consists of isolated points. We let

$$R^* = \tilde{R}(b^*)$$

so that $f^* = \frac{1}{2}(R^*)^T R^*$.

The trial steps in the Levenberg–Marquardt algorithm now have the more general form

$$s_t = -\left(\nu I + J(p_c)^T J(p_c)\right)^\dagger g(p_c). \tag{2.3}$$

The following assumptions were motivated by the applications in [7, 18]. These are somewhat stronger than the standard assumptions for convergence of the Levenberg–Marquardt algorithm for full-rank problems and also stronger than those for the classical results. We use them in many ways in our convergence proof for the rank-deficient case of interest in this paper.

Assumption 2.1. There is $\delta > 0$ such that in the set

$$\mathcal{Z}_\delta = \{p \,|\, \|p - p^*\| \leq \delta \text{ for some } p^* \in \mathcal{Z} \ \}$$

the following conditions are satisfied:

1. B and $\tilde{R}$ are uniformly Lipschitz continuously differentiable,
2. $\min_{p \notin \mathcal{Z}_\delta} \|g(p)\| > 0$ for all $\delta > 0$,
3. the k singular values of B' are uniformly bounded away from zero,
4. the k singular values of the Jacobian of $\tilde{R}$ are uniformly bounded away from zero, and
5. the optimal value of f is small, but not necessarily zero.

Because $k \leq N \leq M$, Assumption 2.1 implies that $J(p)$ has rank $k \leq N$ for all p. Assumption 2.1 also implies that J is Lipschitz continuous. Let γ denote the Lipschitz constant of J, so that

$$\|J(p_1) - J(p_2)\| \leq \gamma \|p_1 - p_2\|. \tag{2.4}$$

At last, Assumption 2.1 implies that the k nonzero singular values of J are bounded from above and below in $\mathcal{Z}_\delta$. Let $\bar{\sigma}_1$ and $\bar{\sigma}_k$ denote upper and lower bounds for these singular values, and define for $\nu \geq 0$,

$$\eta(\nu) = \max_{\bar{\sigma}_k \leq \sigma \leq \bar{\sigma}_1} \frac{\sigma}{\nu + \sigma^2}. \tag{2.5}$$

In section 2.3 we show that the Levenberg parameter remains bounded, even in the rank-deficient case. Then we show that if the norm of the residual is sufficiently small, the Levenberg–Marquardt iteration converges locally q-linearly to a point on the set $\mathcal{Z}$.

As in [13] we define

$$d(p) = \min_{z \in \mathcal{Z}} \|z - p\|.$$

The following lemma below expresses the trial step (2.3) in terms of the error at the current point.

LEMMA 2.1. *Let δ be small enough so that Assumption* 2.1 *holds in $\mathcal{Z}_\delta$. Let $p_c \in \mathcal{Z}_\delta$, and let p^* be a point in $\mathcal{Z}$ closest to p_c so that*

$$d(p_c) = \|e_c\|, \qquad \textit{where} \quad e_c = p_c - p^*.$$

Then for all $\nu \geq 0$,

$$s_t = -(\nu I + J(p_c)^T J(p_c))^\dagger J(p_c)^T J(p_c) e_c + \Delta_S, \tag{2.6}$$

where

$$\|\Delta_S\| \leq \frac{\gamma \eta(\nu)}{2} \|e_c\|^2 + \frac{\gamma \|R^*\|}{\nu + \bar{\sigma}_k^2} \|e_c\|. \tag{2.7}$$

Proof. With (1.2) we can express (2.3) as

$$s_t = -\left(\nu I + J(p_c)^T J(p_c)\right)^\dagger J(p_c)^T R(p_c). \tag{2.8}$$

First consider $R(p_c)$. As is standard, we combine Taylor's theorem, the fundamental theorem of calculus, and (2.4) to obtain

$$R(p_c) = R(p^*) + \int_0^1 J(p^* + t e_c) e_c \, dt = R^* + J(p_c) e_c + \Delta_R, \tag{2.9}$$

where

$$\Delta_R = \int_0^1 (J(p^* + te_c) - J(p_c))e_c \, dt$$

and

$$\|\Delta_R\| \le \gamma \|e_c\|^2 \int_0^1 (1-t)\, dt = \frac{\gamma}{2}\|e_c\|^2.$$

Now consider the Moore–Penrose inverse in (2.8). With $\mathcal{P}$ being the orthogonal projector onto the range of $J(p_c)^T$, we can write

$$-\left(\nu I + J(p_c)^T J(p_c)\right)^\dagger J(p_c)^T = -\left(\nu I + J(p_c)^T J(p_c)\right)^\dagger \mathcal{P} J(p_c)^T,$$

and abbreviate

$$D(\nu) = \left(\nu I + J(p_c)^T J(p_c)\right)^\dagger \mathcal{P}.$$

Putting this abbreviation and the Taylor expansion (2.9) into (2.8) gives

$$\begin{aligned} s_t &= -D(\nu)J(p_c)^T R(p_c) = -D(\nu)J(p_c)^T \left[R^* + J(p_c)e_c + \Delta_R\right] \\ &= -D(\nu)J(p_c)^T J(p_c)e_c + \Delta_S, \end{aligned}$$

where

$$\Delta_S = -D(\nu)J(p_c)^T \left[R^* + \Delta_R\right].$$

This proves the expansion for s_t in (2.6).

We still need to show the bound for Δ_S in (2.7). Inserting $J(p^*)^T R^* = J(p^*)^T R(p^*) = g(p^*) = 0$ into the above expression for Δ_S yields

$$\Delta_S = -D(\nu)\left(J(p_c) - J(p^*)\right)^T R^* - D(\nu)J(p_c)^T \Delta_R.$$

Due to the presence of $\mathcal{P}$ in the expression for $D(\nu)$ we have $\|D(\nu)\| \le \frac{1}{\nu + \bar{\sigma}_k^2}$. Hence

$$\|D(\nu)(J(p_c) - J(p^*))^T\| \le \frac{\gamma\|e_c\|}{\nu + \bar{\sigma}_k^2} \qquad \text{and} \qquad \|D(\nu)J(p_c)^T\| \le \eta(\nu). \qquad \square$$

2.3. The iterates approach the solution manifold. We show that for sufficiently small residuals and Levenberg–Marquardt iterates sufficiently close to the solution manifold, the Levenberg–Marquardt parameter ν remains bounded. The boundedness of ν, which we prove in Theorem 2.2, eliminates a failure mode from the iteration and implies that we will always find an acceptable step after at most finitely many changes to ν. This implies that the search for an acceptable step $s_n = p_{n+1} - p_n$ will succeed and hence the sequence $\{p_n\}$ is infinite. We will use this to show that $d(p_n) \to 0$, which implies that the iterates $\{p_n\}$ approach the solution manifold $\mathcal{Z}$.

THEOREM 2.2. *Let Assumption* 2.1 *hold. There are* $\delta_0 > 0$, $r^* > 0$, *and* $\nu_{max} > 0$ *such that if* $\delta \le \delta_0$, $p_c \in \mathcal{Z}_\delta$, $\|R^*\| \le r^*$, *and* $\nu \ge \nu_{max}$, *then*

$$ared > \mu_{high} pred.$$

Hence if the Levenberg–Marquardt iterates satisfy $\{p_n\} \subset \mathcal{Z}_\delta$*, then the Levenberg–Marquardt parameters satisfy*

$$\limsup \nu_n \leq \omega_{up}\nu_{max}.$$

Proof. Let δ be small enough so that Assumption 2.1 holds in $\mathcal{Z}_\delta$. Taylor's theorem implies

$$R(p_t) = R(p_c) + J(p_c)s_t + \Delta_R^c, \qquad \text{where} \qquad \|\Delta_R^c\| \leq \frac{\gamma}{2}\|s_t\|^2. \tag{2.10}$$

Taking norms and squaring gives

$$\|R(p_t)\|^2 = \|R(p_c)\|^2 + s_t^T J(p_c)^T J(p_c)s_t + 2s_t^T g(p_c) + \Delta_{pred}, \tag{2.11}$$

where

$$\Delta_{pred} = 2\left(R(p_c) + J(p_c)^T s_t\right)^T \Delta_R^c + \|\Delta_R^c\|^2. \tag{2.12}$$

With (1.1) we can express (2.11) as

$$ared = -\frac{1}{2}s_t^T J(p_c)^T J(p_c)s_t - s_t^T g(p_c) - \frac{\Delta_{pred}}{2}. \tag{2.13}$$

We simplify the first two summands by noting that

$$\begin{aligned} pred &= -\frac{1}{2}s_t^T g(p_c) \\ &= \frac{1}{2}s_t^T\left(\nu I + J(p_c)^T J(p_c)\right)s_t = \frac{\nu}{2}\|s_t\|^2 + \frac{1}{2}s_t^T J(p_c)^T J(p_c)s_t \end{aligned}$$

so that

$$ared = pred + \frac{\nu}{2}\|s_t\|^2 - \frac{\Delta_{pred}}{2}.$$

For the expression Δ_{pred} in (2.12) we seek an estimate of the form

$$|\Delta_{pred}| \leq \gamma\left[C_{pred}\delta + \|R^*\|\right]\|s_t\|^2. \tag{2.14}$$

To this end, we combine

$$\|R(p_c)\| = \left\|R^* + \int_0^1 J(p_c + te_c)e_c\,dt\right\| \leq \|R^*\| + \bar{\sigma}_1\|e_c\|,$$

with $\|e_c\| \leq \delta$, $\|J(p_c)^T s_t\| \leq \bar{\sigma}_1\|s_t\|$, and (2.10) to conclude

$$2\left(R(p_c) + J(p_c)^T s_t\right)^T \Delta_R^c \leq \gamma\left[(\bar{\sigma}_1\delta + \|R^*\|) + \bar{\sigma}_1\|s_t\|\right]\|s_t\|^2.$$

We substitute this into (2.12) and obtain

$$|\Delta_{pred}| \leq \gamma\left[\bar{\sigma}_1(\|s_t\| + \delta) + \|R^*\| + \frac{\gamma}{4}\|s_t\|^2\right]\|s_t\|^2. \tag{2.15}$$

Next we bound $\|s_t\|$. From Lemma 2.1 and

$$\left\|\left(\nu I + J(p_c)^T J(p_c)\right)^\dagger J(p_c)^T J(p_c)e_c\right\| \leq \|e_c\|$$

follows

$$\|s_t\| \leq \|e_c\| \left(1 + \frac{\gamma}{2}\eta(\nu)\delta + \frac{\gamma}{\nu + \bar{\sigma}_k^2}\|R^*\|\right). \tag{2.16}$$

Let $\beta > 0$ and assume, by reducing δ if necessary, that the first interesting term in (2.16) is bounded by

$$\frac{\gamma}{2}\eta(\nu)\delta \leq \beta.$$

This will certainly be the case if $\delta \leq \delta_0 \equiv 2\beta\bar{\sigma}_k/\gamma$, as then

$$\frac{\gamma}{2}\eta(\nu)\delta \leq \beta\frac{\bar{\sigma}_k\eta(\nu)}{2} \leq \beta.$$

For the second interesting term in (2.16) we assume $\|R^*\| \leq r^* \equiv \beta\bar{\sigma}_k^2/\gamma$, so that

$$\frac{\gamma}{\nu + \bar{\sigma}_k^2}\|R^*\| \leq \beta.$$

Since both interesting terms in (2.16) are bounded above by β, and $\|e_c\| \leq \delta$, we get

$$\|s_t\| \leq (1 + 2\beta)\|e_c\| \leq (1 + 2\beta)\delta.$$

Using the above bound for the terms in (2.15) yields

$$\bar{\sigma}_1\|s_t\| \leq \bar{\sigma}_1(1 + 2\beta)\delta \qquad \text{and} \qquad \frac{\gamma}{4}\|s_t\|^2 \leq \gamma(1 + 2\beta)^2\delta^2. \tag{2.17}$$

Reducing δ again if needed so that

$$\delta < \frac{1}{\gamma(1 + 2\beta)}$$

and adding the terms in (2.17) gives

$$\bar{\sigma}_1\|s_t\| + \frac{\gamma}{4}\|s_t\|^2 \leq (1 + \bar{\sigma}_1)(1 + 2\beta)\delta.$$

Putting this into (2.15) completes the derivation of (2.14) with

$$C_{pred} = \bar{\sigma}_1 + (1 + 2\beta)(1 + \bar{\sigma}_1).$$

Now that we have established the validity of the bound (2.14), we can substitute it into (2.13),

$$ared \geq pred + \frac{\|s_t\|^2}{2}\left(\nu - \gamma\left[C_{pred}\delta + \|R^*\|\right]\right).$$

If we assume that $\nu \geq \nu_{max} \equiv \gamma(C_{pred}\delta + \|R^*\|)$, as well as $\mu_{high} = 3/4$ from (2.1), we can finally conclude that $ared \geq pred > \mu_{high}pred$. This proves the first assertion.

As for the second assertion, $ared > \mu_{high}pred$ implies that condition 2(d) in Algorithm `levmar_step` is fulfilled and ν is reduced to $\omega_{down}\nu$. From this and steps 2(b) and 2(c) we conclude that if ν is ever increased beyond ν_{max}, it will be reduced. Hence $\nu \leq \omega_{up}\nu_{max}$. □

The boundedness of the Levenberg–Marquardt parameters ν allows us to show below that the Levenberg–Marquardt iterates approach the solution manifold, provided they are sufficiently close to start with.

COROLLARY 2.3. *Let $\mu_0 > 0$ and let the assumptions of Theorem 2.2 hold. Then there exists $\delta < \delta_0$ so that if $p_0 \in \mathcal{Z}_\delta$, then all subsequent Levenberg–Marquardt iterates $\{p_n\}$ remain in $\mathcal{Z}_{\delta_0}$ and $d(p_n) \to 0$.*

Proof. Assume that the iterates are close enough to a minimum. That is, let $\epsilon > 0$ be small enough so that

$$\mathcal{E} = \{p \,|\, \|R(p)\| \le \|R^*\| + \epsilon\} \subset \mathcal{Z}_{\delta_0},$$

and let δ be small enough so that $\mathcal{Z}_\delta \subset \mathcal{E}$. Since $ared > 0$ for all accepted trial steps, the function values $f(p_n) = \|R(p_n)\|^2/2$ are nonincreasing. Hence the iterates $\{p_n\}$ remain in $\mathcal{E} \subset \mathcal{Z}_{\delta_0}$.

Moreover, since $\nu_n \le \nu_{max}$ by Theorem 2.2, each acceptable step s_n must satisfy

$$\begin{aligned} f(p_n) - f(p_{n+1}) &= ared > \mu_0 pred \\ &= -\mu_0 g(p_n)^T s_n = \mu_0 g(p_n) \left(\nu_n I + J(p_n)^T J(p_n)\right)^\dagger g(p_n) \\ &\ge \mu_0 \frac{\|g(p_k)\|^2}{\nu_{max} + \bar{\sigma}_k^2}. \end{aligned}$$

Therefore $\|g(p_n)\| \to 0$ and part 2 of Assumption 2.1 implies $d(p_n) \to 0$. □

2.4. The iterates converge. Corollary 2.3 gives conditions under which $d(p_n) \to 0$ so that the Levenberg–Marquardt iterates $\{p_n\}$ approach the solution manifold $\mathcal{Z}$. However, this does not imply that the iterates $\{p_n\}$ actually converge to a point in $\mathcal{Z}$, nor does $p_c \in \mathcal{Z}_\delta$ imply that $p_+ \in \mathcal{Z}_\delta$.

To prove this we will show that for all p_c with $d(p_c)$ sufficiently small there is $\alpha < 1$ such that if $p^* \in \mathcal{Z}$ and $\|p_c - p^*\| = d(p_c)$, then

$$d(p_t) \le \|p_t - p^*\| \le \alpha \|p_c - p^*\| \equiv \alpha d(p_c). \tag{2.18}$$

If indeed (2.18) holds, then

$$\|p_t - p_c\| \le \|p_c - p^*\| + \|p_t - p^*\| \le (1+\alpha)\|p_c - p^*\| \equiv (1+\alpha) d(p_c). \tag{2.19}$$

Combining (2.18) and (2.19) yields

$$\|p_n - p_{n-1}\| \le (1+\alpha) d(p_{n-1}) \le 2\alpha^{n-1} d(p_0). \tag{2.20}$$

Therefore $\{p_n\}$ is a Cauchy sequence and must converge to a point $p_\infty \in \mathcal{Z}$.

We need a technical lemma to prove convergence. The lemma shows that if $\|p_c - p^*\| = d(p_c)$, then $p_c - p^*$ has only a small contribution in the null space of $J(p_c)$.

LEMMA 2.4. *Let Assumption 2.1 hold, let $\mathcal{P}$ be the orthogonal projection onto the range of $J(p_c)^T$, and let $\|p_c - p^*\| = d(p_c)$, then*

$$(I - \mathcal{P})(p_c - p^*) = \mathcal{O}(d(p_c)^{3/2}).$$

Proof. It suffices to show that there is $q^* \in \mathcal{Z}$ such that

$$q^* = p^* + \mathcal{O}(d(p_c)^2) \quad \text{and} \quad (I - \mathcal{P})(p_c - q^*) = \mathcal{O}(d(p_c)^{3/2}).$$

If there is indeed such a q^*, then with $e_c = p_c - p^*$ and $d(p_c) = \|e_c\|$ we obtain

$$(I - \mathcal{P})e_c = (I - \mathcal{P})(p_c - q^*) + \mathcal{O}(d(p_c)^2) = \mathcal{O}(d(p_c)^{3/2}),$$

as asserted.

From $R(p_c) = \tilde{R}(B(p_c))$ follows

$$J(p_c) = \tilde{R}'(B(p_c))B'(p_c).$$

Hence the range of $J(p_c)^T$ is contained in the range of $B'(p_c)^T$. This, in turn, implies that the null space of $B'(p_c)$ is contained in the null space of $J(p_c)$ which is equal to the range of $I - \mathcal{P}$; hence

$$B'(p_c)(I - \mathcal{P}) = 0. \tag{2.21}$$

To see that q^* exists define a map $G : \mathbb{R}^k \to \mathbb{R}^k$ by

$$G(\sigma) = B(p_c - \mathcal{P}e_c + V_1\sigma) - b^*,$$

where V_1 is an orthonormal basis for the range of $J(p_c)^T$, so that $\mathcal{P} = V_1 V_1^T$.

Expanding the first term in the expression for $G(\sigma)$ gives

$$B(p_c - \mathcal{P}e_c) = B(p^* + (I - \mathcal{P})e_c) = B(p^*) + B'(p^*)(I - \mathcal{P})e_c + E_B^1, \tag{2.22}$$

where

$$\|E_B^1\| \le \frac{\gamma_B}{2}\|e_c\|^2 = \frac{\gamma_B}{2}d(p_c)^2$$

and γ_B is the Lipschitz constant of B'. Further expanding the second term on the right-hand side of (2.22) gives

$$B'(p^*)(I - \mathcal{P})e_c = B'(p_c)(I - \mathcal{P})e_c + E_B^2, \qquad \text{where} \qquad \|E_B^2\| \le \frac{\gamma_B}{2}d(p_c)^2.$$

Since $B'(p_c)(I - \mathcal{P})e_c = 0$ from (2.21), we obtain $B'(p^*)(I - \mathcal{P})e_c = E_B^2$. Substituting this into (2.22) gives

$$\begin{aligned} G(0) = B(p_c - \mathcal{P}e_c) - b^* &= B(p^* + (I - \mathcal{P})e_c) - B(p^*) \\ &= E_B^1 + E_B^2 = \mathcal{O}(d(p_c)^2). \end{aligned}$$

The Kantorovich theorem [12, 16] and Assumption 2.1 imply that there is a solution σ^* of $G(\sigma) = 0$ with $\|\sigma^*\| = \mathcal{O}(d(p_c)^2)$. Setting

$$q^* = p_c - \mathcal{P}(e_c) + V_1\sigma^* = p_c + \mathcal{P}(p^* - p_c) + V_1\sigma^*$$

implies that $B(q^*) = b^*$ and hence $q^* \in \mathcal{Z}$, as well as

$$(I - \mathcal{P})(p_c - q^*) = -V_1\sigma^*.$$

Since $\mathcal{P}V_1 = V_1$, the last equality implies that

$$(I - \mathcal{P})(p_c - q^*) = 0.$$

Hence we have shown that there exists a $q^* \in \mathcal{Z}$ so that $(I-\mathcal{P})(p_c - q^*) = 0$.

It remains to show that $q^* - p^* = \mathcal{O}(d(p_c)^2)$. From $q^* = p^* + (I-\mathcal{P})e_c + V_1\sigma^*$ and $\|\sigma^*\| = \mathcal{O}(d(p_c)^2)$ follows

$$\mathcal{P}(q^* - p^*) = \mathcal{P}V_1\sigma^* = \mathcal{O}(d(p_c)^2).$$

The theorem of Pythagoras then implies

$$\|q^* - p^*\|^2 = \|(I-\mathcal{P})(q^* - p^*)\|^2 + \|\mathcal{P}(q^* - p^*)\|^2 = \|(I-\mathcal{P})(q^* - p^*)\|^2 + \mathcal{O}(d(p_c)^4).$$

To bound $\|(I-\mathcal{P})(q^* - p^*)\|$ we observe that $(I-\mathcal{P})(q^* - p^*) = (I-\mathcal{P})e_c$. Using this in

$$\|\mathcal{P}e_c\|^2 + \|(I-\mathcal{P})e_c\|^2 = d(p_c)^2 \leq \|p_c - q^*\|^2$$

gives

$$\|(I-\mathcal{P})(q^* - p^*)\|^2 = \|(I-\mathcal{P})e_c\|^2 \leq \|p_c - q^*\|^2 - \|\mathcal{P}e_c\|^2.$$

Writing $p_c - q^* = \mathcal{P}e_c + V_1\sigma^*$ gives

$$\begin{aligned}\|(I-\mathcal{P})(q^* - p^*)\|^2 &\leq \|\mathcal{P}e_c + V_1\sigma^*\|^2 - \|\mathcal{P}e_c\|^2 \leq 2\|\mathcal{P}e_c\|\|V_1\sigma^*\| + \|V_1\sigma^*\|^2 \\ &= \mathcal{O}(d(p_c)^3) + \mathcal{O}(d(p_c)^4) = \mathcal{O}(d(p_c)^3),\end{aligned}$$

which completes the proof. □

The theorem below exploits the fact that the error has a small enough contribution in the null space of the Jacobian, so that the nonzero singular values of the Jacobian mitigate the potential ill-conditioning caused by a small Levenberg–Marquardt parameter ν.

THEOREM 2.5. *Let the assumptions of Theorem* 2.2 *hold. Let* $\|p_* - p_c\| = d(p_c)$, *then* (2.18) *holds, that is,*

$$d(p_t) \leq \alpha d(p_c) \qquad \textit{for some } \alpha \in (0,1). \tag{2.23}$$

Proof. Lemma 2.1 implies that

$$s_t = -D(\nu)J(p_c)^T J(p_c)e_c + \Delta_S,$$

where $D(\nu) = \left(\nu I + J(p_c)^T J(p_c)\right)^\dagger \mathcal{P}$, $e_c = p_c - p^*$, and

$$\|\Delta_S\| \leq \gamma\left(\frac{\|e_c\|}{2}\eta(\nu) + \frac{\|R^*\|}{\nu + \bar{\sigma}_k^2}\right)\|e_c\| \leq \gamma\left(\frac{\delta_0}{2}\eta(\nu) + \frac{r^*}{\nu + \bar{\sigma}_k^2}\right) d(p_c). \tag{2.24}$$

Since $\mathcal{P}$ is the orthogonal projector onto the range of $J(p_c)^T$,

$$s_t = -D(\nu)J(p_c)^T J(p_c)\mathcal{P}e_c + \Delta_S.$$

Let $p_t = p_c + s_t$ and $e_t = p_t - p^*$. Then the expression for s_t implies

$$e_t = e_c + s_t = e_c - D(\nu)J(p_c)^T J(p_c)\mathcal{P}e_c + \Delta_S.$$

Partitioning $e_c = (I-\mathcal{P})e_c + \mathcal{P}e_c$ gives

$$e_t = (I-\mathcal{P})e_c + \left[I - D(\nu)J(p_c)^T J(p_c)\right]\mathcal{P}e_c + \Delta_S.$$

From

$$\left[I - D(\nu)J(p_c)^T J(p_c)\right]\mathcal{P}e_c = \nu\left(\nu I + J(p_c)^T J(p_c)\right)^\dagger \mathcal{P}e_c$$

follows

$$e_t = (I-\mathcal{P})e_c + \nu\left(\nu I + J(p_c)^T J(p_c)\right)^\dagger \mathcal{P}e_c + \Delta_S.$$

Bounding this gives

$$d(p_t) \le \|e_t\| \le \|(I-\mathcal{P})e_c\| + \frac{\nu}{\nu+\bar{\sigma}_k^2}\|\mathcal{P}e_c\| + \|\Delta_S\|.$$

Lemma 2.4 implies $\|(I-\mathcal{P})e_c\| \le C_k\|e_c\|^{3/2}$ for some constant C_k, hence

$$\begin{aligned} d(p_t) &\le C_k\|e_c\|^{3/2} + \tfrac{\nu}{\nu+\bar{\sigma}_k^2}\|\mathcal{P}e_c\| + \|\Delta_S\| \\ &\le \left(C_k\delta_0^{1/2} + \tfrac{\nu_{max}}{\nu_{max}+\bar{\sigma}_k^2}\right)\|e_c\| + \|\Delta_S\|. \end{aligned} \tag{2.25}$$

Substituting $\|e_c\| = d(p_c)$ and the above bound for $\|\Delta_S\|$ gives

$$d(p_t) \le \left[C_k\delta_0^{1/2} + \frac{\nu_{max}}{\nu_{max}+\bar{\sigma}_k^2} + \gamma\left(\frac{\delta_0}{2}\eta(\nu) + \frac{r^*}{\nu+\bar{\sigma}_k^2}\right)\right] d(p_c).$$

Hence, $d(p_t) \le \alpha d(p_c)$ in (2.18) holds if

$$\alpha \equiv C_k\delta_0^{1/2} + \frac{\nu_{max}}{\nu_{max}+\bar{\sigma}_k^2} + \gamma\left(\frac{\delta_0}{2}\eta(\nu) + \frac{r^*}{\nu+\bar{\sigma}_k^2}\right) < 1. \qquad \square \tag{2.26}$$

One can extract convergence rate estimates from Theorem 2.5 and its proof. The results above imply that the trial step will be accepted for some $\nu \le \nu_{max}$ and then $p_+ = p_t$. Therefore $d(p_n)$ converges q-linearly to zero in the small residual case. The middle inequalities in (2.24) and (2.25) imply that if $r^* = 0$ and $\nu_n \to 0$, then

$$d(p_{n+1}) = o(d(p_n)), \tag{2.27}$$

and hence $d(p_n)$ will converge to zero q-superlinearly in that case. The estimates (2.23) and (2.27) imply r-linear or r-superlinear convergence of p_n to some $p_\infty \in \mathcal{Z}$. We will state this as a formal corollary.

Corollary 2.6. *Let the assumptions of Theorem* 2.5 *hold. Let $\alpha \in (0,1)$ be the constant from* (2.23)*, and let p_∞ be the limit of the Levenberg–Marquardt sequence $\{p_n\}$. Then*

$$\|p_n - p_\infty\| \le \frac{1+\alpha}{1-\alpha}d(p_n),$$

and hence $p_n \to p_\infty$ r-linearly. Moreover the convergence is r-superlinear if (2.27) *holds.*

Proof. We begin with (2.20), which together with (2.23) implies that

$$\begin{aligned} \|p_n - p^*\| &\le \textstyle\sum_{k=n}^\infty \|p_{k+1}-p_k\| \le (1+\alpha)\sum_{k=n}^\infty d(p_k) \\ &\le (1+\alpha)d(p_n)\textstyle\sum_{k=n}^\infty d(p_k)/d(p_n) \le \frac{1+\alpha}{1-\alpha}d(p_n), \end{aligned}$$

as asserted. $\square$

3. Computing the trial step in finite precision.

We analyze two methods for computing the trial step s_t (2.3) in finite precision: truncated SVD and subset selection. We show that computing s_t via a truncated SVD may not be accurate if the numerical rank of the Jacobian $J(p_c)$ is not well defined. Note that if the Jacobian has less than full column rank, then a perturbation can not only decrease the rank but also increase it.

Abbreviate

$$J \equiv J(p_c), \qquad R \equiv R(p_c), \qquad s \equiv s_t,$$

where J is $M \times N$ with $M \geq N$ and has rank k. As in section 2.2, we use the bounds

$$0 < \bar{\sigma}_k \leq \sigma_i \leq \bar{\sigma}_1$$

for the k nonzero singular values σ_i of J. The exact Levenberg–Marquardt trial step is

$$s = -\left(\nu I + J^T J\right)^{\dagger} J^T R.$$

To put things in perspective, we first analyze the ordinary Levenberg–Marquardt algorithm, where the matrix $\nu I + J^T J$ is nonsingular.

3.1. Full-rank case.

We assume that $\nu I + J^T J$ is nonsingular, and denote the perturbed Jacobian by $\tilde{J} = J + E$, where $\nu I + \tilde{J}^T \tilde{J}$ is also nonsingular. The corresponding computed trial step is

$$\tilde{s} = -\left(\nu I + \tilde{J}^T \tilde{J}\right)^{-1} \tilde{J}^T R.$$

We show below that the difference $\|\tilde{s} - s\|$ can be bounded by a multiple of $\|E\|$, and that the nonlinear residual can amplify the ill-conditioning of J. However, a large value of ν can dampen the influence of the nonlinear residual.

THEOREM 3.1. *If J and $\tilde{J} = J + E$ have full column rank $k = N$ and $\nu \geq 0$, or if $\nu > 0$, then*

$$\|\tilde{s} - s\| \leq \left[\eta(\nu)\|\tilde{s}\| + \frac{\|R\|}{\nu + \bar{\sigma}_k^2}\right] \|E\|.$$

Proof. Let us define $s = -s$ and $\tilde{s} = -\tilde{s}$. Then s solves the full-rank least squares problem

$$\min_x \|Ax - b\|, \qquad \text{where} \qquad A = \begin{pmatrix} J \\ \sqrt{\nu} I \end{pmatrix}, \quad b = \begin{pmatrix} R \\ 0 \end{pmatrix},$$

so that $s = A^{\dagger} b$. Similarly, $\tilde{s} = (A + F)^{\dagger} b$ is the solution of the full-rank problem

$$\min_x \|(A + F)x - b\|, \qquad \text{where} \qquad F = \begin{pmatrix} E \\ 0 \end{pmatrix},$$

with residual $r = (A + F)\tilde{s} - b$. We derive the perturbation bound as in [11, Fact 5.14] by observing that $(A + F)^T r = 0$ because the residual is orthogonal to the column space. Hence $A^T r = -F^T r$. Multiply by $(A^T A)^{-1}$ to obtain

$$-(A^T A)^{-1} F^T r = (A^T A)^{-1} A^T r = A^{\dagger} r = A^{\dagger}\left((A + F)\tilde{s} - b\right) = \tilde{s} - s + A^{\dagger} F \tilde{s}.$$

Hence

$$\|\tilde{s}-s\| \leq \|A^{\dagger}F\|\|\tilde{s}\| + \|(A^TA)^{-1}\|\|F^Tr\|.$$

Now write

$$A^{\dagger}F = \left(\nu I + J^TJ\right)^{-1} J^TE, \qquad (A^TA)^{-1} = \left(\nu I + J^TJ\right)^{-1},$$

and $F^Tr = F^T\left[I-(A+F)(A+F)^{\dagger}\right]b$. Since $I-(A+F)(A+F)^{\dagger}$ is an orthogonal projector, $\|r\| \leq \|b\|$. Hence

$$\|\tilde{s}-s\| \leq \left\{\|(\nu I + J^TJ)^{-1}J^T\|\|\tilde{s}\| + \|(\nu I + J^TJ)^{-1}\|\|R\|\right\} \|E\|.$$

The result now follows from $\|(\nu I + J^TJ)^{-1}J^T\| \leq \nu(\eta)$ and $\|(\nu I + J^TJ)^{-1}\| \leq 1/(\nu + \bar{\sigma}_k^2)$. □

3.2. Truncated SVD. We consider the case when $\nu I + J^TJ$ can be singular. Let the $M \times N$ Jacobian J with $M \geq N$ have rank k, and denote its thin SVD by

$$J = \begin{pmatrix} U_1 & U_{21} \end{pmatrix} \begin{pmatrix} \Sigma_1 & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} V_1 & V_2 \end{pmatrix}^T = U_1\Sigma_1V_1^T,$$

where the $M \times N$ matrix $(U_1 \quad U_{21})$ has orthonormal columns, the $N \times N$ matrix $(V_1 \quad V_2)$ is an orthogonal matrix, and the $k \times k$ diagonal matrix Σ_1 contains the nonzero singular values. In particular, $(\Sigma_1)_{ii} \geq \bar{\sigma}_k$.

Let the SVD of the perturbed matrix be

$$J + E = \begin{pmatrix} \tilde{U}_1 & \tilde{U}_{21} \end{pmatrix} \begin{pmatrix} \tilde{\Sigma}_1 & 0 \\ 0 & \tilde{\Sigma}_2 \end{pmatrix} \begin{pmatrix} \tilde{V}_1 & \tilde{V}_2 \end{pmatrix}^T,$$

where the $k \times k$ matrix $\tilde{\Sigma}_1$ contains the largest singular values, and the $(N-k)\times(N-k)$ matrix $\tilde{\Sigma}_2$ contains the smallest singular values. The $M \times N$ matrix $(\tilde{U}_1 \quad \tilde{U}_{21})$ has orthonormal columns, and the $N \times N$ matrix $(\tilde{V}_1 \quad \tilde{V}_2)$ is orthogonal.

We compute the SVD of the perturbed matrix $J+E$, then truncate it by setting the $N-k$ smallest singular values to zero. The resulting matrix is

$$\tilde{J} = \tilde{U}_1\tilde{\Sigma}_1\tilde{V}_1^T,$$

and from this we compute the trial step,

$$\tilde{s} = -(\nu I + \tilde{J}^T\tilde{J})^{\dagger}\tilde{J}^TR.$$

Below we bound the difference between computed and exact step in terms of $\|E\|$, $\bar{\sigma}_k^2 + \nu$, and $\eta(\nu)$ as defined in (2.5). The bound implies that the computed step is close to the exact step, if the computed singular vectors are accurate, if the Jacobian is far from being rank-deficient, and if the nonlinear residual is small.

THEOREM 3.2. *If the rank of J is $k < N$, if $\nu \geq 0$, and if*

$$\theta\|E\|_F < 1, \qquad \textit{where} \qquad \theta = \frac{2}{\bar{\sigma}_k - 2\|E\|} < 1,$$

then

$$\|s-\tilde{s}\| \leq \left[2\eta(\nu)\|\tilde{s}\| + \left(\frac{1}{\nu+\bar{\sigma}_k^2} + 2\eta(\nu)\theta\right)\|R\|\right]\|E\|_F + \omega\|E\|_F^2,$$

where

$$\omega = \frac{\|\tilde{s}\|}{\nu + \bar{\sigma}_k^2} + 2\eta(\nu)\theta^2 \|R\|.$$

Proof. Choose U_{22} and $\tilde{U}_{22}$ so that

$$U = \begin{pmatrix} U_1 & U_{21} & U_{22} \end{pmatrix} \qquad \text{and} \qquad \tilde{U} = \begin{pmatrix} \tilde{U}_1 & \tilde{U}_{21} & \tilde{U}_{22} \end{pmatrix}$$

are $M \times M$ orthogonal matrices, and set $\tilde{V} = (\tilde{V}_1 \ \ \tilde{V}_2)$. This gives the full SVDs

$$J = U \begin{pmatrix} \Sigma_1 & 0 \\ 0 & 0 \end{pmatrix} V^T, \qquad \tilde{J} = \tilde{U} \begin{pmatrix} \tilde{\Sigma}_1 & 0 \\ 0 & 0 \end{pmatrix} \tilde{V}^T.$$

According to [21, section 6] there exists a $(N-k) \times k$ matrix P that rotates V into $\tilde{V}$,

$$\tilde{V} = V \begin{pmatrix} I_k & -P^T \\ P & I_{N-k} \end{pmatrix} \begin{pmatrix} (I_k + P^T P)^{-1/2} & \\ & (I_{N-k} + PP^T)^{-1/2} \end{pmatrix}$$

and a $(M-k) \times k$ matrix Q that rotates U into $\tilde{U}$,

$$\tilde{U} = U \begin{pmatrix} I_k & Q^T \\ -Q & I_{M-k} \end{pmatrix} \begin{pmatrix} (I_k + Q^T Q)^{-1/2} & \\ & (I_{M-k} + QQ^T)^{-1/2} \end{pmatrix}.$$

By assumption $\theta\|E\|_F < 1$, so that P and Q satisfy $\|(P \ \ Q)\|_F < \theta$ [21, Theorem 6.4]. Hence $\|P\| < \theta\|E\|_F$ and $\|Q\| < \theta\|E\|_F$.

Since Σ_1 is symmetric positive definite, $\nu I + \Sigma_1^2$ is nonsingular for $\nu \geq 0$, so that we can write

$$s = -V \begin{pmatrix} C & 0 \\ 0 & 0 \end{pmatrix} U^T R, \qquad \text{with} \qquad C = (\nu I_k + \Sigma_1^2)^{-1}\Sigma_1.$$

Setting

$$\phi_P = \begin{pmatrix} (I_k + P^T P)^{-1/2} & \\ & (I_{N-k} + PP^T)^{-1/2} \end{pmatrix},$$
$$\phi_Q = \begin{pmatrix} (I_k + Q^T Q)^{-1/2} & \\ & (I_{M-k} + QQ^T)^{-1/2} \end{pmatrix}$$

allows us to write

$$\begin{aligned} s &= -\tilde{V}\,\tilde{V}^T V \begin{pmatrix} C & 0 \\ 0 & 0 \end{pmatrix} U^T \tilde{U}\,\tilde{U}^T R \\ &= -\tilde{V}\phi_P \begin{pmatrix} C & CQ^T \\ -PC & -PCQ^T \end{pmatrix} \phi_Q\,\tilde{U}^T R \\ &= -\tilde{V}\left[\phi_P \begin{pmatrix} C & 0 \\ 0 & 0 \end{pmatrix} \phi_Q + \phi_P \begin{pmatrix} 0 & CQ^T \\ -PC & -PCQ^T \end{pmatrix} \phi_Q\right] \tilde{U}^T R. \end{aligned}$$

Now consider the computed step $\tilde{s}$. From the assumption $\theta\|E\|_F < 1$ follows $\bar{\sigma}_k > 4\|E\|$. Hence we obtain for the kth singular value of $J + E$,

$$\tilde{\sigma}_k \geq \bar{\sigma}_k - \|E\| > 3\|E\|,$$

where the first inequality follows from the well-conditioning of singular values [8, Corollary 8.6.2]. Since $\tilde{\sigma}_k > 0$, the matrix $\tilde{\Sigma}_1$ must by positive definite, so that the computed step can be expressed as

$$\tilde{s} = -\tilde{V} \begin{pmatrix} \tilde{C} & 0 \\ 0 & 0 \end{pmatrix} \tilde{U}^T R, \qquad \text{where} \qquad \tilde{C} = (\nu I_k + \tilde{\Sigma}_1^2)^{-1} \tilde{\Sigma}_1.$$

The expressions for s and $\tilde{s}$ imply

$$\begin{aligned} \tilde{s} - s &= \tilde{V} \left[\phi_P \begin{pmatrix} C & 0 \\ 0 & 0 \end{pmatrix} \phi_Q + \phi_P \begin{pmatrix} 0 & CQ^T \\ -PC & -PCQ^T \end{pmatrix} \phi_Q - \begin{pmatrix} \tilde{C} & 0 \\ 0 & 0 \end{pmatrix} \right] \tilde{U}^T R \\ &= \tilde{V} \left[\begin{pmatrix} Z_p C Z_Q - \tilde{C} & 0 \\ 0 & 0 \end{pmatrix} + \phi_P \begin{pmatrix} 0 & CQ^T \\ -PC & -PCQ^T \end{pmatrix} \phi_Q \right] \tilde{U}^T R, \end{aligned}$$

where

$$Z_P = (I + P^T P)^{-1/2}, \qquad Z_Q = (I + Q^T Q)^{-1/2}.$$

We partition the difference $\tilde{s} - s$ into two terms, where

$$\begin{aligned} B_1 &= \tilde{V} \left[\begin{pmatrix} Z_p C Z_Q - \tilde{C} & 0 \\ 0 & 0 \end{pmatrix} \right] \tilde{U}^T R, \\ B_2 &= \tilde{V} \left[\phi_P \begin{pmatrix} 0 & CQ^T \\ -PC & -PCQ^T \end{pmatrix} \phi_Q \right] \tilde{U}^T R. \end{aligned} \tag{3.1}$$

This leaves us to bound $\|\tilde{s} - s\| \le \|B_1\| + \|B_2\|$.

We bound $\|B_2\|$ first, because it is easier. Using $\|\phi_P\| \le 1$, $\|\phi_Q\| \le 1$, and $\|C\| = \eta(\nu)$ in (3.1) yields

$$\|B_2\| \le \eta(\nu) \ (\|P\| + \|Q\| + \|P\|\|Q\|).$$

Since $\|P\| \le \theta \|E\|_F$ and $\|Q\| \le \theta \|E\|_F$ we obtain

$$\|B_2\| \le \eta(\nu) \ \left(2\theta + \theta^2 \|E\|_F\right) \ \|R\| \|E_F\|. \tag{3.2}$$

Now we bound $\|B_1\|$. We write the (1,1) element of the middle matrix in B_1 in (3.1) as

$$Z_P C Z_Q - \tilde{C} = C - \tilde{C} + Z_P C (Z_Q - I) + (Z_P - I) C. \tag{3.3}$$

We start with the term $C - \tilde{C}$. Singular value perturbation bounds [22, Theorem 4.11 in section IV] imply that $\tilde{\Sigma}_1 = \Sigma_1 + D$ holds for some diagonal matrix D with $\|D\| \le \|E\| \le \|E\|_F$. This, together with the abbreviations $A = \nu I + \Sigma_1^2$ and $B = \nu I + \tilde{\Sigma}_1^2$, allows us to write

$$\tilde{C} = A^{-1} \tilde{\Sigma}_1 \qquad \text{and} \qquad C = B^{-1} \Sigma_1 = B^{-1} \tilde{\Sigma}_1 - B^{-1} D.$$

From $B^{-1} = A^{-1} - B^{-1}(B - A) A^{-1}$ [8, eq. (2.1.3)] follows for the first summand in C

$$\begin{aligned} B^{-1} \tilde{\Sigma}_1 &= \tilde{C} - B^{-1}(B - A) \tilde{C} = \tilde{C} - \left(\nu I + \Sigma_1^2\right)^{-1} \left(\Sigma_1^2 - \tilde{\Sigma}_1^2\right) \tilde{C} \\ &= \tilde{C} + \left(\nu I + \Sigma_1^2\right)^{-1} (2\Sigma_1 + D) D \tilde{C} = \tilde{C} + \left(2C + \left(\nu I + \Sigma_1^2\right)^{-1} D\right) D \tilde{C}. \end{aligned}$$

Putting this into the above expression for C gives

$$\begin{aligned} C &= \tilde{C} + \left(\nu I + \Sigma_1^2\right)^{-1} (2\Sigma_1 + D)\, D\tilde{C} - \left(\nu I + \Sigma_1^2\right)^{-1} D \\ &= \tilde{C} + \left(2C + \left(\nu I + \Sigma_1^2\right)^{-1} D\right) D\tilde{C} - \left(\nu I + \Sigma_1^2\right)^{-1} D. \end{aligned}$$

In turn substituting this into (3.3) gives

$$Z_P C Z_Q - \tilde{C} = F_1 \tilde{C} + F_2,$$

where

$$\begin{aligned} F_1 &= \left(2C + \left(\nu I + \Sigma_1^2\right)^{-1} D\right) D, \\ F_2 &= -\left(\nu I + \Sigma_1^2\right)^{-1} D + Z_p C(Z_Q - I) + (Z_P - I)C. \end{aligned}$$

At last we substitute this into (3.1) to obtain

$$B_1 = \tilde{V} \begin{pmatrix} F_1 & 0 \\ 0 & 0 \end{pmatrix} \tilde{V}^T \underbrace{\tilde{V} \begin{pmatrix} \tilde{C} & 0 \\ 0 & 0 \end{pmatrix} \tilde{U}^T R}_{\tilde{s}} + \tilde{V} \begin{pmatrix} F_2 & 0 \\ 0 & 0 \end{pmatrix} \tilde{U}^T R.$$

Hence

$$\|B_1\| \le \|F_1\| \|\tilde{s}\| + \|F_2\| \|R\|. \tag{3.4}$$

From $\|C\| = \eta(\nu)$ and $\|D\| \le \|E\|_F$ follows

$$\|F_1\| \le \left(2\nu(\eta) + \frac{\|E\|_F}{\nu + \bar{\sigma}_k^2}\right) \|E\|_F.$$

In order to bound $\|F_2\|$ we use the fact that $(1 + x^2)^{-1/2} - 1 \le x^2/2$ for $|x| < 1$. Hence

$$\|Z_Q - I\| \le \frac{\|Q\|^2}{2} \le \frac{\theta^2}{2} \|E\|_F^2, \qquad \|Z_P - I\| \le \frac{\|P\|^2}{2} \le \frac{\theta^2}{2} \|E\|_F^2.$$

Putting the above inequalities into the formula for F_2 along with $\|Z_P\| \le 1$ gives

$$\|F_2\| \le \frac{\|E\|_F}{\nu + \bar{\sigma}_k^2} + \eta(\nu)\theta^2 \|E\|_F^2.$$

Putting the bounds for $\|F_1\|$ and $\|F_2\|$ into (3.4) gives

$$\|B_1\| \le \left[2\eta(\nu)\|\tilde{s}\| + \frac{\|R\|}{\nu + \bar{\sigma}_k^2}\right] \|E\|_F + \left[\frac{\|\tilde{s}\|}{\nu + \bar{\sigma}_k^2} + \eta(\nu)\theta^2 \|R\|\right] \|E\|_F^2.$$

Putting this bound and (3.2) into $\|\tilde{s} - s\| \le \|B_1\| + \|B_2\|$ gives the desired result. □

The first assumption in Theorem 3.2 requires that the kth singular value of J is sufficiently large compared to the perturbation E. If that is indeed the case, then θ indicates the accuracy of the computed singular vectors. In particular, θ is a bound on the tangent between the column spaces of V_1 and $\tilde{V}_1$, and also a bound on the tangent between the column spaces of U_1 and $\tilde{U}_1$. The quantity θ is small if $\bar{\sigma}_k \gg 0$, that is, the numerical rank of J is well defined. Furthermore, the denominator of θ can also be interpreted as the gap between the kth and the $(k+1)$st singular values of $J+E$. If this gap is large, then the computed singular vectors are close to the exact singular vectors, and θ is small. If the gap is small, then the singular vectors are sensitive and it is difficult to compute them accurately.

We can compare the bound for $\|\tilde{s}-s\|$ in Theorem 3.2, when J is rank-deficient, to the bound in Theorem 3.1, where J can have full-rank. We see that the first order term in Theorem 3.2 contains an additional factor $2\eta(\nu)\theta$ that amplifies the nonlinear residual. This factor reflects the conditioning of J and how well defined the rank of J is. The comparison of the two bounds shows that when J is rank-deficient the nonlinear residual affects the accuracy of the step $\tilde{s}$ even more than in the full-rank case. The influence of the residual is even more amplified by the size of the kth singular value.

In summary, the step computed by the truncated SVD is affected by the accuracy of the computed singular vectors, and this, in turn, depends on whether the numerical rank of J is well defined. If J is close to a matrix of lower rank, then the step computed by the truncated SVD may have no accuracy.

3.3. Subset selection. Subset selection chooses k linearly independent columns J_1 from the Jacobian J, and approximates the trial step by

$$\hat{s} = (\nu I + J_1^T J_1)^{-1} J_1^T R.$$

This avoids the difficulties of the truncated SVD for several reasons: First, it replaces a potentially ill-posed problem with a rank-deficient matrix J by a problem with a full column rank matrix J_1. Second, it does not require the computation of potentially sensitive singular vectors.

Subset selection on the matrix J produces a permutation matrix Π, which brings the k columns J_1 to the front, i.e., $J\Pi = (J_1 \;\; J_2)$. We use the strong rank revealing QR (SRRQR) algorithm by Gu and Eisenstat [9, Algorithm 4] in the exact version (with parameter $f=1$). This SRRQR algorithm assures, among other things, for the singular values $\sigma_i(J_1)$ of J_1 that

$$\frac{\sigma_i}{\sqrt{1+k(N-k)}} \le \sigma_i(J_1) \le \sigma_i, \qquad 1 \le i \le k. \tag{3.5}$$

We assume that the perturbed matrix $J+E$ also has rank at least k, and that subset selection brings the same group of k columns to the front, i.e., $(J+E)\Pi = (\tilde{J}_1 \;\; \tilde{J}_2)$, where $\tilde{J}_1$ has k columns. Denote the computed step by

$$\tilde{s} = (\nu I + \tilde{J}_1^T \tilde{J}_1)^{-1} \tilde{J}_1^T R.$$

THEOREM 3.3. *If the rank of J is $k < N$, if $\nu \ge 0$, and if the k columns of J_1 are selected by the SRRQR algorithm so that* (3.5) *holds, then*

$$\|\tilde{s} - \hat{s}\| \le \left[\tilde{\eta}(\nu)\|\tilde{s}\| + \frac{\|R\|}{\nu + \bar{\sigma}_k^2/(1+k(N-k))} \right] \|E\|,$$

where

$$\tilde{\eta}(\nu) = \max_{\bar{\sigma}_k/\sqrt{1+k(N-k)} \le \sigma \le \bar{\sigma}_1} \frac{\sigma}{\nu + \sigma^2}.$$

Proof. This follows from applying Theorem 3.1 to the full-rank matrices J_1 and $J_1 + E_1$, where $\|E_1\| \leq \|E\|$, and using the bound (3.5). □

The bound in Theorem 3.3 is, up to factors of $\sqrt{1 + k(N - k)}$, identical to the one in Theorem 3.1. This means, by applying the Levenberg–Marquardt algorithm to k specially selected columns we solve a full-rank, well-posed least squares problem and avoid the computation of potentially sensitive singular vectors. As a result, the nonlinear residual has less damaging effect on the accuracy of the computed step than in the rank-deficient case in Theorem 3.2.

We apply subset selection to nonlinear problems by using the permutation matrix Π to prune some of the design parameters. After applying subset selection to the Jacobian at the initial iterate $R'(p_0)$, we set

$$\Pi p = \left[\begin{array}{c} p^1 \\ p^2 \end{array} \right],$$

where p^1 corresponds to the columns of J_1. We propose fixing the remaining variables p^2 to nominal values, as we did in [18], and solving a full-rank nonlinear least squares problem for $p^1 \in \mathbb{R}^k$. We only apply subset selection at the initial iterate, and at the end of the optimization verify the first k columns of $R'\Pi$ are a well-conditioned $M \times k$ matrix.

4. Numerical examples.

The purpose of our experiments is to reproduce, in the simplest possible framework, the essential conditions of the cardiovascular models [7, 10, 17, 18] that lead to the failure of the truncated SVD.

To this end, we consider a parameter-dependent system of ordinary differential equations

$$y' = F(y, p), \qquad y(0) = y_0. \tag{4.1}$$

The goal is to approximate the parameter vector $p \in \mathbb{R}^N$ with a least squares fit to data. The nonlinear least squares problem is to minimize

$$f(p) = \frac{1}{2} \sum_{i=1}^{M} (\tilde{y}(t_i, p) - d_i)^2, \tag{4.2}$$

where $\tilde{y}(t_i, p)$ are approximations to the solution of the initial value problem evaluated at M points in time and $\{d_i\}$ are the data at those time points. We computed $\tilde{y}$ with the MATLAB `ode15s` integrator [20]. We manufacture the data vector d by either solving the initial value problem to very high precision or using an analytic solution, and then, for some of our examples, applying a random perturbation.

The elements of the residual vector $R \in \mathbb{R}^M$ are $R_i(p) = \tilde{y}(t_i, p) - d_i$, and we compute the columns of the Jacobian by computing the sensitivities,

$$w_j = \partial y / \partial p_j, \qquad 1 \leq j \leq N.$$

Here w_j is the solution of the initial value problem

$$w_j' + F_y(y, k) w_j + \frac{\partial F}{\partial p_j}(y, p) = 0, \qquad w_p(0) = 0.$$

We can solve for the sensitivities at the same time as we compute y, and thereby recover the Jacobian with the same accuracy as that of the initial value problem.

Had we elected to approximate Jacobians with differences, we would have produced significantly less accurate Jacobians.

We use a Levenberg–Marquardt code [5, 13, 14, 15] to solve the nonlinear least squares problems. Our implementation enforces simple bound constraints. Our subset selection algorithm is the strong rank revealing QR algorithm [9, Algorithm 4].

We assign equal relative and absolute error tolerances to `ode15s.m`,

$$atol = rtol = \tau_{ivp}. \tag{4.3}$$

The tolerance for the initial value problem determines both the termination criterion for the nonlinear least squares solver and, most of the time, the number of columns requested from subset selection. We terminate the least squares iteration when the Levenberg–Marquardt iterates $\{p_n\}_{n\geq 0}$ satisfy

$$\|\nabla f(p_n)\| = \|J(p_n)^T R(p_n)\| \leq 10\tau_{ivp} \qquad \text{or} \qquad |f(p_n) - f(p_{n-1})| < 100\tau_{ivp}^2. \tag{4.4}$$

Finally, we request k columns from subset selection, where k is chosen so that

$$\sigma_{k+1} \leq 10\tau_{ivp}\sigma_1 < \sigma_k, \tag{4.5}$$

and σ_k are the singular values of the initial Jacobian $J(p_0)$.

Our examples are based on a driven harmonic oscillator. The corresponding equation is the first order system which is equivalent to the second order initial value problem

$$my'' + cy' + ky = A\sin(\omega t), \qquad y(0) = y_0, \qquad y'(0) = y_0'. \tag{4.6}$$

Here m is the mass, c is the damping coefficient, and k_0 is the spring constant. We integrate the initial value problem from $t = 0$ to $t = 10$. The data are samples of the exact solution at the 100 equally spaced points $t = j/10$, $1 \leq j \leq 100$.

4.1. Forced harmonic oscillator: Low resolution. We set $A = 2$ and $\omega = 5$ in (4.6). This is sufficient to fit the three parameters $p = (m \;\; c \;\; k_0)^T$ if the tolerances for the integrator are sufficiently tight.

We control the Levenberg–Marquardt iteration and the subset selection algorithm with (4.4) and (4.5). In this example, however, we set the tolerance for the integrator we set $\tau_{ivp} = .05$. This led to $k = 2$, and so the Levenberg–Marquardt iteration without subset selection used a truncated SVD, setting the small singular value σ_3 to zero.

Subset selection with $k = 2$ picks c and k_0 as the important variables. Hence the optimization after subset selection had only two unknowns and the Jacobian was 100×2. For the computation using subset selection, we fixed m at its initial value.

We used $p^* = (1 \;\; 1 \;\; 2)^T$ to generate the data and let $p_0 = (1 \;\; .5 \;\; .2)^T$ as the initial iterate. The initial iterate is not terribly far from the solution, but the low tolerance for the integrator will lead to an incorrect computation of the step and cause problems for both traditional Levenberg–Marquardt and the truncated SVD approach. In this example the residual is small, but nonzero. The problem is that the error in the computation of R and R' is large relative to the smallest singular value.

We plot iteration histories for the truncated SVD approach, the method based on subset selection, and a basic Levenberg–Marquardt method where all three singular values are kept. In Figure 4.1 the dashed line is the least squares error and the solid line is the gradient norm.

In the far left and far right plots of Figure 4.1 the least squares errors and the gradient norms decrease for a few iterations and then stagnate because the Levenberg–Marquardt parameter increased rapidly. The iterations failed to converge and terminated when the Levenberg–Marquardt parameter exceeded 10^7.

With subset selection and fixing $m = 1$, on the other hand, the iteration converges and terminates when the gradient norm falls below 10^{-2}. One can see in the central plot of Figure 4.1 the rapid decrease in the least squares error. The converged result for the subset selection algorithm is $(1 \quad 1.003 \quad 2.005)^T$, which is correct up to the tolerances in the integrator. One can also see the concavity in the gradient norm in the terminal phase of the iteration and in the least squares error before the error stagnates at the (nonzero) residual. These are indicators of the rapid convergence one expects for a full-rank problem with a small residual.

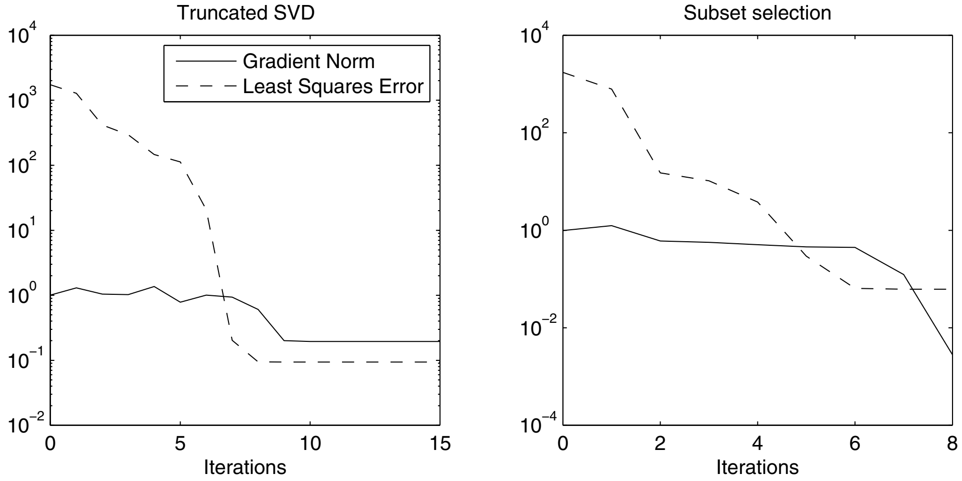

FIG. 4.1. *Forced oscillator: low resolution.*

4.2. Forced harmonic oscillator: Small perturbation in mass. We seek to identify c, k_0, and a small perturbation of the mass. Think of a unit mass with a fly sitting on it; our task is to weigh the fly. In particular, we set $m = 1 + 10^{-3}\delta_m$ and seek to find δ_m. The purpose of the scaling factor in front of δ_m is to ensure that all parameters have the same order of magnitude. This is standard practice and not doing it would make the contribution of subset selection unclear. We also add a redundant parameter by replacing c by a sum $c_1 + c_2$. This has the effect of duplicating a column in the Jacobian and gives subset selection some well-defined work to do. The resulting parameter vector to be fitted is

$$\hat{p} = \begin{pmatrix}\delta_m & c_1 & c_2 & k_0\end{pmatrix}^T \in \mathbb{R}^4.$$

The least squares residual $\hat{R}$ that is a function of $\hat{p}$ and the original residual R are related by

$$\hat{R}(\hat{p}) = R(p), \qquad \text{where} \qquad p = \begin{pmatrix}1 + 10^{-3}\delta_m & c_1 + c_2 & k_0\end{pmatrix}^T$$

and therefore,

$$\hat{R}'(\hat{p}) = \begin{pmatrix} 10^{-3}R_m & R_c & R_c & R_{k_0} \end{pmatrix}(p) = J(p)\begin{pmatrix} 10^{-3} & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}.$$

The first column of the Jacobian is small, which indicates that δ_m is a difficult parameter to resolve. The data come from the exact solution with $\hat{p}^* = (1.23\ \ 1\ \ 0\ \ 1)^T$, but one should keep in mind that any c_1 and c_2 which sum to one will give the same result. We perturbed the data componentwise by multiplying each component by $1 + 10^{-4}r$, where r is a uniformly distributed random vector obtained with the MATLAB `rand` command. The initial iterate was $\hat{p}_0 = (0\ \ 1\ \ 1\ \ .3)^T$. We set the tolerance for the integrator to $\tau_{ivp} = 10^{-8}$ and used the tolerances in (4.4) and (4.5) for the least squares and subset selection computations.

This is no longer a zero-residual problem, so one would not expect the least squares error to converge to zero. Without subset selection the result is $(.091\ \ .5\ \ .5\ \ .998)^T$, and δ_m is completely wrong, even though the convergence history indicates rapid convergence and the iteration terminates with a small gradient norm. In contrast, subset selection produces $(1.28\ \ 0\ \ 1\ \ 1)$, which is a reasonable fit. The iteration history is plotted in Figure 4.2. As in the previous section, the dashed line is the least squares error and the solid line the norm of the gradient. For this example the basic Levenberg–Marquardt iteration is essentially the same as the method using the truncated SVD.

In the previous example the advantage of subset selection was convergence speed and avoiding stagnation. In this example, the convergence is fast for all methods, and the advantage for the subset selection approach is that we get a much more accurate result.

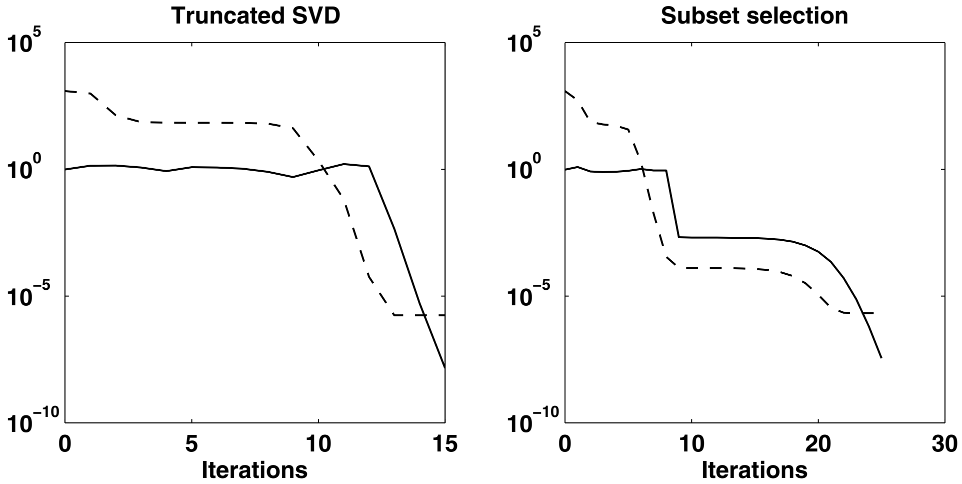

FIG. 4.2. *Forced oscillator: perturbed mass.*

REFERENCES

[1] A. BEN-ISRAEL, *A Newton-Raphson method for the solution of systems of equations*, J. Math. Anal. Appl., 15 (1966), pp. 243–252.

[2] P. T. Boggs, *The convergence of the Ben-Israel iteration for nonlinear least squares problems*, Math. Comp., 30 (1976), pp. 512–522.

[3] S. L. Campbell and C. D. Meyer, *Generalized Inverses of Linear Transformations*, Dover Publications, New York, 1991.

[4] A. R. Conn, N. I. M. Gould, and P. L. Toint, *Trust Region Methods*, MPS-SIAM Ser. Optim. 1, SIAM, Philadelphia, 2000.

[5] J. E. Dennis, Jr., and R. B. Schnabel, *Numerical Methods for Unconstrained Optimization and Nonlinear Equations*, Classics Appl. Math. 16, SIAM, Philadelphia, 1996.

[6] P. Deuflhard and G. Heindl, *Affine invariant convergence theorems for Newton's method and extensions to related methods*, SIAM J. Numer. Anal., 16 (1979), pp. 1–10.

[7] L. M. Ellwein, S. R. Pope, A. Xie, J. Batzel, C. T. Kelley, and M. S. Olufsen, *Modeling cardiovascular and respiratory dynamics in congestive heart failure*, preprint.

[8] G. H. Golub and C. F. Van Loan, *Matrix Computations*, 3rd ed., The Johns Hopkins University Press, Baltimore, MD, 1996.

[9] M. Gu and S. C. Eisenstat, *Efficient algorithms for computing a strong rank-revealing QR factorization*, SIAM J. Sci. Comput., 17 (1996), pp. 848–869.

[10] T. Heldt, *Computational Models of Cardiovascular Response to Orthostatic Stress*, Ph.D. thesis, Massachusetts Institute of Technology, Cambridge, MA, 2004.

[11] I. C. F. Ipsen, *Numerical Matrix Analysis*, SIAM, Philadelphia, 2009.

[12] C. T. Kelley, *Iterative Methods for Linear and Nonlinear Equations*, Frontiers Appl. Math. 16, SIAM, Philadelphia, 1995.

[13] C. T. Kelley, *Iterative Methods for Optimization*, Frontiers Appl. Math. 18, SIAM, Philadelphia, 1999.

[14] K. Levenberg, *A method for the solution of certain non-linear problems in least squares*, Quart. Appl. Math., 4 (1944), pp. 164–168.

[15] D. W. Marquardt, *An algorithm for least squares estimation of nonlinear parameters*, SIAM J. Appl. Math., 11 (1963), pp. 431–441.

[16] J. M. Ortega and W. C. Rheinboldt, *Iterative Solution of Nonlinear Equations in Several Variables*, Academic Press, New York, 1970.

[17] S. Pope, *Parameter Identification in Lumped Compartment Cardiorespiratory Models*, Ph.D. thesis, North Carolina State University, Raleigh, NC, 2009.

[18] S. R. Pope, L. M. Ellwein, C. L. Zapata, V. Novak, C. T. Kelley, and M. S. Olufsen, *Estimation and identification of parameters in a lumped cerebrovascular model*, Math. Biosci. Eng., 6 (2009), pp. 93–115.

[19] R. Schaback, *Convergence analysis of the general Gauss-Newton method*, Numer. Math., 46 (1985), pp. 281–309.

[20] L. F. Shampine and M. W. Reichelt, *The MATLAB ODE suite*, SIAM J. Sci. Comput., 18 (1997), pp. 1–22.

[21] G. W. Stewart, *Error and perturbation bounds for subspaces associated with certain eigenvalue problems*, SIAM Rev., 15 (1973), pp. 727–764.

[22] G. W. Stewart and J.-G. Sun, *Matrix Perturbation Theory*, Academic Press, Boston, 1990.

[23] K. Tanabe, *Continuous Newton-Raphson method for solving an underdetermined system of nonlinear equations*, Nonlinear Analysis, TMA, 3 (1979), pp. 493–501.

[24] K. Tanabe, *Global analysis of continuous analogs of the Levenberg-Marquardt and Newton-Raphson methods for solving nonlinear equations*, Ann. Inst. Statist. Math., 37 (1985), pp. 189–203.