Convergence and Iteration Complexity Analysis of a Levenberg-Marquardt Algorithm for Zero and Non-zero Residual Inverse Problems

E. Bergou * Y. Diouane† V. Kungurtsev ‡

January 28, 2018

Abstract

The Levenberg-Marquardt algorithm is one of the most popular algorithms for the solution of nonlinear least squares problems. In this paper, we propose and analyze the global and local convergence results of a novel Levenberg-Marquardt method for solving general nonlinear least squares problems. The proposed algorithm enjoys strong convergence properties (global convergence as well as local convergence) for least squares problems which do not necessarily have a zero residual solution, all without any additional globalization strategy. Furthermore, we proved worst-case iteration complexity bounds for the proposed algorithm. Preliminary numerical experiments confirm the theoretical behavior of our proposed algorithm.

Keywords: Nonlinear least squares problem, inverse problems, Levenberg-Marquardt method, global and local convergence, worst-case complexity bound, quadratic and linear convergence.

1 Introduction

In this paper we consider the general nonlinear least squares problem

$$\min_{x\in\mathbb{R}^n} \; f(x) = \frac{1}{2}\|F(x)\|^2, \tag{1}$$

where $F : \mathbb{R}^n \to \mathbb{R}^m$ is a (deterministic) vector-valued function, assumed continuously differentiable.We do not assume that there is a solution with zero residual, or that we seek such a solution. In fact, problems of this nature arise in several important practical contexts. One example is inverse problems [16] (e.g., data assimilation [4, 17], full-waveform inversion [20]), where typically an ill-posed nonlinear continuous problem is solved through a discretization. Other examples appear in parameter estimation when a mathematical model approximating a true distribution is fit to given (noisy) data [16, 20]. In all these cases, the resulting least squares problems do not necessarily have a zero residual at any point but may be small.

*MaIAGE, INRA, Université Paris-Saclay, 78350 Jouy-en-Josas, France (elhoucine.bergou@inra.fr).

†Institut Supérieur de l'Aéronautique et de l'Espace (ISAE-SUPAERO), Université de Toulouse, 31055 Toulouse Cedex 4, France (youssef.diouane@isae.fr).

‡Department of Computer Science, Faculty of Electrical Engineering, Czech Technical University in Prague. Support for this author was provided by the Czech Science Foundation project 17-26999S (vyacheslav.kungurtsev@fel.cvut.cz).

Recall that the Gauss-Newton method is an iterative procedure for solving (1) and where at each iterate x_j a step is computed as a solution to the linearized least squares subproblem

$$\min_{s\in\mathbb{R}^n} \ \frac{1}{2}\|F_j + J_j s\|^2,$$

where $F_j = F(x_j)$ and $J_j = J(x_j)$ denotes the Jacobian of F at x_j. The subproblem has a unique solution if J_j has full column rank, and in that case the step is a descent direction for f.

The Levenberg-Marquardt method [11, 12] (see also [15]) was developed to deal with the rank deficiency of J_j and to provide a globalization strategy for Gauss-Newton. At each iteration a step is considered of the form $-(J_j^\top J_j + \gamma_j I)^{-1} J_j^\top F_j$, corresponding to the unique solution of

$$\min_{s\in\mathbb{R}^n} \ m_j(s) = \frac{1}{2}\|F_j + J_j s\|^2 + \frac{1}{2}\gamma_j\|s\|^2, \tag{2}$$

where $\gamma_j > 0$ is an appropriately chosen regularization parameter.

In this paper, we will present and analyze the global and local convergence results of a novel Levenberg-Marquardt method for solving general nonlinear least squares problems. In particular we present a Levenberg-Marquardt updating strategy that carefully balances the opposing objectives of ensuring global convergence and stabilizing a Newton local convergence regime.

The strongest results for local convergence of Levenberg-Marquardt are given in a series of papers beginning with [21] (such as [8] and [5], see also [6]), wherein it is assumed that the solution satisfies $F(x) = 0$. The algorithm we present matches this rate for zero-residual problems. In the case of non-zero residuals, it has been found that the implementations of Levenberg-Marquardt converge locally at a linear rate if the norm of the residual is sufficiently small and the parameter γ_j goes to zero [10]. Our proof of linear convergence is simpler than in [10] as well as more completely complementing the other convergence results. Furthermore [10] only show superlinear convergence for zero-residual problems and only present global convergence for exact solutions of subproblems. They also present additional analysis in regards for finite arithmetic that is different in scope of our work, however.

It should be noted that in general the goals of encouraging global and local convergence compete against each other. Namely, the regularization parameter appearing in the subproblem should be allowed to become arbitrarily large in order to encourage global convergence, to ensure the local accuracy of the linearized subproblem, but the parameter must approach zero in order to function as a stabilizing regularization that encourages fast local convergence.

For example, in the original presentation of the Levenberg-Marquardt method in [11, 12], γ_j is not permitted to go to zero, and only global convergence is considered. By contrast, in [21], for instance, there is a two-phase method where quadratic decline in the residual is tested with each step that is otherwise globalized by a line-search procedure. Two phase methods, while also less mathematically elegant, are practically inefficient and challenging to implement in the sense that it can be difficult to properly ascertain when the region of local convergence is reached.

Our parameter updating strategy is inspired by [7], which presents a Levenberg-Marquardt method inspired from trust-region algorithm for zero residual least squares problems. The trust-region radius is updated as $\Delta_{j+1} = \mu\|F_{j+1}\|^\delta$, with $\delta \in (1/2; 1)$ and μ is updated according to classical global convergence updating strategies, the residual F_{j+1} is included to enforce local convergence. They show global and superlinear local convergence properties for their method. We extend the results outlined above in scope in the sense of showing the convergence properties for residual problems which do not necessarily have a zero residual solution, as well as in elegance in that the method is purely a Levenberg-Marquardt method, with no additional globalization

strategies, and thus is seamless and is an extension of the standard classical approach to least-squares problems, as well as improve the convergence rate to quadratic.

Furthermore, we establish a worst-case complexity analysis of our proposed algorithm. In fact, given a tolerance $\epsilon \in (0, 1)$, we aim at estimating the number of iterations needed to reach an iterate x_j such that

$$\|\nabla f(x_j)\| < \epsilon. \tag{3}$$

Worst-case iteration complexity bounds of Levenberg-Marquardt methods applied to non-linear least squares problems using specific schemes of update for γ_j can be found in [18, 19, 22]. Up to a logarithmic factor, we show that our proposed algorithm has a complexity bounds that matches results previously cited Levenberg-Marquardt algorithms. Precisely, we obtain an iteration complexity bound in $\tilde{\mathcal{O}}\left(\epsilon^{-2}\right)$, where the notation $\tilde{\mathcal{O}}(\cdot)$ indicates the presence of logarithmic factors in ϵ. The logarithmic factor in our complexity analysis is due to our strategy of updating γ_j to ensure both global convergence and fast local convergence. We note that in [18, 19, 22] only the global convergence is shown.

1.1 Summary of Contributions

Our contribution amounts to the following. Whereas with the state of the art, in choosing a Levenberg-Marquardt method, one has four choices with regards to convergence and complexity guarantees, in particular

1. global convergence with the exact subproblem solution and a quadratic local convergence for zero-residual problems, or
2. global convergence with the exact and inexact subproblem solution for any (zero or nonzero residual) inverse problem, or,
3. global convergence and iteration complexity results for any inverse problem, or
4. global convergence, linear local convergence for non-zero residual problems, and superlinear convergence for zero-residual problems.

In this paper, we present a *seamless* (i.e., no separate phases) Levenberg-Marquardt method that simultaneously achieves,

1. global convergence for exact and inexact subproblem solutions for any inverse problem
2. iteration complexity results for any inverse problem,
3. linear local convergence rate for non-zero residual inverse problems,
4. quadratic local convergence rate for zero-residual inverse problems,

and thus match all of the best state of the art results in the literature with just one algorithm. This is with a relatively simple procedure and standard proof techniques. In addition we present a thorough set of numerical results demonstrating the order of convergence as well as the global convergence properties of the method.

1.2 Outline and Notation

The outline of this paper is as follows. In Section 2 we present the proposed Levenberg-Marquardt algorithm for solving general nonlinear least squares problems. Section 3 addresses the inexact solution of the linearized least squares subproblems arising within the Levenberg-Marquardt method. In Section 4, we show the global convergence of our algorithm. Section 5 describes a worst-case complexity analysis of the proposed method. In Section 6 we derive the overall local convergence analysis of the proposed algorithm. In Section 7, preliminary numerical experiments with basic implementations are presented that show the good behavior of our novel algorithm. Finally, in Section 8 we draw some perspectives and conclusions.

Throughout this paper $\|\cdot\|$ will denote the vector or matrix l_2-norm.

2 A novel Levenberg-Marquardt algorithm

In deciding whether to accept a step s_j generated by the subproblem (2), the Levenberg-Marquardt method can be seen as precursor of the trust-region method [3]. In fact, it seeks to determine when the Gauss-Newton step is applicable (in which case the regularization parameter is set to zero) or when it should be replaced by a slower but safer steepest descent step (corresponding to a sufficiently large regularization parameter). For that purpose, one considers the ratio between the actual reduction $f(x_j) - f(x_j + s_j)$ attained in the objective function and the reduction $m_j(0) - m_j(s_j)$ predicted by the model, given by

$$\rho_j \;=\; \frac{f(x_j) - f(x_j + s_j)}{m_j(0) - m_j(s_j)}.$$

Then, if ρ_j is sufficiently above zero, the step is accepted and γ_j is possibly decreased. Otherwise the step is rejected and γ_j is increased.

In this paper we consider the choice of the regularization parameter as $\gamma_j = \mu\|\nabla f(x_j)\|^2$. where μ is updated according to the ratio ρ_j. The considered Levenberg-Marquardt algorithm is described below.

Algorithm 1: Levenberg-Marquardt algorithm.

Initialization

Choose the constants $\eta \in (0,1)$, $\mu_{\min} > 0$ and $\lambda > 1$. Select x_0 and $\mu_0 \geq \mu_{\min}$. Set $\gamma_0 = \mu_0\|\nabla f(x_0)\|^2$ and $\bar{\mu} = \mu_0$.

For $j = 0, 1, 2, \ldots$

1. Solve (or approximately solve) (2), and let s_j denote such a solution.
2. Compute $\rho_j = \frac{f(x_j) - f(x_j + s_j)}{m_j(0) - m_j(s_j)}$.
3. If $\rho_j \geq \eta$, then set $x_{j+1} = x_j + s_j$ and $\mu_{j+1} \in [\max(\mu_{\min}, \bar{\mu}/\lambda), \bar{\mu}]$ and $\bar{\mu} = \mu_{j+1}$. Otherwise, set $x_{j+1} = x_j$ and $\mu_{j+1} = \lambda\mu_j$.
4. Compute $\gamma_{j+1} = \mu_{j+1}\|\nabla f(x_{j+1})\|^2$.

A brief remark is warranted regarding the new step acceptance criteria. Note that we have an auxiliary parameter $\bar{\mu}$, that represents the last good parameter. This is necessary in order to

balance the requirements of global and local convergence. If the model is relatively inaccurate, then μ_j is driven higher, however, when we reach a region of local convergence, we need the parameter γ_j to encourage local convergence, and thus bound μ_j so that the component $\|\nabla f(x)\|$ dominates the behavior of the parameter γ_j.

3 Inexact solution of the linearized subproblems

Step 1 of Algorithm 1 requires the approximate solution of subproblem (2). As in trust-region methods, there are different techniques to approximate the solution of this subproblem yielding a globally convergent step. For the purposes of global convergence it is sufficient to compute a step s_j that provides a reduction in the model as good as the one produced by the so-called Cauchy step (defined as the minimizer of the model along the negative gradient).

The Cauchy step is defined by minimizing $m_j(x_j - t\nabla f(x_j))$ when $t > 0$ and is given by

$$s_j^{\mathrm{c}} \;=\; -\frac{\|\nabla f(x_j)\|^2}{\nabla f(x_j)^\top (J_j^\top J_j + \gamma_j I)\nabla f(x_j)} \nabla f(x_j). \tag{4}$$

The corresponding Cauchy decrease of the model is

$$m_j(0) - m_j(s_j^{\mathrm{c}}) \;=\; \frac{1}{2} \frac{\|\nabla f(x_j)\|^4}{\nabla f(x_j)^\top (J_j^\top J_j + \gamma_j I)\nabla f(x_j)}.$$

Since $\nabla f(x_j)^\top (J_j^\top J_j + \gamma_j I)\nabla f(x_j) \leq \|\nabla f(x_j)\|^2(\|J_j\|^2 + \gamma_j)$, we conclude that

$$m_j(0) - m_j(s_j^{\mathrm{c}}) \;\geq\; \frac{1}{2} \frac{\|\nabla f(x_j)\|^2}{\|J_j\|^2 + \gamma_j}.$$

The Cauchy step (4) is cheap to calculate as it does not require any system solve. Moreover, the Levenberg-Marquardt method will be globally convergent if it uses a step that attains a reduction in the model as good as a multiple of the Cauchy decrease. Thus we will impose the following assumption on the step calculation:

Assumption 3.1 *For every step j,*

$$m_j(0) - m_j(s_j) \;\geq\; \frac{\theta_{fcd}}{2} \frac{\|\nabla f(x_j)\|^2}{\|J_j\|^2 + \gamma_j}$$

for some constant $\theta_{fcd} > 0$.

Despite providing a sufficient reduction in the model and being cheap to compute, the Cauchy step is a particular form of steepest descent. In practice, a version of Algorithm 1 solely based on the Cauchy step would suffer from the same drawbacks as the steepest descent algorithm on ill-conditioned problems. One can see that the Cauchy step depends on $J_j^\top J_j$ only in the step length. Faster convergence can be expected if the matrix $J_j^\top J_j$ also influences the step direction.

Since the Cauchy step is the first step of the conjugate gradient method (CG) when applied to the minimization of the quadratic $s \to m_j(s)$, it is natural to propose to run the CG further and stop only when the residual becomes relatively small. The truncated-CG step is of the form:

$$s_j^{\mathrm{cg}} \;=\; V_j \left(V_j^\top (J_j^\top J_j + \gamma_j I) V_j \right)^{-1} V_j^\top \nabla f(x_j), \tag{5}$$

where V_j is a given unitary matrix whose first column is given by $-\nabla f(x_j)/\|\nabla f(x_j)\|$.

Since the CG generates iterates by minimizing the quadratic model over nested Krylov subspaces, and the first subspace is the one generated by $\nabla f(x_j)$ (see, e.g., [14, Theorem 5.2]), the decrease attained at the first CG iteration (i.e., by the Cauchy step) is kept by the remaining iterations. Thus Assumption 3.1 holds for all the iterates $s_j^{\rm cg}$ generated by the truncated-CG whenever it is initialized by the null vector.

The following lemma is similar to [1, Lemma 5.1] and will be useful for our global convergence analysis.

Lemma 3.1 *For the three steps proposed (exact, Cauchy, and truncated-CG), one has that*

$$\|s_j\| \;\leq\; \frac{\|\nabla f(x_j)\|}{\gamma_j} \;=\; \frac{1}{\mu_j\|\nabla f(x_j)\|}$$

and

$$|s_j^\top(\gamma_j s_j + \nabla f(x_j))| \;\leq\; \frac{\|J_j\|^2\|\nabla f(x_j)\|^2}{\gamma_j^2} \;=\; \frac{\|J_j\|^2}{\mu_j^2\|\nabla f(x_j)\|^2}.$$

Proof. We will omit the indices j in the proof. We note that the truncated-CG step can be seen as a generalized step of both exact and Cauchy steps. In fact, the first CG iteration produces the Cauchy step while the last iteration gives the exact one.

Thus, without loss of generality, for the three proposed steps there exists a unitary matrix V with first column given by $-\nabla f(x)/\|\nabla f(x)\|$ and such that

$$s \;=\; V\left(V^\top(J^\top J+\gamma I)V\right)^{-1}V^\top\nabla f(x) \;=\; V\left(V^\top J^\top JV+\gamma I\right)^{-1}\|\nabla f(x)\|e_1,$$

where e_1 is the first vector of the canonical basis of $\mathbb{R}^n$. From the positive semidefiniteness of $V^\top J^\top JV$, we immediately obtain $\|s\| \leq \|\nabla f(x)\|/\gamma$.

To prove the second inequality of Lemma 3.1, we apply the Sherman–Morrisson–Woodbury formula and obtain

$$s \;=\; V\left(\frac{1}{\gamma}I - \frac{1}{\gamma^2}(JV)^\top\left(I+\frac{(JV)(JV)^\top}{\gamma}\right)^{-1}(JV)\right)\|\nabla f(x)\|e_1.$$

Since $Ve_1 = -\nabla f(x)/\|\nabla f(x)\|$,

$$\gamma s + \nabla f(x) \;=\; -\frac{1}{\gamma}V(JV)^\top\left(I+\frac{(JV)(JV)^\top}{\gamma}\right)^{-1}(JV)\|\nabla f(x)\|e_1.$$

Now, from the fact that $(JV)(JV)^\top/\gamma$ is positive semidefinite, the norm of the inverse of $I + (JV)(JV)^\top/\gamma$ is less than one, and thus (since V is unitary)

$$\|\gamma s + \nabla f(x)\| \;\leq\; \frac{\|J\|^2\|\nabla f(x)\|}{\gamma}.$$

Finally,

$$|s^\top(\gamma s+\nabla f(x))| \;\;\leq\;\; \|s\|\|\gamma s+\nabla f(x)\| \;\;\leq\;\; \frac{\|J\|^2\|\nabla f(x)\|^2}{\gamma^2}.$$

■

4 Global convergence

We start by giving some classical assumptions and then state and prove some lemmas that later will appear in the global convergence analysis.

Assumption 4.1 *The function f is continuously differentiable in an open set containing $L(x_0) = \{x \in \mathbb{R}^n : f(x) \leq f(x_0)\}$ with Lipschitz continuous gradient on $L(x_0)$ and corresponding constant $\nu > 0$.*

Assumption 4.2 *The Jacobian J of F is uniformly bounded, i.e., there exists $\kappa_J > 0$ such that $\|J_j\| \leq \kappa_J$ for all j.*

Note that from the previous two assumptions we conclude that, the gradient of f is uniformly bounded, i.e., there exists $\kappa_g > 0$ such that $\|\nabla f(x_j)\| \leq \kappa_g$ for all j.

The next lemma says that, if we suppose that the gradient norm is bounded below by a non zero constant $g_{\min}$, then for a value of the parameter μ_j sufficiently large, the step is accepted by the acceptance criterion.

Lemma 4.1 *Let Assumptions 3.1, 4.1, and 4.2 hold. Suppose that, for all the iterations j, there exists a bound $g_{\min} > 0$ such that $\|\nabla f(x_j)\| \geq g_{\min}$. Then , one has*

$$\lim_{\mu_j \to \infty} \rho_j = 2.$$

Proof. By applying a Taylor expansion, one has

$$\begin{aligned} 1 - \frac{\rho_j}{2} &= \frac{m_j(0) - f(x_j) + f(x_j + s_j) - m_j(s_j) + m_j(0) - m_j(s_j)}{2[m_j(0) - m_j(s_j)]} \\ &= \frac{R - s_j^\top \nabla f(x_j) - s_j^\top (J_j^\top J_j + \gamma_j I) s_j}{2[m_j(0) - m_j(s_j)]} \\ &= \frac{R - s_j^\top (J_j^\top J_j) s_j - s_j^\top (\gamma_j s_j + \nabla f(x_j))}{2[m_j(0) - m_j(s_j)]} \qquad \text{where } R \leq \frac{\nu}{2}\|s\|^2. \end{aligned}$$

Then, using Lemma 3.1, Assumptions 3.1 and 4.1, one gets

$$|1 - \frac{\rho_j}{2}| \leq \frac{\frac{\nu}{2}\|s_j\|^2 + \|J_j\|^2\|s_j\|^2 + |s_j^\top(\gamma_j s_j + \nabla f(x_j))|}{\frac{\theta_{fcd}\|\nabla f(x_j)\|^2}{\|J_j\|^2 + \gamma_j}} \leq \frac{\left(\frac{\nu}{2} + 2\kappa_J^2\right)}{\theta_{fcd}} \frac{\kappa_J^2 + \gamma_j}{\gamma_j^2} \tag{6}$$

Since $\gamma_j = \mu_j \|\nabla f(x_j)\|^2$, one deduces that

$$\begin{aligned} |1 - \frac{\rho_j}{2}| &\leq \frac{\left(\frac{\nu}{2} + 2\kappa_J^2\right)}{\theta_{fcd}} \frac{\kappa_J^2 + \mu_j\|\nabla f(x_j)\|^2}{\mu_j^2 \|\nabla f(x_j)\|^4} \\ &\leq \frac{\left(\frac{\nu}{2} + 2\kappa_J^2\right)}{\theta_{fcd}} \frac{\kappa_J^2 + \mu_j \kappa_g^2}{\mu_j^2 g_{\min}^4}. \end{aligned}$$

■

Now, we can show our main global convergence result

Theorem 4.1 *Under Assumptions 3.1, 4.1, and 4.2, the sequence $\{x_j\}$ generated by Algorithm 1 satisfies*

$$\liminf_{j\to\infty} \|\nabla f(x_j)\| \;=\; 0.$$

Proof. By contradiction, if the theorem is not true, then there exists a bound $g_{\min} > 0$ such that

$$\|\nabla f(x_j)\| \geq g_{\min}, \quad \forall j \geq 0.$$

Define $\mathbb{S} = \{j \in \mathbb{N} |\;\; \rho_j \geq \eta\}$ as the set of successful iterations. Hence for $j \in \mathbb{S}$, one has

$$\begin{aligned}
\|F_j\|^2 - \|F_{j+1}\|^2 &\geq \eta(m_j(0) - m_j(s_j)) \\
&\geq \eta\frac{\theta_{fcd}}{2}\frac{\|\nabla f(x_j)\|^2}{\|J_j\|^2 + \gamma_j} \qquad \text{(using Assumption 3.1)} \\
&\geq \frac{\eta\theta_{fcd}}{2}\frac{\|\nabla f(x_j)\|^2}{\kappa_J^2 + \mu_j\|\nabla f(x_j)\|^2} \\
&\geq \frac{\eta\theta_{fcd}}{2}\frac{g_{\min}^2}{\kappa_J^2 + \mu_j\kappa_g^2}
\end{aligned}$$

If $\mathbb{S}$ is infinite, then since $\sum_{j\in\mathbb{S}} \|F_j\|^2 - \|F_{j+1}\|^2$ is finite, we deduce that

$$\lim_{j\to\infty} \frac{\eta\theta_{fcd}}{2}\frac{g_{\min}^2}{\kappa_J^2 + \mu_j\kappa_g^2} = 0,$$

hence $\lim_{j\to\infty} \mu_j = +\infty$.

Otherwise (i.e., $\mathbb{S}$ is finite), from Algorithm 1 we have $\mu_{j+1} = \lambda\mu_j$ for all sufficiently large j. Since $\lambda > 1$, we deduce that $\lim_{j\to\infty} \mu_j = +\infty$.

Thus, using Lemma 4.1, we have $\lim_{j\to\infty} \rho_j = 2$. Since when $\rho_j \geq \eta$ we decrease μ_j then, there exists a positive constant $\mu_{\max}$ such that $\mu_j \leq \mu_{\max}$ holds for all sufficiently large j. Which leads to a contradiction with the fact that μ_j goes to infinity when j goes to $+\infty$. ■

5 Worst-case complexity results

We now establish a worst-case complexity bound of Algorithm 1. We begin by deriving a condition on the parameter μ_j for an iteration to be successful.

Lemma 5.1 *Let Assumptions 3.1, 4.1, and 4.2 hold. Suppose that at the j-th iteration of Algorithm 1, one has*

$$\mu_j \;>\; \frac{\kappa}{\|\nabla f(x_j)\|^2} \tag{7}$$

where

$$\kappa = \frac{a + \sqrt{a^2 + 4a\kappa_J^2(1-\eta_1)}}{2(1-\eta_1)}, \quad a = \frac{\frac{\nu}{2} + 2\kappa_J^2}{\theta_{fcd}}.$$

Then, the iteration is successful.

Proof. Recall that by classical Taylor expansion formulas, one has:

$$\begin{aligned}
f(x_j + s_j) - m_j(x_j + s_j) &= f(x_j + s_j) - f(x_j) - \nabla f(x_j)^\top s_j - \frac{1}{2}s_j^\top J_j^\top J_j s_j - \frac{\gamma_j}{2}\|s_j\|^2 \\
&\leq \frac{\nu}{2}\|s_j\|^2 - \frac{1}{2}s_j^\top J_j^\top J_j s_j - \frac{\gamma_j}{2}s_j^\top s_j.
\end{aligned}$$

In addition, the definition of $m_j(\cdot)$ yields:

$$m_j(x_j) - m_j(x_j + s_j) = -\nabla f(x_j)^\top s_j - \frac{1}{2} s_j^\top J_j^\top J_j s_j - \frac{\gamma_j}{2} s_j^\top s_j.$$

As a result, similarly to (6) in the proof of Lemma 4.1, one has

$$|1 - \frac{\rho_j}{2}| \quad \leq \quad \frac{\frac{\nu}{2} + 2\kappa_J^2}{\theta_{fcd}} \frac{\kappa_J^2 + \gamma_j}{\gamma_j^2}.$$

Considering the later expression and using $a = \frac{\frac{\nu}{2}+2\kappa_J^2}{\theta_{fcd}}$, we have

$$\frac{\frac{\nu}{2} + 2\kappa_J^2}{\theta_{fcd}} \frac{\kappa_J^2 + \gamma_j}{\gamma_j^2} \geq (1 - \eta_1) \;\Rightarrow\; 0 \geq (1 - \eta_1)\gamma_j^2 - a\gamma_j - a\kappa_J^2.$$

Treating the right-hand side as a second-order polynomial in γ_j yields

$$\gamma_j \leq \frac{a + \sqrt{a^2 + 4a\kappa_J^2(1 - \eta_1)}}{2(1 - \eta_1)} \iff \mu_j \leq \frac{\kappa}{\|\nabla f(x_j)\|^2},$$

which contradicts (7). We thus conclude that

$$\frac{\frac{\nu}{2} + 2\kappa_J^2}{\theta_{fcd}} \frac{\kappa_J^2 + \gamma_j}{\gamma_j^2} < (1 - \eta_1),$$

which in turns implies that $|1 - \frac{\rho_j}{2}| \leq 1 - \eta_1$, hence $\rho_j \geq \eta_1$, therefore the iteration is successful. ■

Our next result states that when the gradient norm stays bounded away from zero, the parameter μ_j cannot grow infinitely. Without loss of generality we assume that $\epsilon \leq \sqrt{\frac{\lambda\kappa}{\mu_0}}$, where κ is the same as in the previous lemma.

Lemma 5.2 *Under Assumptions 3.1, 4.1, and 4.2, let j be a given iteration index such that for every $l \leq j$, $\|\nabla f(x_l)\| > \epsilon$ where $\epsilon \in (0, 1)$. Then, for every $l \leq j$, one also has*

$$\mu_l \leq \mu_{\max} := \max\left\{\mu_0, \frac{\lambda\kappa}{\epsilon^2}\right\} = \frac{\lambda\kappa}{\epsilon^2}, \tag{8}$$

Proof. We prove this result by contradiction. Suppose that $l \geq 1$ is the first index such that

$$\mu_l > \frac{\lambda\kappa}{\epsilon^2}. \tag{9}$$

By the updating rules on μ_l, one has either the iteration $l - 1$ is successful hence $\mu_l \leq \mu_0 \leq \frac{\lambda\kappa}{\epsilon^2}$ which contradicts (9), or the iteration $l - 1$ is unsuccessful hence

$$\mu_l = \lambda\mu_{l-1} \quad \Rightarrow \quad \mu_{l-1} = \frac{\mu_l}{\lambda} > \frac{\kappa}{\epsilon^2} > \frac{\kappa}{\|\nabla f(x_l)\|^2},$$

therefore using Lemma 7 implies that the $l - 1$-th iteration is successful which leads to contradiction again. ■

Thanks to Lemma 5.2, we can now bound the number of successful iterations needed to drive the gradient norm below a given threshold.

Proposition 5.1 *Under Assumptions 3.1, 4.1, and 4.2, let $\epsilon \in (0,1)$ and let k_ϵ be the first iteration index such that $\|\nabla f(x_{k_\epsilon+1})\| < \epsilon$.*

Then, if S_ϵ is the set of indexes of successful iterations prior to k_ϵ, one has:

$$|S_\epsilon| \;\leq\; \mathcal{C}\epsilon^{-2}, \tag{10}$$

with

$$\mathcal{C} = \frac{2\left(\kappa_J^2 + \lambda\kappa\right)}{\eta_1\theta_{fcd}} f(x_0).$$

Proof. For any $j \in S_\epsilon$, one has

$$\begin{aligned} f(x_j) - f(x_{j+1}) &\geq \eta_1 \left(m_j(x_j) - m_j(x_{j+1})\right) \\ &\geq \eta_1 \frac{\theta_{fcd}}{2} \frac{\|\nabla f(x_j)\|^2}{\kappa_J^2 + \gamma_j} \\ &\geq \eta_1 \frac{\theta_{fcd}}{2} \frac{\|\nabla f(x_j)\|^2}{\kappa_J^2 + \mu_j\|\nabla f(x_j)\|^2} \\ &\geq \eta_1 \frac{\theta_{fcd}}{2} \frac{\|\nabla f(x_j)\|^2}{\kappa_J^2 + \frac{\lambda\kappa}{\epsilon^2}\|\nabla f(x_j)\|^2}. \end{aligned}$$

Using now the assumption that $\|\nabla f(x_j)\| \geq \epsilon$, we arrive at

$$\begin{aligned} f(x_j) - f(x_{j+1}) &\geq \eta_1 \frac{\theta_{fcd}}{2} \frac{\|\nabla f(x_j)\|^2}{\left(\kappa_J^2 + \lambda\kappa\right)\frac{\|\nabla f(x_j)\|^2}{\epsilon^2}} \\ &= \eta_1 \frac{\theta_{fcd}}{2} \frac{\epsilon^2}{\kappa_J^2 + \lambda\kappa}. \end{aligned}$$

Consequently, by summing on all iteration indices within S_ϵ and using the fact that f is bounded below by 0, we obtain

$$f(x_0) - 0 \geq \sum_{j=0}^{k_\epsilon} f(x_j) - f(x_{j+1}) \geq \sum_{j\in S_\epsilon} f(x_j) - f(x_{j+1}) \geq |S_\epsilon| \frac{\eta_1\theta_{fcd}}{2\left(\kappa_J^2 + \lambda\kappa\right)}\epsilon^2,$$

hence the result. ■

Lemma 5.3 *Under the assumptions of Proposition 5.1, let U_ϵ denote the set of unsuccessful iterations of index less than or equal to k_ϵ. Then,*

$$|U_\epsilon| \;\leq\; \log_\lambda\left(\frac{\kappa}{\mu_{\min}\epsilon^2}\right) |S_\epsilon|. \tag{11}$$

Proof. Note that we necessarily have $k_\epsilon \in S_\epsilon$ (otherwise $\|\nabla f(x_{k_\epsilon})\| < \epsilon$, which would contradict the definition of k_ϵ).

Our objective is to bound the number of unsuccessful iterations between two successful ones. Let thus $\{j_0, \ldots, j_t = k_\epsilon\}$ be an ordering of S_ϵ, and $i \in \{0, t-1\}$.

Due to the updating formulas for μ_j on successful iterations, we have:

$$\mu_{j_i+1} \geq \max\{\mu_{\min}, \bar{\mu}/\lambda\} \geq \mu_{\min}.$$

Moreover, we have $\|\nabla f(x_{j_i+1})\| \geq \epsilon$ by assumption.

By Lemma 5.1, for any unsuccessful iteration $j \in \{j_i + 1, \ldots, j_{i+1} - 1\}$, we then must have:

$$\mu_j \leq \frac{\kappa}{\epsilon^2},$$

since otherwise $\mu_j > \frac{\kappa}{\epsilon^2} \geq \frac{\kappa}{\|\nabla f(x_j)\|^2}$ and the iteration would be successful.

Using the updating rules for μ_j on unsuccessful iterations, we obtain:

$$\forall j = j_i + 1, \ldots, j_{i+1} - 1, \qquad \mu_j = \lambda^{j-j_i-1}\mu_{j_i+1} \geq \lambda^{j-j_i-1}\mu_{\min}.$$

Therefore, the number of unsuccessful iterations between j_i and j_{i+1}, equal to $j_{i+1} - j_i - 1$, satisfies:

$$j_{i+1} - j_i - 1 \;\leq\; \log_\lambda \left(\frac{\kappa}{\mu_{\min}\epsilon^2} \right). \tag{12}$$

By considering (12) for $i = 0, \ldots, t-1$, we arrive at

$$\sum_{i=0}^{t-1} (j_{i+1} - j_i - 1) \;\leq\; \log_\lambda \left(\frac{\kappa}{\mu_{\min}\epsilon^2} \right) [|S_\epsilon| - 1]. \tag{13}$$

What is left to bound is the number of possible unsuccessful iterations between the iteration of index 0 and the first successful iteration j_0. Since $\mu_0 \geq \mu_{\min}$, a similar reasoning as the one used to obtain (12) leads to

$$j_0 - 1 \leq \log_\lambda \left(\frac{\kappa}{\mu_{\min}\epsilon^2} \right). \tag{14}$$

Putting (13) and (14) the expected result. ■

By combining the results from Proposition 5.1 and Lemma 5.3, we thus get the following complexity estimate.

Theorem 5.1 *Let the assumptions of Proposition 5.1 hold, and let $\epsilon \in (0, 1)$. Then, the first index k_ϵ for which $\|\nabla f(x_{k_\epsilon+1})\| < \epsilon$ is bounded above by*

$$\mathcal{C}\left(1 + \log_\lambda \left[\frac{\kappa}{\mu_{\min}\epsilon^2} \right] \right) \epsilon^{-2}, \tag{15}$$

where $\mathcal{C}$ is the constant defined in Proposition 5.1.

For the Levenberg-Marquardt method proposed in this paper, we thus obtain an iteration complexity bound in $\tilde{\mathcal{O}}\left(\epsilon^{-2}\right)$, where the notation $\tilde{\mathcal{O}}(\cdot)$ indicates the presence of logarithmic factors in ϵ. Note that the evaluation complexity bounds are of the same order.

In the case where the problem has zero residuals as a consequence of Theorem 5.1, if the Jacobian matrix is uniformly non singular over the iterate sequence, we can also provide a complexity bound on the number of iterations needed to drive the residual below a given threshold. We note that in this case, the non-singularity if the Jacobian matrix at the solution implies that the minimization problem has zero residuals. A similar result is given in [18].

Corollary 5.1 *Let the assumptions of Theorem 5.1 hold, and suppose further that there exists $\sigma > 0$ such that $\forall j,\ \lambda_{\min}(J_j^\top J_j) \geq \sigma^2$. Then, for any $\hat{\epsilon} \in (0, 1)$, the number of iterations required by Algorithm 1 to reach an iterate for which $\|F(x_j)\| < \hat{\epsilon}$ is at most*

$$\mathcal{C}\left(1 + \log_\lambda \left[\frac{\kappa}{\mu_{\min}\sigma^2\hat{\epsilon}^2} \right] \right) \sigma^{-2}\hat{\epsilon}^{-2}. \tag{16}$$

Proof. By assumption, for every iterate x_j, one has:

$$\|\nabla f(x_j)\| = \left\| J_j^\top F(x_j) \right\| \geq \sigma \|F(x_j)\|.$$

Therefore, letting $\epsilon = \sigma\,\hat{\epsilon}$, we have

$$\|\nabla f(x_j)\| < \epsilon \;\Rightarrow\; \|F(x_j)\| < \hat{\epsilon}.$$

Applying Theorem 5.1 with this particular choice of ϵ yields the desired result. ■

Up to a logarithmic factor, the complexity bound obtained in Theorem 5.1 matches results previously obtained bounds for the Levenberg-Marquardt algorithms of Ueda and Yamashita [18] as well as that of Zhao and Fan [22]. Note that the latter also uses the regularization $\gamma_j = \mu_j \|\nabla f(x_j)\|$, while the former directly updates the γ_j parameter (without decrease). However, the updating rules used in Algorithm 1 do not relate the value of μ_j on successful and very successful iterations, which seems to be the cause for the additional logarithmic factor. In particular, for a successful iteration, μ_j is kept unchanged if it is seen as large compared to the norm of the gradient of the merit function. Our updating rule of μ_j could be modified in the same manner to be closer to standard methods [18, 22], in order to get rid of the logarithmic dependence; however, the method will not behave purely as Levenberg-Marquardt method (with no additional globalization strategies) with strong local convergence properties. In the next section we present the local convergence analysis of Algorithm 1.

6 Local convergence

As mentioned in Section 1, the state of the art for problems with zero residual at the solution is that local convergence at a quadratic rate holds under a Lipschitz continuity and error bound assumptions. We will replicate this in a manner appropriate for problems without a zero residual at the solution, and subsequently require only one separate split in the analysis with the final Lemma to distinguish the convergence rate for zero and non-zero residual problems. In the sequel of this section, the considered step is the exact solution of subproblem (2).

6.1 Assumptions

In this setting, $\|F(x)\|$ is no longer an appropriate measure for the distance to the solution. Stationarity is associated with a zero gradient, and, as can be gleaned from the form of our update for the regularization parameter γ_j in Algorithm 1, this is what we use for the regularization to encourage fast convergence.

In the case of zero-residual problems, there is a set of solutions to $F(x) = 0$ and the purpose of the algorithm is to obtain a point at which the residual is zero. In this case, we seek a stationary solution where $\nabla f(x) = 0$, however there can be multiple sets of stationary points, with varying objective values. As the behavior of the algorithm is such that both descent of $f(x)$ encouraged as well as a solution to stationarity is sought, for a clear picture of the convergence, we instead propose to consider a particular subset with a constant value of the objective.

Assumption 6.1 *There exists a connected isolated set X^* composed of stationary points to* (1)*, and Algorithm 1 generates a sequence with an accumulation point $x^* \in X^*$.*

We shall denote by $\bar{F}$ the value of F at any $\bar{x} \in X^*$. Note that this is unique, as X^* is a connected set of stationary points, so there is no direction of ascent for $f(x)$ among the set of directions feasible within X^*.

Henceforth, from the global convergence analysis, we can assume, without loss of generality, that there exists a subsequence approaching this X^*. This subsequence need not be unique, i.e., there may be more than one subsequence converging to separate connected sets of stationary points. We shall see that eventually, one of these sets shall "catch" the subsequence and result in direct convergence to the solution set at a quadratic rate.

In the sequel $N(x, \delta)$ denotes the closed ball of center x (a given vector) and radius $\delta > 0$. $\text{dist}(x, X^*)$ denotes the distance between the vector x and the set X^*, i.e.,

$$\text{dist}(x, X^*) = \min_{y \in X^*} \|x - y\|.$$

Next, we detail the required assumptions to guarantee the good local convergence rate of our proposed algorithm.

Assumption 6.2 *It holds that $F(x)$ and $J(x)$ are both locally Lipschitz continuous around $x^* \in X^*$ with x^* satisfying Assumption 6.1. In particular this implies, letting $\bar{x} = \text{argmin}_{y \in X^*} \|x - y\|$, that there exists $\delta_1 > 0$ such that for $x \in N(x^*, \delta_1)$,*

$$\|\nabla f(x)\|^2 = \|J(x)^\top F(x)\|^2 = \|J(x)^\top F(x) - J(\bar{x})^\top \bar{F}\|^2 \leq L_1 \,\text{dist}(x, X^*)^2, \tag{17}$$

$$\|F(x) - \bar{F}\| \leq L_2 \,\text{dist}(x, X^*), \tag{18}$$

and that for all $x, y \in N(x^*, \delta)$,

$$\|F(y) - F(x) - J(x)(y - x)\| \leq L_3 \|y - x\|^2, \tag{19}$$

where L_1, L_2, and L_3 are positive constants.

From the triangle inequality and assuming (18), we get

$$\|F(x)\| - \|\bar{F}\| \leq \|F(x) - \bar{F}\| \leq L_2 \,\text{dist}(x, X^*). \tag{20}$$

We introduce then the following additional assumption,

Assumption 6.3 *There exists a $\delta_3 > 0$ and $M > 0$ such that for $x \in N(x^*, \delta_3)$,*

$$\text{dist}(x, X^*) \leq M \|F(x) - \bar{F}\|.$$

As the function $x \to F(x) - \bar{F}$ is zero residual, the proposed error bound assumption can be seen as a generalization of the zero residual case [21, 8, 5, 6]. We note that locally the nonsingularity of $J(x)$ or a standard second order sufficient optimality conditions would imply this error bound.

6.2 Convergence Proof

From the global convergence results, we have established that there is a subsequence of successful iterations converging to a solution set X^*. In this section, we begin by considering the subsequence of iterations that succeed the successful iterations, i.e., we consider the subsequence $\mathcal{K} = \{j + 1 : j \in \mathbb{S}\}$. We shall present the results with a slight abuse of notation that simplifies the presentation without sacrificing accuracy or generality: in particular every time we denote a quantity a_j, the index j corresponds to an element of this subsequence $\mathcal{K}$ denoted above, thus when we say a particular statement holds eventually, this means that it holds for all $j \in \mathbb{S} + 1$ with j sufficiently large.

We shall denote $\hat{\mu}$ as an upper bound for $\bar{\mu}$. Note that this exists for $\{\mu_j\}_{j\in\mathcal{K}}$ by the formulation of Algorithm 1. In addition we shall denote δ as $\delta = \min(\delta_1, \delta_2, \delta_3)$, with $\{\delta_i\}_{i=1,2,3}$ defined in the Assumptions.

In the proof we follow the structure of the local convergence proof in [21], with the additional point that the step is accepted by the globalization procedure. For all j, we define

$$\bar{x}_j = \text{argmin}_{y\in X^*} \|x_j - y\|,$$

meaning that

$$\|x_j - \bar{x}_j\| = \text{dist}(x_j, X^*).$$

The next lemma is similar to [21, Lemma 2.1].

Lemma 6.1 *Suppose that Assumptions 6.1 and 6.2 are satisfied.*
If $x_j \in N(x^*, \frac{\delta}{2})$, then the solution s_j to (2) satisfies,

$$\|J_j s_j + F_j\| - \|\bar{F}\| \leq C_1 \,\text{dist}(x_j, X^*)^2, \tag{21}$$

where C_1 is a positive constant independent of j.

Proof. We have,

$$\begin{aligned}
\|J_j s_j + F_j\|^2 &\leq 2m_j(s_j) \leq 2m_j(\bar{x}_j - x_j) = \|J_j(\bar{x}_j - x_j) + F_j\|^2 + \gamma_j\|\bar{x}_j - x_j\|^2 \\
&\leq \|J_j(\bar{x}_j - x_j) + F_j\|^2 + \mu_j L_1 \|x_j - \bar{x}_j\|^4 \qquad \text{(using Assumption 6.2)} \\
&\overset{(a)}{\leq} \left(\left(\|\bar{F}\| + L_3\|x_j - \bar{x}_j\|^2\right)^2 + \mu_j L_1 \|x_j - \bar{x}_j\|^4\right) \\
&= \left(\left(L_3^2 + \mu_j L_1\right)\| x_j - \bar{x}_j\|^4 + 2L_3\|\bar{F}\|\|x_j - \bar{x}_j\|^2 + \|\bar{F}\|^2\right) \\
&\leq \left(\left(L_3^2 + \mu_j L_1\right)\| x_j - \bar{x}_j\|^4 + 2L_3\sqrt{\frac{L_3^2 + \mu_j L_1}{L_3^2}}\|\bar{F}\|\|x_j - \bar{x}_j\|^2 + \|\bar{F}\|^2\right) \\
&= \left(\sqrt{L_3^2 + \hat{\mu} L_1}\|x_j - \bar{x}_j\|^2 + \|\bar{F}\|\right)^2
\end{aligned}$$

where (a) is from the triangle inequality as well as Assumption 6.2, i.e.,

$$\begin{aligned}
\|J_j(\bar{x}_j - x_j) + F_j\| &\leq \|\bar{F}\| + \|J_j(\bar{x}_j - x_j) + F_j - \bar{F}\| \\
&\leq \|\bar{F}\| + L_3\|x_j - \bar{x}_j\|^2.
\end{aligned}$$

■

Lemma 6.2 *Suppose that Assumptions 6.1 and 6.2 are satisfied.*
If $x_j \in N(x^*, \frac{\delta}{2})$, then the solution s_j to (2) satisfies,

$$\|s_j\| \leq C_2 \,\text{dist}(x_j, X^*), \tag{22}$$

where C_2 is a positive constant independent of j.

Proof. The solution of the classical Levenberg-Marquardt subproblem, for zero-residual problems proposed in [21] , when solving $F(x) - \bar{F} = 0$, satisfies

$$\begin{aligned}
&\text{argmin}_s \quad \tfrac{1}{2}\|F_j - \bar{F} + J_j s\|^2 + \tfrac{1}{2}\mu_j\|F_j - \bar{F}\|^2\|s\|^2, \\
=\ &\text{argmin}_s \quad \tfrac{1}{2}\left(2F_j^\top J_j s - 2\bar{F}^\top J_j s + s^\top J_j^\top J_j s + \mu_j\|F_j - \bar{F}\|^2 s^\top s\right)
\end{aligned}$$

From the first inequality in [21, Lemma 2.1] it holds that the solution to this problem, $\hat{s}_j$, satisfies $\|\hat{s}_j\| \leq \hat{C}\,\mathrm{dist}(x_j, X^*)$, where $\hat{C}$ is a positive constant independent of j.

Now define the function

$$G(s,u,v) = \frac{1}{2}\left(2F_j^\top J_j s + u^\top s + s^\top J_j^\top J_j s + v s^\top s\right). \tag{23}$$

We may consider the model in (2) as a perturbation of (23) if we set in the expression of G, u and v to $u_0 = 0$ and $v_0 = \gamma_j$, respectively. The classical Levenberg-Marquardt model for zero residual problems is the same as (23) with $u_1 = -2J_j^\top \bar{F}$ and $v_1 = \mu_j \|F_j - \bar{F}\|^2$.

In other words, if we change in the expression of G the values of $u = u_1$ and $v = v_1$ by $u = u_0$ and $v = v_0$, respectively, we modify the model from the classical Levenberg-Marquardt subproblem for zero residual problems to the one defined in (2).

Since the function G is quadratic with respect to s and linear with respect to (u, v), the conditions of [2, Proposition 4.36] are satisfied, and thus,

$$\begin{aligned}
\|s_j - \hat{s}_j\| &\leq \hat{D}_1\|u_1 - u_0\| + \hat{D}_2\|v_1 - v_0\| \\
&\leq 2\hat{D}_1\|\bar{F}^\top J_j\| + \hat{D}_2\mu_j|\|F_j - \bar{F}\|^2 - \|F_j^\top J_j\|^2| \\
&\leq 2\hat{D}_1\left(\|F_j J_j\| + \|F_j - \bar{F}\|\|J_j\|\right) + \hat{D}_2\max(L_1, L_2^2)\hat{\mu}\,\mathrm{dist}(x_j, X^*)^2 \\
&\leq 4\hat{D}_1\max(\sqrt{L_1}, B_J L_2)\,\mathrm{dist}(x_j, X^*) + \hat{D}_2\max(L_1, L_2^2)\hat{\mu}\,\mathrm{dist}(x_j, X^*)^2 \\
&\leq \hat{D}\,\mathrm{dist}(x_j, X^*),
\end{aligned}$$

where $\hat{D}_1$, $\hat{D}_2$ and $\hat{D}$ are positive constants independent of j. The second inequality follows from Assumption 6.2, the triangle and Cauchy-Schwartz inequalities, and the final inequality again follows from Assumption 6.2 and the boundedness of J_j by B_J (due to the convergence of x_j).

Thus, from the triangle inequality,

$$\|s_j\| \leq \|\hat{s}_j\| + \|s_j - \hat{s}_j\| \leq \left(\hat{C} + \hat{D}\right)\mathrm{dist}(x_j, X^*).$$

■

Lemma 6.3 *Suppose that Assumptions 6.1 and 6.2 are satisfied. Then for j sufficiently large, one has $\rho_j \geq \eta$.*

Proof. Consider first the case of $\|\bar{F}\| > 0$.

It holds that,

$$\begin{aligned}
m_j(0) - m_j(s_j) &= \|F_j\|^2 - \|F_j + J_j s_j\|^2 - \gamma_j\|s_j\|^2 \\
&\overset{(a)}{\geq} \|F_j\|^2 - \|F_j + J_j(\bar{x}_j - x_j)\|^2 - \gamma_j\|\bar{x}_j - x_j\|^2 \\
&= \left(\|F_j\| + \|F_j + J_j(\bar{x}_j - x_j)\|\right)\left(\|F_j\| - \|F_j + J_j(\bar{x}_j - x_j)\|\right) - \gamma_j\|\bar{x}_j - x_j\|^2 \\
&\geq \left(\|F_j\| + \|F_j + J_j(\bar{x}_j - x_j)\|\right)\left(\|F_j\| - \|\bar{F}\| - L_3\|\bar{x}_j - x_j\|^2\right) - \gamma_j\|\bar{x}_j - x_j\|^2 \\
&\overset{(b)}{\geq} \|F_j\|\left(L_2\|\bar{x}_j - x_j\| - L_3\|\bar{x}_j - x_j\|^2\right) - \gamma_j\|\bar{x}_j - x_j\|^2 \\
&\geq \|\bar{F}\|L_2\|\bar{x}_j - x_j\| - \left(\|F_0\|L_3 + \gamma_j\right)\|\bar{x}_j - x_j\|^2 \\
&= \|\bar{F}\|L_2\|\bar{x}_j - x_j\| - \left(\tilde{L}_3 + \gamma_j\right)\|\bar{x}_j - x_j\|^2,
\end{aligned}$$

where (a) arises from the optimality of s_j for m_j, and for (b) we note that for j sufficiently large $\|\bar{x}_j - x_j\|^2 \ll \|\bar{x}_j - x_j\|$ and thus $L_2\|\bar{x}_j - x_j\| - L_3\|\bar{x}_j - x_j\|^2 \geq 0$.

Now we write,

$$\begin{aligned}
|1-\rho_j| &= \left|\frac{m_j(0)-f(x_j)+f(x_j+s_j)-m_j(s_j)}{m_j(0)-m_j(s_j)}\right| \\
&= \left|\frac{\|F(x_j+s_j)\|^2-\|F_j+J_js_j\|^2-\gamma_j\|s_j\|^2}{m_j(0)-m_j(s_j)}\right| \\
&= \left|\frac{(\|F(x_j+s_j)\|-\|F_j+J_js_j\|)\,(\|F(x_j+s_j)\|+\|F_j+J_js_j\|)-\gamma_j\|s_j\|^2}{m_j(0)-m_j(s_j)}\right| \\
&\leq \frac{L_3\|s_j\|^2\,(\|F_j\|+\|J_j\|\|s_j\|)+\gamma_j\|s_j\|^2}{\|\bar{F}\|L_2\|\bar{x}_j-x_j\|-\left(\tilde{L}_3+\gamma_j\right)\|\bar{x}_j-x_j\|^2} \\
&\to 0 \quad \text{when } j \text{ goes to } +\infty.
\end{aligned}$$

The last limit is zero because $\|s_j\| = O(\|x_j - \bar{x}_j\|) \to 0$ by Lemma 6.2, γ_j is bounded, and $\|\bar{F}\| > 0$.

Now, if $\|\bar{F}\| = 0$, we have, from the same derivation,

$$m_j(0) - m_j(s_j) \geq \|F_j\|L_2\|\bar{x}_j - x_j\| - \left(\tilde{L}_3+\gamma_j\right)\|\bar{x}_j - x_j\|^2,$$

and, using Assumptions 6.3 and 6.2 as well as Lemma 6.2,

$$\begin{aligned}
|1-\rho_j| &\leq \frac{L_3\|s_j\|^2\,(\|F_j\|+\|J_j\|\|s_j\|)+\gamma_j\|s_j\|^2}{\|F_j\|L_2\|\bar{x}_j-x_j\|-\left(\tilde{L}_3+\gamma_j\right)\|\bar{x}_j-x_j\|^2} \\
&\leq \frac{C_5\|x_j-\bar{x}_j\|^2\,(\|x_j-\bar{x}_j\|+\|J_j\|\|x_j-\bar{x}_j\|)+\gamma_j\|x_j-\bar{x}_j\|^2}{\tilde{L}_2\|x_j-\bar{x}_j\|\|\bar{x}_j-x_j\|-\left(\tilde{L}_3+\gamma_j\right)\|\bar{x}_j-x_j\|^2},
\end{aligned}$$

thus the power for $\|x_j - \bar{x}_j\|$ is larger in the numerator and the fraction converges to zero. ■

Proposition 6.1 *Suppose that Assumptions 6.1, 6.2, and 6.3 are satisfied. Let x_j, $x_{j+1} \in N(x^*, \delta/2)$. One has,*

$$\left(1-\sqrt{m}L_1M^2\|\bar{F}\|\right)\operatorname{dist}(x_{j+1},X^*)^2 \leq C_3^2\operatorname{dist}(x_j,X^*)^4 + \hat{C}_3^2\|\bar{F}\|\operatorname{dist}(x_j,X^*)^2 \tag{24}$$

where the constants C_3 and $\hat{C}_3$ are given by

$$C_3 = M\sqrt{C_1^2+2L_3C_1C_2^2+L_3^2} \quad \textit{and} \quad \hat{C}_3 = M\sqrt{2C_1+2L_3C_2^2}.$$

Proof. Indeed, using Assumption 6.3 and Lemma 6.1, one has

$$\begin{aligned}
\|x_{j+1}-\bar{x}_{j+1}\|^2 &\leq M^2\|F(x_j+s_j)-\bar{F}\|^2 \\
&\leq M^2\left(\|F(x_j+s_j)\|^2-2F(x_j+s_j)^\top\bar{F}+\|\bar{F}\|^2\right) \\
&\leq M^2\left(\left(\|J(x_j)s_j+F_j\|+L_3\|s_j\|^2\right)^2-2F(x_j+s_j)^\top\bar{F}+\|\bar{F}\|^2\right) \\
&\leq M^2\Big(\|J(x_j)s_j+F_j\|^2+2L_3\|J(x_j)s_j+F_j\|\|s_j\|^2+L_3^2\|s_j\|^4 \\
&\qquad -2F(x_j+s_j)^\top\bar{F}+\|\bar{F}\|^2\Big) \\
&\leq M^2\Big(C_1^2\|x_j-\bar{x}_j\|^4+2C_1\|x_j-\bar{x}_j\|^2\|\bar{F}\|+\|\bar{F}\|^2+2L_3C_1\|x_j-\bar{x}_j\|^2\|s_j\|^2 \\
&\qquad +2L_3\|\bar{F}\|\|s_j\|^2+L_3^2\|s_j\|^4-2F(x_j+s_j)^\top\bar{F}+\|\bar{F}\|^2\Big).
\end{aligned}$$

Therefore, using Lemma 6.2, one gets

$$\|x_{j+1} - \bar{x}_{j+1}\|^2 \leq C_3^2 \|x_j - \bar{x}_j\|^4 + \hat{C}_3^2 \|\bar{F}\| \|x_j - \bar{x}_j\|^2 + 2M^2 |F(x_j + s_j)^\top \bar{F} - \|\bar{F}\|^2|, \tag{25}$$

where $C_3 = M\sqrt{C_1^2 + 2L_3 C_1 C_2^2 + L_3^2}$ and $\hat{C}_3 = M\sqrt{2C_1 + 2L_3 C_2^2}$ are positive constants.

Moreover, by applying a Taylor expansion to $x \to F(x)^\top \bar{F}$ at the point $x_{j+1} = x_j + s_j$ around $\bar{x}_{j+1}$, there exists $R > 0$ such that

$$|F(x_j + s_j)^\top \bar{F} - \|\bar{F}\|^2| = \|(J(\bar{x}_{j+1})^\top \bar{F})^\top (x_{j+1} - \bar{x}_{j+1})\| + R\|x_{j+1} - \bar{x}_{j+1}\|^2 = R\|x_{j+1} - \bar{x}_{j+1}\|^2.$$

Note that the Hessian of $x \to F(x)^\top \bar{F}$ is equal to $\sum_{i=1}^m F_i(\bar{x}) \nabla^2 F_i(x)$, and from Assumption 6.2 we have $\nabla^2 F_i(x)$ are bounded. Hence, the constant R is bounded as follows

$$R \leq L_1 \sum_{i=1}^m |\bar{F}_i| \leq \sqrt{m} L_1 \|\bar{F}\|.$$

Combining the obtained Taylor expansion and (25) gives

$$\|x_{j+1} - \bar{x}_{j+1}\|^2 \leq C_3^2 \|x_j - \bar{x}_j\|^4 + \hat{C}_3^2 \|\bar{F}\| \|x_j - \bar{x}_j\|^2 + \sqrt{m} L_1 M^2 \|\bar{F}\| \|x_{j+1} - \bar{x}_{j+1}\|^2.$$

Which completes this proof. ■

In the next lemma, we show that, as long as the iterates $\{x_j\}_j$ lie sufficiently near to x^*, the sequence $\{\text{dist}(x_j, X^*)\}_j$ converges to 0 quadratically if the problem has a zero residual, or linearly when the residual is small.

Lemma 6.4 *Suppose that Assumptions 6.1, 6.2, and 6.3 are satisfied. Let x_j, $x_{j+1} \in N(x^*, \delta/2)$.*

If the problem has a zero residual, i.e., $\|\bar{F}\| = 0$, then

$$\text{dist}(x_{j+1}, X^*) \leq C_3 \, \text{dist}(x_j, X^*)^2, \tag{26}$$

where C_3 is a positive constant independent of j.

If the problem has a small non-zero residual, i.e., $\|\bar{F}\| < \min\left\{\frac{1}{\sqrt{m}L_1 M^2}, \frac{1 - C_3^2 \delta}{\hat{C}_3^2 + \sqrt{m} L_1 M^2}\right\}$, then

$$\text{dist}(x_{j+1}, X^*) \leq C_4 \, \text{dist}(x_j, X^*), \tag{27}$$

where $C_4 \in (0, 1)$ is a positive constant independent of j.

Proof. Indeed, Under the zero residual case, i.e., $\bar{F} = 0$, then Proposition 6.1 is being equivalent to

$$\text{dist}(x_{j+1}, X^*) \leq C_3 \, \text{dist}(x_j, X^*)^2.$$

If the problem has a small non-zero residual such as $\|\bar{F}\| < \min\left\{\frac{1}{\sqrt{m}L_1 M^2}, \frac{1 - C_3^2 \delta}{\hat{C}_3^2 + \sqrt{m} L_1 M^2}\right\}$, then Proposition 6.1 will be equivalent to

$$\text{dist}(x_{j+1}, X^*)^2 \leq \frac{C_3^2 \delta^2 + \hat{C}_3^2 \|\bar{F}\|}{1 - \sqrt{m} L_1 M^2 \|\bar{F}\|} \, \text{dist}(x_j, X^*)^2 = C_4^2 \, \text{dist}(x_j, X^*)^2,$$

where $C_4 = \sqrt{\frac{C_3^2 \delta^2 + \hat{C}_3^2 \|\bar{F}\|}{1 - \sqrt{m} L_1 M^2 \|\bar{F}\|}}$. Since $\|\bar{F}\| < \frac{1 - C_3^2 \delta}{\hat{C}_3^2 + \sqrt{m} L_1 M^2}$, one has $C_4 \in (0, 1)$. Which completes the proof. ■

Theorem 6.1 *Suppose that Assumptions 6.1, 6.2, and 6.3 are satisfied.*

If $\|\bar{F}\| = 0$ then Algorithm 1 converges locally quadratically to X^*. Otherwise, if the problem has a small non-zero residual as in Lemma 6.4, Algorithm 1 converges locally linearly to X^*.

Proof. From the previous results, it can be seen that eventually for $j \in \mathcal{K} = \mathbb{S} + 1$, it holds that there is a step s_j such that $x_j + s_j$ is quadratically (or linearly, depending on the value of $\|\bar{F}\|$) closer to the solution, and is accepted. In particular, by the same argument as given in [21, Lemma 2.3] $x_j + s_j$ is always at least as close to x^* as x_j, and thus $x_{j+1} = x_j + s_j$ lies in a ball around x^* for which all of the local assumptions hold as well.

But then this implies that $j + 1 \in \mathcal{K}$ as well, and all of the previous results apply to it. Thus, proceeding inductively we get that for sufficiently large $j \in \mathcal{K}$, it holds that all subsequent iterations are in $\mathcal{K}$ and the entire sequence of iterates $\{x_j\}$ (no longer subsequence) locally converges to X^*, quadratically if $\|\bar{F}\| = 0$ and linearly for small non-zero residual. ■

Using [9, Lemma 2.9], one can deduce the previous lemma results hold with respect to $\{\|x_j - \hat{x}\|\}_j$ for some limit point $\hat{x} \in X^*$.

Corollary 6.1 *Suppose that Assumptions 6.1, 6.2, and 6.3 are satisfied. Let x_j be a sequence generated by the proposed Algorithm. Then, there are $\delta_1 > 0$ and $C_6 > 0$ such that if $x_0 \in N(x^*, \delta_1)$ implies that $(x_j)_j$ converges to some $\hat{x} \in X^*$ as follows:*

If the problem has a zero residual, then

$$\|x_{j+1} - \hat{x}\| \ \leq C_6 \|x_j - \hat{x}\|^2.$$

Otherwise, if the problem has a small non-zero residual as in Lemma 6.4, then

$$\|x_{j+1} - \hat{x}\| \ \leq C_6 \|x_j - \hat{x}\|.$$

Proof. Let $0 < \delta_1 \leq \min\left\{\frac{\delta}{2C_3}, C_4\right\}$, where δ, C_3 and C_4 are the same as in Lemma 6.4.

Let $w_j = x_j$, $r_j = \operatorname{dist}(x_j, X^*)$ and $R = C_2$.

If $\|\bar{F}\| = 0$ let $\tau = 2$, $r = \delta_1 C_3 < 1$ and $c = C_3$. Otherwise if the residual is small as in Lemma 6.4, we set $\tau = 1$, $r = C_4 < 1$ and $c = C_4$.

We have from Lemma 6.2

$$\|s_j\| \ = \ \|w_{j+1} - w_j\| \ \leq \ C_2 \operatorname{dist}(x_j, X^*) = R r_j,$$

moreover from Theorem 6.1 we have

$$r_{j+1} = \operatorname{dist}(x_{j+1}, X^*) \leq r \operatorname{dist}(x_j, X^*) = r r_j, \text{ and } r_{j+1} \leq c r_j^{\tau}.$$

Therefore by using [9, Lemma 2.9] (with the same notation except for the index j which is replaced by index k), we conclude that $(w_j)_j$ converges to some limit point $\hat{w} := \hat{x}$ and

$$\|w_{j+1} - \hat{w}\| = \|x_{j+1} - \hat{x}\| \ \leq \ \frac{cR}{1-r}\|w_j - \hat{w}\|^{\tau} = \frac{cR}{1-r}\|x_j - \hat{x}\|^{\tau}.$$

■

7 Numerical results

In this section we present some results of numerical experiments performed implementing Algorithm 1. We performed our experiments using the well known 33 Moré/Garbow/Hillstrom test problems [13]. All the tested problems are smooth and have a least-squares structure. The residual function F and the Jacobian matrix for all the test problems [13] are implemented in MATLAB. Some of these problems have a non-zero value at the optimum and thus are consistent with the scope of the paper. To have a large test set, we create additional 14 optimization problems by varying the problem dimension n when it is possible. In our results the names of the latter problems will be followed by " * " and we will refer to the final test problems by $\mathcal{P}$.

Among all the test problems in $\mathcal{P}$ three problems can be seen as ill-conditioned (meaning that the condition number of their Jacobian at the optimum is larger than 10^6). four of them are even nearly singular (i.e., the condition number of their Jacobian at the optimum is larger than 10^{20}). For all the tested problems, we used the proposed starting points x_0 as in the original optimization test problems [13].

A preliminary implementation of Algorithm 1 was written in MATLAB. The initial parameters defining the implemented algorithm were set as follows:

$$\eta = 0.01,\ \lambda = 5,\ \mu_0 = 1,\ \mu_{\min} = 10^{-16}.$$

At each iteration of Algorithm 1 we solved exactly the subproblem using the backslash MATLAB operator. If the iteration successful, we set the parameter μ_{j+1} equals to $\max(\bar{\mu}/\lambda, \mu_{\min})$. The algorithm is stopped when

$$\|\nabla f(x_j)\| \leq \epsilon \qquad \text{with} \qquad \epsilon = 10^{-5}.$$

If it did not converge within a maximum number of iterations $j_{\max} = 10000$, then Algorithm 1 was considered to have failed.

The order of convergence was estimated by the quantity

$$\text{EOC} = \log\left(\frac{\|\nabla f(x_{j_f})\|}{\max(1, \|\nabla f(x_0)\|)}\right) / \log\left(\frac{\|\nabla f(x_{j_f-1})\|}{\max(1, \|\nabla f(x_0)\|)}\right), \tag{28}$$

where j_f denotes the index of the final computed iterate. The obtained values of EOC will help to estimate the speed of convergence. For the problems with an EOC ≥ 1.8 Algorithm 1 will be said to be quadratically converging. For test problems with $1.8 >$ EOC ≥ 1.1 the algorithm can be seen as super-linearly converging. Otherwise (i.e., EOC < 1.1), Algorithm 1 will be estimated to converge linearly or worse.

Table 1: Order of convergence for problems in $\mathcal{P}$, zero residual case.

	Order of convergence			Total
	Linear or worse	Super-linear	Quadratic	
Well-conditioned	2	7	16	25
Poorly conditioned	0	1	2	3
Total	2	8	18	28

Tables 1 and 2 gives the distribution of the order of convergence over our test set $\mathcal{P}$, separated by whether the solution has a zero or non-zero residual, with the threshold 10^{-5}. The detailed results are presented in Tables 3 and 4. Within a maximum number of iterations $j_{\max} = 10000$,

Table 2: Order of convergence for problems in $\mathcal{P}$, non-zero residual case.

	Order of convergence			Total
	Linear or worse	Super-linear	Quadratic	
Well-conditioned	4	5	1	10
Poorly conditioned	3	2	4	9
Total	7	7	5	19

Table 3: The complete results for Algorithm 1 for zero residual test problems.

Problem	n	$f(x_{j_f})$	$\|\nabla f(x_{j_f-1})\|$	$\|\nabla f(x_{j_f})\|$	Cond(J_{j_f})	γ_{j_f}	EOC
`rosen`	2	8.096e-21	7.569e+00	6.355e-11	5.008e+01	2.647e-32	10.3
`badscp`	2	6.311e-30	1.305e-03	3.235e-10	8.295e+08	1.047e-35	1.96
`badscb`	2	0.000e+00	4.274e-05	0.000e+00	1.000e+06	0.000e+00	Inf
`beale`	2	7.026e-25	1.792e-05	2.485e-12	1.275e+01	1.581e-29	2.16
`helix`	3	1.526e-21	7.884e-04	8.852e-10	2.222e+01	3.210e-27	1.98
`gauss`	3	5.640e-09	3.726e-03	3.260e-08	7.162e+00	2.126e-16	3.08
`gulf`	3	2.904e-09	4.356e-05	3.557e-06	1.047e+05	3.317e-24	1.21
`box`	3	2.440e-13	9.171e-04	1.012e-06	9.133e+01	8.391e-22	1.6
`sing`	4	8.157e-09	5.303e-05	6.628e-06	6.102e+02	1.799e-19	1.14
`wood`	4	6.630e-19	5.684e-03	2.535e-08	3.739e+01	6.425e-32	1.87
`biggs`	6	7.708e-15	6.281e-04	2.843e-07	1.134e+03	8.084e-30	2.01
`watson`	9	6.999e-07	6.049e-04	1.110e-06	4.132e+04	3.155e-18	1.53
`watson*`	20	2.913e-15	4.611e-04	2.139e-06	2.776e+13	1.171e-17	1.42
`rosex`	10	2.496e-30	3.066e+00	2.234e-15	5.008e+01	1.047e-44	8.85
`rosex*`	20	1.440e-25	1.046e+01	2.498e-13	5.008e+01	6.542e-40	9.81
`singx`	4	8.157e-09	5.303e-05	6.628e-06	6.102e+02	1.799e-19	1.14
`singx*`	20	2.539e-09	1.477e-05	1.846e-06	1.220e+03	5.582e-22	1.12
`pen2`	4	4.711e-06	1.234e-05	8.531e-06	8.527e+03	7.278e-27	1.03
`vardim`	10	3.857e-29	3.241e-04	1.725e-13	1.965e+01	1.561e-39	1.94
`vardim*`	20	1.741e-24	5.237e-03	9.998e-11	5.358e+01	9.996e-37	1.72
`trig*`	20	2.329e-06	1.574e-05	9.646e-06	2.048e+03	4.764e-17	1.04
`bv`	10	1.260e-14	6.502e-05	2.793e-08	3.792e+01	3.121e-17	1.8
`bv*`	20	2.324e-08	5.596e-03	7.427e-06	1.392e+02	1.103e-11	2.28
`ie`	10	2.631e-11	1.595e-02	9.249e-06	1.289e+00	3.422e-12	2.8
`ie*`	20	5.499e-22	4.084e-05	4.262e-11	1.292e+00	1.453e-23	2.36
`trid`	10	5.119e-14	5.157e-03	9.827e-07	3.098e+00	6.180e-17	2.01
`trid*`	20	3.145e-12	1.667e-02	7.554e-06	3.155e+00	3.652e-15	2.04
`lin*`	20	3.361e-22	4.327e-03	2.593e-11	1.000e+00	4.303e-26	3.48

all the tested problems converged except two `meyer` and `bd` where our algorithm stagnate at the point x_{j_f} with gradient norm equals to 1.248×10^{-4} and 1.393×10^{-4} respectively.

For zero-residual problems (see Table 1), the obtained results confirm our theoretical analysis in two senses. First, for all the tested problems, we observe the global convergence of our proposed algorithm. Second, the obtained results show quadratic or super-linear convergence rates on most of the problems tested. In fact, among the 28 zero-residual tested problems, for 18 problems we observe quadratic convergence, 8 problems are super-linearly converging, and

Table 4: The complete results for Algorithm 1 for non-zero residual test problems.

Problem	n	$f(x_{j_f})$	$\|\nabla f(x_{j_f-1})\|$	$\|\nabla f(x_{j_f})\|$	Cond(J_{j_f})	γ_{j_f}	EOC
froth	2	2.449e+01	2.300e-05	7.641e-06	3.554e+08	5.838e-27	1.06
jensam	2	6.218e+01	4.652e-05	6.465e-06	4.166e+08	4.179e-27	1.1
bard	3	4.107e-03	7.274e-04	1.601e-06	6.687e+01	6.559e-18	1.56
meyer	3	4.397e+01	8.258e+01	1.248e-04	9.970e+07	5.035e+37	-
kowosb	4	1.538e-04	2.248e-05	7.625e-06	6.927e+01	2.977e-17	1.1
bd	4	4.291e+04	2.149e-04	1.393e-04	1.084e+02	6.269e+37	-
osb1	5	2.732e-05	5.139e-05	2.803e-08	4.884e+04	7.858e-32	1.49
osb2	11	2.007e-02	1.921e-05	3.291e-06	9.020e+01	2.219e-19	1.15
pen1	4	1.125e-05	2.398e-05	8.658e-06	3.162e+02	7.496e-27	1.06
pen1*	20	7.889e-05	3.013e-05	5.564e-06	3.163e+02	3.096e-27	1.07
pen2*	10	1.468e-04	6.296e-05	7.166e-06	1.458e+04	8.414e-21	1.14
trig	10	1.398e-05	1.609e-05	6.412e-06	1.105e+03	3.368e-20	1.08
band	10	1.340e+00	4.767e-05	9.948e-06	5.621e+06	2.594e-23	1.1
band*	20	1.340e+00	1.200e-05	8.143e-06	4.678e+07	2.173e-21	1.02
lin	10	5.000e+00	1.178e-05	3.331e-16	1.000e+00	7.100e-36	2.84
lin1	10	2.317e+00	2.528e-03	5.368e-11	9.172e+63	4.720e-31	1.84
lin1*	20	2.317e+00	3.574e+03	9.083e-07	8.640e+49	2.703e-23	3.43
lin0	10	3.068e+00	7.850e+00	4.949e-11	Inf	4.012e-31	3
lin0*	20	3.068e+00	3.349e+00	1.449e-09	Inf	1.376e-29	2.32

only 2 converge linearly or worse.

For non-zero residual problem (see Table 2), we note that the percentage of the problems quadratically converging is no fewer. In fact, quadratic convergence was observed only for linear residual test cases, mainly `lin`, `lin0`, `lin0*`, `lin1` and `lin1*` test problems. Super-linear convergence was observed for 7 tests cases while the method shows a linear convergence rate for 5 problems.

8 Conclusion

In this paper, we presented and analyzed a novel Levenberg-Marquardt method for solving nonlinear least-squares problems. In particular, we were interested in the class of problems for which there is no zero-residual solution (or such a solution is not sought after) and the problem is potentially ill-conditioned. Without the use of ancillary procedures to enforce globalization, we were able to formulate a globally convergent Levenberg-Marquardt method and explicite its worst-case iteration complexity bounds. The proposed method is convergent locally at quadratic rate for zero residual problems and a linear rate for small residuals. Preliminary numerical results confirmed the theoretical behavior. Future research can include problems with constraints as well as those with noisy data.

Acknowledgement

We would like to thank Clément Royer for assistance with respect to the complexity results in the paper, as well as the anonymous referees for their recommendations for the paper.

References

[1] E. Bergou, S. Gratton, and L. N. Vicente. Levenberg-Marquardt Methods Based on Probabilistic Gradient Models and Inexact Subproblem Solution, with Application to Data Assimilation. *SIAM/ASA J. Uncertainty Quantification*, 4:924–951, 2016.

[2] J. F. Bonnans and A. Shapiro. *Perturbation Analysis of Optimization Problems*. Springer Science & Business Media, 2013.

[3] A. R. Conn, N. I. M. Gould, and Ph. L. Toint. *Trust-Region Methods*. SIAM, Philadelphia, PA, USA, 2000.

[4] P. Courtier, J.-N Thépaut, and A. Hollingsworth. A Strategy for Operational Implementation of 4D-Var, Using an Incremental Approach. *Q. J. R. Meteorol. Soc.*, 120:1367–1388, 1994.

[5] H. Dan, N. Yamashita, and M. Fukushima. Convergence Properties of the Inexact Levenberg-Marquardt Method under Local Error Bound Conditions. *Optim. Methods Softw.*, 17:605–626, 2002.

[6] F. Facchinei, A. Fischer, and M. Herrich. A Family of Newton Methods for Nonsmooth Constrained Systems with Nonisolated Solutions. *Math. Methods Operations Research*, 77:433–443, 2013.

[7] J. Fan. Convergence Rate of the Trust Region Method for Nonlinear Equations under Local Error Bound Condition. *Comput. Optim. Appl.*, 34:215–227, 2006.

[8] J. Fan and Y. Yuan. On the Quadratic Convergence of the Levenberg-Marquardt Method without Nonsingularity Assumption. *Computing*, 74:23–39, 2005.

[9] A. Fischer, P.K. Shukla, and M. Wang. On the inexactness level of robust Levenberg-Marquardt methods. *Optimization*, 59:273–287, 2010.

[10] I. C. F. Ipsen, C. T. Kelley, and S. R. Pope. Rank-deficient Nonlinear Least Squares Problems and Subset Selection. *SIAM J. Numer. Anal.*, 49:1244–1266, 2011.

[11] K. Levenberg. A Method for the Solution of Certain Problems in Least Squares. *Quart. Appl. Math.*, 2:164–168, 1944.

[12] D. Marquardt. An Algorithm for Least-Squares Estimation of Nonlinear Parameters. *SIAM J. Appl. Math.*, 11:431–441, 1963.

[13] J. J. Moré, B. S. Garbow, and K. E. Hillstrom. Testing Unconstrained Optimization Software. *ACM Trans. Math. Softw.*, 7:17–41, 1981.

[14] J. Nocedal and S. J. Wright. *Numerical Optimization*. Springer-Verlag, Berlin, second edition, 2006.

[15] M. R. Osborne. Nonlinear Least Squares – The Levenberg Algorithm Revisited. *J. Austral. Math. Soc. Ser. B*, 19:343–357, 1976.

[16] A. Tarantola. *Inverse Problem Theory and Methods for Model Parameter Estimation*. SIAM, Philadelphia, 2005.

[17] Y. Trémolet. Model Error Estimation in 4D-Var. *Q. J. R. Meteorol. Soc.*, 133:1267–1280, 2007.

[18] K. Ueda and N. Yamashita. On a global complexity bound of the Levenberg-Marquardt method. *J. Optim. Theory Appl.*, 147:443–453, 2010.

[19] K. Ueda and N. Yamashita. Global Complexity Bound Analysis of the Levenberg-Marquardt Method for Nonsmooth Equations and Its Application to the Nonlinear Complementarity Problem. *J. Optim. Theory Appl.*, 152:450–467, 2012.

[20] J. Virieux and S. Operto. An Overview of Full-Waveform Inversion in Exploration Geophysics. *Geophysics*, 74:WCC1–WCC26, 2009.

[21] N. Yamashita and M. Fukushima. On the Rate of Convergence of the Levenberg-Marquardt Method. In *Topics in numerical analysis*, pages 239–249. Springer, 2001.

[22] R. Zhao and J. Fan. Global complexity bound of the Levenberg-Marquardt method. *Optim. Methods Softw.*, 31:805–814, 2016.