Levenberg-Marquardt methods for the solution of noisy nonlinear least squares problems

PhD Candidate : Elisa Riccietti

Università degli Studi di Firenze
Dipartimento di Matematica e Informatica 'Ulisse Dini'
Institut National Polytechnique (INP), Toulouse

Supervisor: Stefania Bellavia, French supervisor: Serge Gratton

PhD defence, 26/02/2018.

Noisy least-squares problems

Nonlinear least-squares problems

Given $R : \mathbb{R}^n \to \mathbb{R}^m$, $m \geq n$, nonlinear, continuously differentiable solve

$$\min_{x \in \mathbb{R}^n} \Phi(x) = \frac{1}{2}\|R(x)\|^2.$$

Let x^* be a solution of the problem.

Noisy least-squares problems

Nonlinear least-squares problems

Given $R : \mathbb{R}^n \to \mathbb{R}^m$, $m \geq n$, nonlinear, continuously differentiable solve

$$\min_{x \in \mathbb{R}^n} \Phi(x) = \frac{1}{2}\|R(x)\|^2.$$

Let x^* be a solution of the problem.

Noisy least-squares problems

We assume that Φ and its derivatives are not available. We look for an approximation to x^* considering a sequence of approximations to the objective function:

$$\Phi_{\delta_k} \sim \Phi$$

Noisy least-squares problems

Nonlinear least-squares problems

Given $R : \mathbb{R}^n \to \mathbb{R}^m$, $m \geq n$, nonlinear, continuously differentiable solve

$$\min_{x \in \mathbb{R}^n} \Phi(x) = \frac{1}{2}\|R(x)\|^2 \to \textit{unperturbed problem}$$

Let x^* be a solution of the problem.

Noisy least-squares problems

We assume that Φ and its derivatives are not available. We look for an approximation to x^* considering a sequence of approximations to the objective function:

$$\Phi_{\delta_k} \sim \Phi$$

We are interested into two classes of such problems:

- **Ill-posed problems.** Data fitting problems with noisy data such that the solution **does not depend continuously on the data**.
  The noise is fixed and arises from measurements errors: $\Phi_{\delta_k} \equiv \Phi_\delta$ for each k.
  AIM: design stable methods for their solution.

We are interested into two classes of such problems:

- **Ill-posed problems.** Data fitting problems with noisy data such that the solution **does not depend continuously on the data**.
  The noise is fixed and arises from measurements errors: $\Phi_{\delta_k} \equiv \Phi_\delta$ for each k.
  AIM: design stable methods for their solution.

- **Large scale noisy problems.** **Objective function is expensive to compute**, we want to use cheaper approximations.
  The approximation can be improved reducing the noise level.
  AIM: design fast methods for the solution of the unperturbed problem considering a sequence of function approximations of increasing accuracy.

We are interested into two classes of such problems:

- **Ill-posed problems.** Data fitting problems with noisy data such that the solution **does not depend continuously on the data**.
  The noise is fixed and arises from measurements errors: $\Phi_{\delta_k} \equiv \Phi_\delta$ for each k.
  AIM: design stable methods for their solution.

- **Large scale noisy problems.** **Objective function is expensive to compute**, we want to use cheaper approximations.
  The approximation can be improved reducing the noise level.
  AIM: design fast methods for the solution of the unperturbed problem considering a sequence of function approximations of increasing accuracy. $\rightarrow$ *study performed in collaboration with Prof. Serge Gratton in Toulouse.*

Outline

- Background material: introduction to Levenberg-Marquardt and trust-region methods.
- I part: Ill-posed problems
  - regularizing method for zero residual problems,
  - regularizing method for non-zero residual problems.
- II part: Large scale problems with expensive objective function.
- Conclusions and perspectives.
- Research outputs.

Levenberg-Marquardt method

It is an iterative method for solving a least-squares problem. It builds the sequence of solution approximations as $x_{k+1} = x_k + p_k$ where p_k is the solution of:

$$\min_{p\in\mathbb{R}^n} m_k^{LM}(p) = \frac{1}{2}\|R(x_k) + J(x_k)p\|^2 + \frac{1}{2}\lambda_k\|p\|^2$$

where J is the Jacobian matrix of R and $\lambda_k \geq 0$ is a regularization parameter.

Remark

p_k is the solution of

$$(B_k + \lambda_k I)p_k = -g_k$$

with $B_k = J(x_k)^T J(x_k)$, $g_k = J(x_k)^T R(x_k)$.

Classical Levenberg-Marquardt method

- Given $x_k \in \mathbb{R}^n$ and $\lambda_k \geq 0$, find the step $p_k \in \mathbb{R}^n$ minimizing

$$m_k^{LM}(p) = \frac{1}{2}\|R(x_k) + J(x_k)p\|^2 + \frac{1}{2}\lambda_k\|p\|^2.$$

- Set $\Phi(x) = \frac{1}{2}\|R(x)\|^2$, and compute

$$\rho_k(p_k) = \frac{\Phi(x_k) - \Phi(x_k + p_k)}{m_k^{LM}(0) - m_k^{LM}(p_k)}.$$

- Step acceptance. Given $\eta \in (0, 1)$:
  - If $\rho_k < \eta$ reject the step: $x_{k+1} = x_k$ and increase λ_k.
  - If $\rho_k \geq \eta$ accept the step: $x_{k+1} = x_k + p_k$.

Trust-region methods

- Given x_k and the trust-region radius $\Delta_k > 0$ find the step p_k solving

$$\min_p m_k^{TR}(p) = \frac{1}{2}\|R(x_k) + J(x_k)p\|^2,$$
$$\text{s.t. } \|p\| \leq \Delta_k$$

- Set $\Phi(x) = \frac{1}{2}\|R(x)\|^2$. Compute

$$\rho_k(p_k) = \frac{\Phi(x_k) - \Phi(x_k + p_k)}{m_k^{TR}(0) - m_k^{TR}(p_k)}.$$

- Step acceptance and trust-region radius update. Given $\eta \in (0, 1)$:
    - If $\rho_k < \eta$ then set $\Delta_{k+1} < \Delta_k$ and $x_{k+1} = x_k$.
    - If $\rho_k \geq \eta$ then set $\Delta_{k+1} \geq \Delta_k$ and $x_{k+1} = x_k + p_k$.

Trust-region methods

Trust-region methods falls into the class of Levenberg-Marquardt methods.

Levenberg-Marquardt - Trust region

- LM: $\min_p m_k^{LM}(p) = \frac{1}{2}\|R(x_k) + J(x_k)p\|^2 + \frac{\lambda_k}{2}\|p\|^2$
- TR: $\min_p m_k^{TR}(p) = \frac{1}{2}\|R(x_k) + J(x_k)p\|^2,$
  s.t. $\|p\| \leq \Delta_k$

It is possible to prove that for TR p_k solves

$$(B_k + \lambda_k I)p_k = -g_k, \qquad B_k = J(x_k)^T J(x_k),\; g_k = J(x_k)^T R(x_k)$$

for some $\lambda_k \geq 0$ such that

$$\lambda_k(\|p_k\| - \Delta_k) = 0.$$

$\Rightarrow$ Trust-region methods are Levenberg-Marquardt methods!

I part: Ill-posed least squares problems

I part: Ill-posed least squares problems

Let us consider the following least squares problem: given $\mathcal{X}, \mathcal{Y}$ Hilbert spaces, $F : \mathcal{X} \to \mathcal{Y}$, nonlinear, continuously differentiable and $y \in \mathcal{Y}$, solve

$$\min_{x \in \mathcal{X}} \Phi(x) = \|F(x) - y\|^2_{\mathcal{Y}}.$$

Definition

The problem is well-posed if:

1. $\forall y \in \mathcal{Y}$ it exists a solution $x \in \mathcal{X}$,
2. the solution is unique,
3. property of stability holds (the solution depends continuously on the data).

The problem is ill-posed if one or more of the previous properties do not hold.

Ill-posed problems

- Let us consider problems of the form

$$\min_{x\in\mathbb{R}^n} \Phi(x) = \|F(x)-y\|^2, \quad x \in (\mathbb{R}^n, \|\cdot\|_2),\ y \in (\mathbb{R}^m, \|\cdot\|_2),$$

with $F : \mathbb{R}^n \to \mathbb{R}^m$ and $m \geq n$, arising from the discretization of an ill-posed problem.

- In a realistic situation the data y are affected by noise, we have at disposal only y^δ such that:

$$\|y - y^\delta\| \leq \delta$$

for some positive δ .

- We can handle only a noisy problem:

$$\min_{x\in\mathbb{R}^n} \Phi_\delta(x) = \|F(x) - y^\delta\|^2.$$

Need for regularization

- As stability does not hold, the solutions of the original problem do not depend continuously on the data.
  $\Longrightarrow$ The solutions of the noisy problem may not be meaningful approximations of the original problem solutions.

Need for regularization

- As stability does not hold, the solutions of the original problem do not depend continuously on the data.
  $\Longrightarrow$ The solutions of the noisy problem may not be meaningful approximations of the original problem solutions.
- Classical methods used for well-posed systems are not suitable in this contest.

$$\Downarrow$$

Need for regularization.

Iterative regularization methods

Iterative regularization methods generate a sequence $\{x_k^\delta\}$.
Regularizing properties arise from:

- construction of the iterates,
- the choice of a suitable stopping criterion.

If the process is stopped at iteration $k^*(\delta)$ the method is supposed to guarantee the following properties, given x^* a solution of the unperturbed problem:

- $x_{k^*(\delta)}^\delta$ is an approximation of x^*;
- $\{x_{k^*(\delta)}^\delta\}$ tends to x^* if δ tends to zero;
- local convergence to x^* in the noise-free case.

Outline

We consider regularizing trust-region approaches

1) Zero-residual problems: $F(x) = y^\delta$

It exists $x^\dagger$ such that $F(x^\dagger) = y$. We propose a regularizing trust-region approach, able to find an approximation to a solution of the unperturbed problem.

2) Non-zero residual problems: $\min_{x\in\mathbb{R}^n} \|F(x) - y^\delta\|^2$

It does not exist $x^\dagger$ such that $F(x^\dagger) - y = 0$.
We extend the trust-region approach designed for zero-residual problem to small residual problems.

Zero-residual problems

We consider

$$F(x) = y^{\delta},$$

with δ fixed noise level, and let $x^{\dagger}$ be a solution of $F(x) = y$.

Standard trust-region

Standard trust-region

The step p_k solves

$$(B_k + \lambda_k I)p_k = -g_k$$

for some $\lambda_k \geq 0$ such that

$$\lambda_k(\|p_k\| - \Delta_k) = 0.$$

- B_k is ill-conditioned.
- In trust-region methods the trust region is eventually inactive: $\|p_k\| < \Delta_k \rightarrow \lambda_k = 0$.
- It is not a regularization method!

How to obtain a regularizing method?

Noisy problem

$$\min_{x\in\mathbb{R}^n} \frac{1}{2}\|F(x) - y^\delta\|^2$$

Exact problem

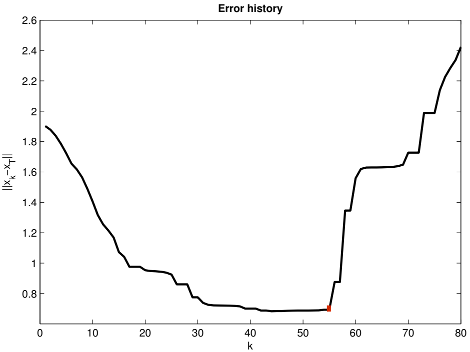
$$\min_{x\in\mathbb{R}^n} \frac{1}{2}\|F(x) - y\|^2$$

1. stopping criterion
2. small steps

1) Stopping criterion: with noisy data the process is stopped at iteration $k^*(\delta)$ such that $x^\delta_{k^*(\delta)}$ satisfies the discrepancy principle:

$$\|F(x^\delta_{k^*(\delta)}) - y^\delta\| \leq \tau\delta < \|F(x^\delta_k) - y^\delta\|$$

for $0 \leq k < k^*(\delta)$ and $\tau > 1$ suitable parameter.

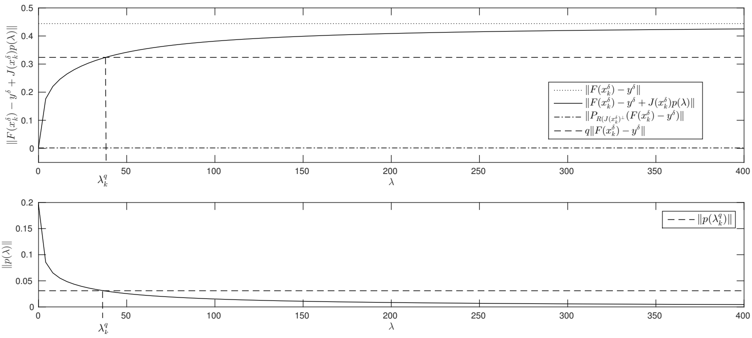


SEMI CONVERGENCE

Plot of the error $\|x^\delta_k - x^\dagger\|$ versus iteration number.

Regularizing trust-region

2) q-condition: $\|F(x_k^\delta) - y^\delta + J(x_k^\delta)p\| \geq q\|F(x_k^\delta) - y^\delta\|$, $q \in (0,1)$



$\rightarrow$ If $\Delta_k \leq \frac{1-q}{\|B_k\|}\|g_k^\delta\|$ then p_k satisfies the q-condition and the trust region is active.

Algorithm : k-th iteration of regularizing trust-region

Given x_k^δ, $\eta \in (0,1)$, $\gamma \in (0,1)$, $0 < C_{\min} < C_{\max}$.

Exact data: y, $q \in (0,1)$.

Noisy data: y^δ, $q \in (0,1)$, $\tau > 1/q$.

1. Compute $B_k = J(x_k^\delta)^T J(x_k^\delta)$ and $g_k^\delta = J(x_k^\delta)^T (F(x_k^\delta) - y^\delta)$.
2. Choose $\Delta_k \in \left[C_{\min} \|g_k^\delta\|, \min\left\{ C_{\max}, \frac{1-q}{\|B_k\|} \right\} \|g_k^\delta\| \right]$
3. Repeat

   3.1 Compute the solution p_k of trust-region problem.

   3.2 Compute

   $$\rho_k(p_k) = \frac{\Phi(x_k^\delta) - \Phi(x_k^\delta + p_k)}{m_k^{TR}(0) - m_k^{TR}(p_k)}$$

   with $\Phi(x) = \frac{1}{2}\|F(x) - y^\delta\|^2$, $m_k^{TR}(p) = \frac{1}{2}\|F(x_k^\delta) + J(x_k^\delta)p\|^2$.

   3.3 If $\rho_k(p_k) < \eta$,set $\Delta_k = \gamma \Delta_k$.

   Until $\rho_k(p_k) \geq \eta$.
4. Set $x_{k+1}^\delta = x_k^\delta + p_k$.

Local analysis

- **Assumption 1:** For index $\bar{k}$ it exist positive ρ and c such that
  1. the system $F(x) = y$ is solvable in $B_\rho(x^\delta_{\bar{k}})$;
  2. for $x, \tilde{x} \in B_{2\rho}(x^\delta_{\bar{k}})$ the following **tangential cone condition** holds,

$$\|F(x) - F(\tilde{x}) - J(x)(x - \tilde{x})\| \leq c\|x - \tilde{x}\|\|F(x) - F(\tilde{x})\|.$$

  For well-posed systems: $\|F(x) - F(\tilde{x}) - J(x)(x - \tilde{x})\| \leq c\|x - \tilde{x}\|^2$.

- **Assumption 2**: It exists positive K_J such that

$$\|J(x)\| \leq K_J$$

for all $x \in \mathcal{L} = \{x \in \mathbb{R}^n \;\; s.t. \;\; \Phi(x) \leq \Phi(x_0)\}$.

[Iterative regularization methods for nonlinear ill-posed problems, Kaltenbacher, Neubauer,Scherzer,2008]

Theoretical results

Lemma

The method generates a sequence $\{x_k^\delta\}$ such that:
1) the trust-region is active, i.e. $\lambda_k > 0$,
2) error decreases monotonically: $\|x_{k+1}^\delta - x^\dagger\| < \|x_k^\delta - x^\dagger\|$,
for $k \geq \bar{k}$, with $\bar{k} < k^*(\delta)$ for noisy data.

Theorem

If $\delta = 0$ the sequence $\{x_k\}$ converges to a solution x^* of $F(x) = y$ such that $\|x^* - x^\dagger\| \leq \rho$.
If $\delta > 0$ the discrepancy principle is satisfied after a finite number of iterations $k^*(\delta)$ and the sequence $\{x_{k^*(\delta)}^\delta\}$ converges to a solution of $F(x) = y$ if δ tends to zero.

$\rightarrow$ **Regularizing method**, [S. Bellavia, B. Morini, E. R., COAP, 2016].

Test problems

- Four nonlinear ill-posed systems arising from the discretization of **nonlinear first-kind Fredholm integral equation** are considered, they model gravimetric and geophysics problems:

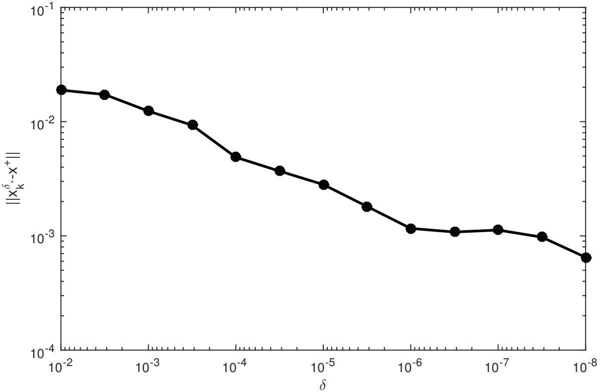
$$\int_0^1 k(t,s,x(s))ds = y(t), \qquad t \in [0,1],$$

  P1,**P2**, [Vogel, 1990], **P3**,**P4** [Kaltenbacher,2007];

- Their kernel is of the form

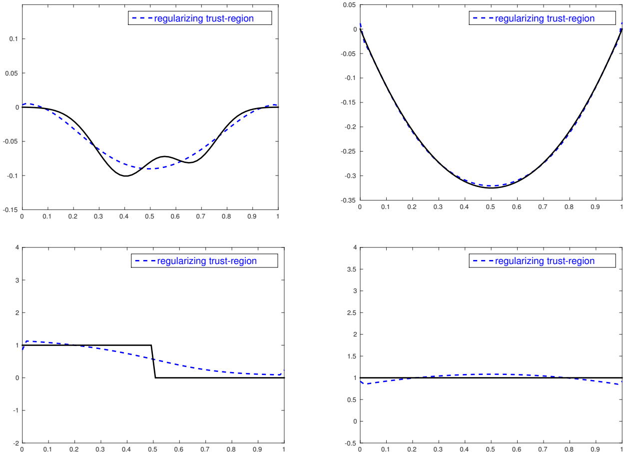
$$\begin{aligned} k(t,s,x(s)) &= log\left(\frac{(t-s)^2+H^2}{(t-s)^2+(H-x(s))^2}\right); \\ k(t,s,x(s)) &= \frac{1}{\sqrt{1+(t-s)^2+x(s)^2}}; \end{aligned}$$

Regularizing properties of the method.



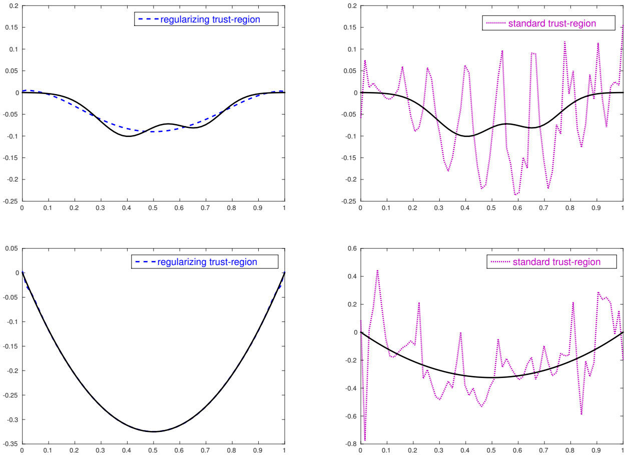
Semilogarithmic plot of the error $||x^{\delta}_{k^*(\delta)} - x^{\dagger}||$ as a function of the noise level δ.

Computed solution approximations



$\delta = 10^{-2}$. Blue: regularizing TR, **Solid line**: solution of the original problem.

Comparison between regularizing and standard trust-region



$\delta = 10^{-2}$. **Left**: regularizing TR, **Right**: standard TR , **Solid line**: solution of the original problem.

Contribution

- Theoretical study and implementation of a Regularizing Trust-region approach
- The methods represents an improvement over the Levenberg-Marquardt method in [Hanke 1996] based on the condition

$$\|F(x_k^\delta) - y^\delta + J(x_k^\delta)p_k(\lambda_k)\| = q\|F(x_k^\delta) - y^\delta\| \tag{1}$$

  which is not ensured to have a solution far from $x^\dagger$, while the condition we adopted can always be satisfied. The proposed method results to be more robust.
- The Trust-region approach is also shown to be less-dependent on the free parameters of the method (q).
- We analyzed the practical implementation of the method in [Hanke 1996] that was not considered in the original paper or in related articles. Specifically we discuss how to solve (1) in a reliable way.

Non-zero residual problems

It does not exist x such that $F(x) - y = 0$, but it exists $x^{\dagger}$ local minimum of the problem

$$\min_x \frac{1}{2}\|F(x) - y\|^2.$$

- Non-zero residual problems frequently appear in applications, especially when a natural phenomenon is represented through a mathematical model.
  - The most part of the literature on ill-posed nonlinear least squares deals with zero residual problems, we are not aware of other contributions on this topic.
- Usually the modelling error is incorporated in the data error and the problem is solved as a zero residual problem
  - Estimation of the modelling error is not required.

Small residual problems

- We extend the approach for zero-residual problems to small residual problems. $\rightarrow$ We propose an **elliptical trust-region approach.**

At a generic iteration k, given $\Delta_k > 0$, the following problem is solved:

$$\min_p m_k(p) := \frac{1}{2}\|F(x_k^\delta) - y^\delta + J(x_k^\delta)p\|^2,$$
$$\text{s.t. } \|(B_k)^{-\frac{1}{2}}p\| \leq \Delta_k.$$

1. **discrepancy principle** :

$$\|J(x_{k^*(\delta)}^\delta)^T(F(x_{k^*(\delta)}^\delta) - y^\delta)\| \leq \tau\delta < \|J(x_k^\delta)^T(F(x_k^\delta) - y^\delta)\|$$

2. **q-condition**:

$$\|J(x_k^\delta)^T(F(x_k^\delta) - y^\delta + J(x_k^\delta)p_k)\| \geq q\|J(x_k^\delta)^T(F(x_k^\delta) - y^\delta)\|$$

Regularizing method [S.Bellavia, E.R., submitted to JOTA (second revision)]

II part: Large scale problems with expensive objective function

6-months collaboration with S. Gratton, INP-ENSEEIHT, Toulouse.

II part: Large scale problems with expensive objective function

- We consider large-scale problems for which the objective function is expensive to evaluate:

$$\min_x \Phi(x) = \frac{1}{2}\|F(x)\|^2$$

- We consider an iterative process that employs a sequence of approximations $\{\Phi_{\delta_k}\}$ of the original objective function

$$\Phi_{\delta_k}(x) = \frac{1}{2}\|F_{\delta_k}(x)\|^2, \quad F_{\delta_k} \sim F$$

- δ_k is the accuracy level of the approximations:

$$|\Phi_{\delta_k}(x_k) - \Phi(x_k)| \leq \delta_k.$$

- We assume that the accuracy level can be improved along the optimization process: $\delta_k \searrow 0$.

Typical applications

Subsampling techniques

- Machine learning, Data assimilation.
- Large set of data at disposal: $\{1, \ldots, N\}$.
  Subsampling: $X_k \subseteq \{1, \ldots, N\}$ such that $|X_k| = K_k \leq N$ is selected.
- $F_{\delta_k} : \mathbb{R}^n \to \mathbb{R}^{K_k}$ such that $(F_{\delta_k})_i = F_j$, $j \in X_k$ is built.
- $\Phi_{\delta_k}(x) = \frac{1}{2}\|F_{\delta_k}(x)\|^2$
- approximation can be improved by considering more observations.

Iterative methods

- Φ is the result of an iterative process (solution of a nonlinear equation or an inversion process) that can be stopped when a certain accuracy level is reached.
- By varying the stopping criterion we vary the accuracy of the approximation.

Levenberg-Marquardt method

- We consider a Levenberg-Marquardt method that at each iteration uses an approximated model employing the approximations to function and derivatives:

$$m_k(p_k) = \frac{1}{2}\|F_{\delta_k}(x_k) + J_{\delta_k}(x_k)p_k\|^2 + \frac{\lambda_k}{2}\|p_k\|^2$$

  for J_{δ_k} an approximation to J.
- At each iteration the step is found minimizing the noisy model, i.e. solving a linear systems of the form:

$$(J_{\delta_k}(x_k)^T J_{\delta_k}(x_k) + \lambda_k I)p_k = -g_{\delta_k}(x_k), \quad g_{\delta_k}(x_k) = J_{\delta_k}(x_k)^T F_{\delta_k}(x_k)$$

Inexact step

Large-scale problems: approximate solution of LM subproblem

p provides the sufficient Cauchy decrease:

$$m_k(0) - m_k(p_k) \geq \frac{\theta}{2} \frac{\|g_{\delta_k}(x_k)\|^2}{\|J_{\delta_k}(x_k)\|^2 + \lambda_k}, \qquad \theta > 0.$$

The Levenberg-Marquardt step computed as

$$(J_{\delta_k}(x_k)^T J_{\delta_k}(x_k) + \lambda_k I)p_k = -g_{\delta_k}(x_k) + r_k$$

for a residual r_k satisfying $\|r_k\| \leq \epsilon_k \|g_{\delta_k}(x_k)\|$, with ϵ_k such that

$$0 \leq \epsilon_k \leq \min \left\{ \frac{\theta_1}{\lambda_k^\alpha}, \sqrt{\theta_2 \frac{\lambda_k}{\|J_{\delta_k}(x_k)\|^2 + \lambda_k}} \right\},$$

where $\theta_1 > 0$, $\theta_2 \in \left(0, \frac{1}{2}\right]$ and $\alpha \in \left[\frac{1}{2}, 1\right)$ achieves the Cauchy decrease.

Step acceptance

- After the step is computed, we have to decide whether to accept the step.
- Step acceptance is based on the ratio:

$$\rho_k^{\delta_k}(p_k) = \frac{\Phi_{\delta_k}(x_k) - \Phi_{\delta_k}(x_k + p_k)}{m_k(0) - m_k(p_k)}.$$

- If the noise is too high, the reduction in Φ_{δ_k} can be just an effect of the presence of the noise.

Step acceptance

- After the step is computed, we have to decide whether to accept the step.
- Step acceptance is based on the ratio:

$$\rho_k^{\delta_k}(p_k) = \frac{\Phi_{\delta_k}(x_k) - \Phi_{\delta_k}(x_k + p_k)}{m_k(0) - m_k(p_k)}.$$

- If the noise is too high, the reduction in Φ_{δ_k} can be just an effect of the presence of the noise.

Need for a strategy to control the noise!!

Noise control

Noise control

[Trust region methods, 2000] Let

$$\max\{|\Phi_{\delta_k}(x_k) - \Phi(x_k)|, |\Phi_{\delta_k}(x_k + p_k) - \Phi(x_k + p_k)|\} \leq \delta_k$$
$$\delta_k \leq \eta_0(m_k(0) - m_k(p_k)).$$

If

$$\rho_k^{\delta_k}(p_k) = \frac{\Phi_{\delta_k}(x_k) - \Phi_{\delta_k}(x_k + p_k)}{m_k(0) - m_k(p_k)} > \eta$$

then also

$$\rho_k(p_k) = \frac{\Phi(x_k) - \Phi(x_k + p_k)}{m_k(0) - m_k(p_k)} > \eta.$$

$\rightarrow$ True reduction in the noise-free objective function Φ

In our approach: $m_k(0) - m_k(p_k) \sim \frac{1}{2}\lambda_k \|p_k\|^2$.

Algorithm : k-th iteration of regularizing Levenberg-Marquardt

Given $\alpha \in \left(\frac{1}{2}, 1\right]$, δ_0, $\eta_1 \in (0,1)$, $\eta_2 > 0$, $\lambda_{\max} > \lambda_{\min} > 0$, $\gamma > 1$, x_0 and $\lambda_{\max} > \lambda_0 \geq \lambda_{\min}$.

Compute $\Phi^{\delta_0}(x_0)$. For $k = 0, 1, 2, \dots$

1. Compute a solution p_k of the LM subproblem.
2. If $\delta_k \leq \kappa_d \frac{1}{2} \lambda_k^{\alpha} \|p_k\|^2$, compute $\Phi_{\delta_k}(x_k + p_k)$, else reduce δ_k and go back to 1.
3. Compute
$$\rho_k^{\delta_k}(p_k) = \frac{\Phi_{\delta_k}(x_k) - \Phi_{\delta_k}(x_k + p_k)}{m_k(0) - m_k(p_k)}.$$
3.1 If $\rho_k^{\delta_k}(p_k) \geq \eta_1$, then set $x_{k+1} = x_k + p_k$ and update λ.
3.2 Otherwise set $x_{k+1}^{\delta_k} = x_k^{\delta_k}$, $\lambda_{k+1} = \gamma \lambda_k$.

Regularization Parameter update

The parameter update is inspired by [Bergou, Gratton, Vicente, 2016] and [Bandeira, Scheinberg, Vicente, 2014]. If success, given $\gamma > 1$:

$$\lambda_{k+1} = \begin{cases} \min\{\gamma\lambda_k, \lambda_{\max}\} & \text{if } \|g_{\delta_k}(x_k^\delta)\| < \eta_2/\lambda_k, \\ \max\{\lambda_k, \lambda_{\min}\} & \text{if } \|g_{\delta_k}(x_k^\delta)\| \geq \eta_2/\lambda_k. \end{cases}$$

Gradient approximations

We can control the accuracy on the gradient approximation:
$\frac{\|g(x_k)\|}{(1+c_k)} \leq \|g_{\delta_k}(x_k)\| \leq \frac{\|g(x_k)\|}{(1-c_k)}$, with $c_k = O\left(\frac{1}{\lambda_k^{1-\alpha/2}}\right)$.

Assumptions

- Assumption 1:
  Function f is continuously differentiable, and it exists $\kappa_J > 0$ such that for all $k \geq 0$ and all $x \in [x_k, x_k + p_k^{LM}]$, $\|J_{\delta_k}(x)\| \leq \kappa_J$.
- Assumption 2: f has Lipschitz continuous gradient:
  $\|g(x) - g(y)\| \leq L\|x - y\|$ for all $x, y \in \mathbb{R}^n$.

Let the residual be small enough, i.e. r_k satisfies $\|r_k\| \leq \epsilon_k \|g_{\delta_k}\|$, with

$$\epsilon_k \leq \min\left\{ \frac{\theta_1}{\lambda_k^\alpha}, \sqrt{\theta_2 \frac{\lambda_k}{\|J_{\delta_k}(x_k)\|^2 + \lambda_k}} \right\}$$

where $\theta_1 > 0$, $\theta_2 \in \left(0, \frac{1}{2}\right]$ and $\alpha \in \left[\frac{1}{2}, 1\right)$.

Lemma

The sequences $\{\delta_k\}$ and $\{x_k\}$ generated by the Algorithm are such that

$$\lim_{k\to\infty} \delta_k \leq \lim_{k\to\infty} \frac{1}{2}\lambda_k^\alpha \|p_k\|^2 = 0 \qquad \lim_{k\to\infty} \|g(x_k)\| = 0.$$

[S.Bellavia, S.Gratton, E.R., submitted to Numerische Mathematik (second revision)].

Local convergence

Asymptotic step behaviour

The LM step asymptotically tends to the direction of the negative perturbed gradient:

$$\lim_{k\to\infty}(p_k^{LM})_i + \frac{\theta}{\kappa_J^2+\lambda_k}(g_{\delta_k}(x_k))_i = 0 \quad \text{for} \quad i = 1, \ldots, n,$$

where $(\cdot)_i$ denotes the i-th vector component.

Lemma

Let $p_k^{SD} = -\frac{\theta}{\kappa_J^2+\lambda_k} g_{\delta_k}(x_k)$ and $x_{k+1} = x_k + p_k^{SD}$. If $x_{\bar{k}} \in B_r(x^*)$ and $\lambda_{\bar{k}}$ big enough,

- $\|x_{k+1} - x^*\| < \|x_k - x^*\|$, for all $k \geq \bar{k}$.
- $\|x_k - x^*\|$ tends to zero.

Complexity analysis

Assumption

Let assume that the procedure is stopped when $\|g_{\delta_k}(x_k)\| \leq \epsilon$.

- The number of successful iterations N_1 is bounded above by:

$$N_1 \leq O(\epsilon^{-2}).$$

- The number of unsuccessful iterations N_3 is bounded above by a constant independent of ϵ:

$$N_3 \leq c(\lambda_{\max}, \lambda_0, \gamma).$$

Complexity

Standard Levenberg-Marquardt methods complexity is preserved:

$$N_T = O(\epsilon^{-2}),$$

- **Data assimilation problem**. Nonlinear wave equation:

$$\frac{\partial^2 u(z,t)}{\partial t^2} - \frac{\partial^2 u(z,t)}{\partial z^2} + \mu e^{\nu u} = 0,$$
$$u(0,t) = u(1,t) = 0, u(z,0) = u_0(z),$$
$$\frac{\partial u(z,0)}{\partial t} = 0,\ 0 \leq t \leq T,\ 0 \leq z \leq 1.$$

  We look for initial state $u_0(z)$.

- **Machine learning problem**. Binary classification problem: $\{(z^i, y^i)\}$ with $z^i \in \mathbb{R}^n$, $y^i \in \{-1,+1\}$ and $i = 1,\ldots,N$.
  Training objective function: logistic loss with l_2 regularization

$$f(x) = \frac{1}{2N}\sum_{i=1}^{N} \log(1+\exp(-y^i x^T z^i)) + \frac{1}{2N}\|x\|^2.$$

Data assimilation

We look for the initial state $u_0(z)$, from the knowledge of observations $u(z_i, t_j)$, $t_j > 0$. Data assimilation problem:

$$\min_{x \in \mathbb{R}^n} \frac{1}{2}\|x - x_b\|^2_{B^{-1}} + \frac{1}{2}\sum_{j=0}^{N_t} \|H_j(x(t_j)) - y_j\|^2_{R_j^{-1}}$$

- $\|x\|^2_M = x^T M x$ for a symmetric positive definite matrix M,
- $x_b \in \mathbb{R}^n$ is the background vector (a priori estimate)
- $y_j \in \mathbb{R}^{m_j}$ is the vector of observations at time t_j, $m_j \leq n$.
- H_j is the operator modelling the observation process at t_j
- $x(t_j)$ the state vector, solution of the nonlinear model at time t_j.

Build the approximations

- We build the approximations through subsampling techniques.
- In both cases

$$\Phi(x) = \sum_{i=1}^{N} \Phi_i(x)^2.$$

- Function approximations:

$$\Phi_{\delta_k}(x) = \sum_{i \in X_k} \Phi_i(x)^2$$

  with $X_k \subset \{1, \ldots, N\}$.
- Increasing the size of X_k we have a better approximation.

	Data Assimilation		Machine learning	
	All samples	Subsampled	All samples	Subsampled
it	9	12	52	38
$\mathbf{cost_f}$	10	3	53	16
$\mathbf{cost_p}$	67	15	808	316
RMSE	1.2e-2	3.8e-2	5.4e-2	6.0e-2
$\mathbf{save_f}$		67%		70%
$\mathbf{save_p}$		78%		61%

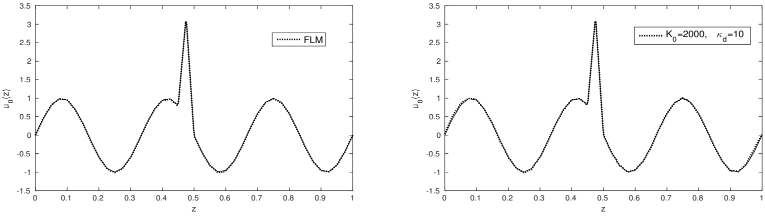


Figure: Solution approximation, Left: all samples, Right: Subsampled

Contribution

- We proposed a method to solve least squares problems with both noisy function and gradients.
- We are not aware of methods for noisy non-zero residual nonlinear least squares problems, for which both local and global convergence is proved.
- The proposed Levenberg-Marquardt method allows considerable savings in terms of function evaluations and matrix-vector products compared to inexact Levenberg-Marquardt methods and Gauss-Newton methods employing the exact objective function and Jacobian.

Development of the code implementing three numerical methods:

- *Regularizing Trust-Region method*. Ill-posed nonlinear least-squares problems with zero-residual.
- *Elliptical regularizing Trust-Region method*. Ill-posed nonlinear least-squares problems with non zero-residual.
- *Levenberg-Marquardt method for large scale problems with dynamic noise*. Large scale problems for least-squares problem with objective function that can be computed with dynamic accuracy.

Research Outputs

Articles related to the thesis:

- S.Bellavia, B.Morini, E.Riccietti, *On an adaptive regularization for ill-posed nonlinear systems and its trust-region implementation* (Computational Optimization and Applications, 2016).
- S.Bellavia, E.Riccietti, *On non-stationary Tikhonov procedures for ill-posed nonlinear least squares problems*, submitted to Journal of Optimization Theory and Applications (second revision).
- S.Bellavia, S.Gratton, E.Riccietti, *A Levenberg-Marquardt method for large nonlinear least squares problems with noisy functions and gradients*, submitted to Numerische Mathematik (second revision).

Other articles:

- E.Riccietti, J.Bellucci, M.Checcucci, M.Marconcini, A.Arnone, *Support Vector Machine classification applied to the parametric design of centrifugal pumps*, (Engineering Optimization, 2017).
- E.Riccietti, S.Bellavia, S.Sello, *Numerical methods for optimization problems arising in energetic districts*, (ECMI proceeding, 2016).
- E.Riccietti, S.Bellavia, S.Sello, *Sequential Linear Programming and Particle Swarm Optimization for the optimization of energy districts*, (Engineering Optimization, 2018).

Perspectives

Solution of large scale ill-posed problems.

- Variant of the elliptical Trust-Region approach. Critical point: cannot compute square root of matrix B_k or solve linear systems exactly: need of iterative solvers that introduces a source of inexactness.
- Extension of the method presented in Part II to allow input spaces of increasing dimensions, to include also multilevel strategies.
  Ideas on which the methods presented in Part I and Part II are based can be coupled, to design a method suitable for handling discrete ill-posed problems arising from a discretization of the input space of an infinite dimensional problem: adaptive choice of mesh size.

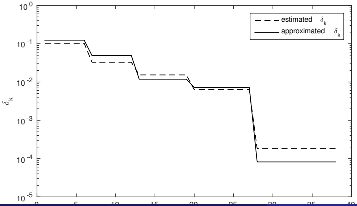
Thank you for your attention!

Noise estimation

- We estimate the noise computing the true objective function when the noise control is not satisfied: $\delta_k \sim |\Phi(x_k) - \Phi_{\delta_k}(x_k)|$.
- We could use an estimate:

$$\delta_k \simeq \frac{\sqrt{2(N-K_k)}}{K_k}, \quad \text{with} \quad K_k = |X_k|.$$

  If the components $F_i(x)$ of $F(x)$ were Gaussian, $\sum_{i=1}^{N-K_k} F_i(x)^2$ would follow a Chi-squared distribution with standard deviation $\sqrt{2(N-K_k)}$.
- non-deterministic estimate: not supported by our theory



Solver	**it**	**cost$_\mathbf{f}$**	**cost$_\mathbf{p}$**	**err**
SSLM$_{est}$	38	15.9	316.7	5.4e-2
SSLM$_{appr}$	37	17.7	318.1	5.7e-2

Comparison between regularizing TR-LM, $\delta = 10^{-2}$.

Problem	x_0	Regularizing TR it	nf	cf	Regularizing LM it	nf	cf
P1	$0\ e$	20	21	6	17	18	4
	$-0.5\ e$	29	30	6	22	23	4
	$-1\ e$	35	36	5	24	25	4
	$-2\ e$	40	41	5	25	26	4
P2	$0\ e$	30	31	5	*	*	*
	$0.5\ e$	25	26	5	*	*	*
	$1\ e$	29	30	5	22	23	5
	$2\ e$	37	39	5	25	26	5
P3	$x_0(1.25)$	15	16	4	12	13	4
	$x_0(1.5)$	17	18	4	14	15	4
	$x_0(1.75)$	19	20	4	15	16	4
	$x_0(2)$	22	23	4	16	17	4
P4	$x_0(1,1)$	17	18	5	10	11	4
	$x_0(0.5,0)$	20	21	4	*	*	*
	$x_0(1.5,1)$	22	23	4	15	16	4
	$x_0(1.5,0)$	26	27	4	*	*	*

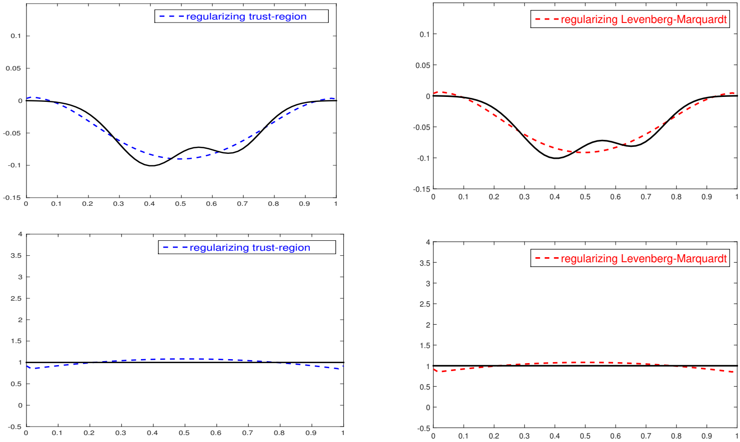
it=iterations, **nf**=function evaluations,

cf=mean number of Cholesky factorizations.

$*$=failure, reached maximum number of iterations or convergence to a solution of the noisy problem

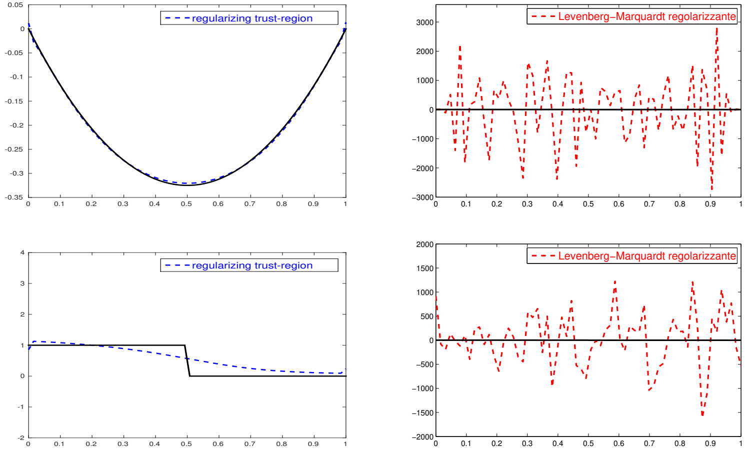
$e = (1,\dots,1)^T$, **P3:** $(x_0(\alpha))_j = (-4\alpha+4)s_j^2 + (4\alpha-4)s_j + 1$, **P4:** $x_0(\beta,\chi) = \beta - \chi s_j$, s_j grid points, $j = 1,\dots,n$.

Comparison between regularizing TR and LM, $\delta = 10^{-2}$



Left: regularizing TR, **Right**: regularizing LM , **Solid line**: solution of the original problem.

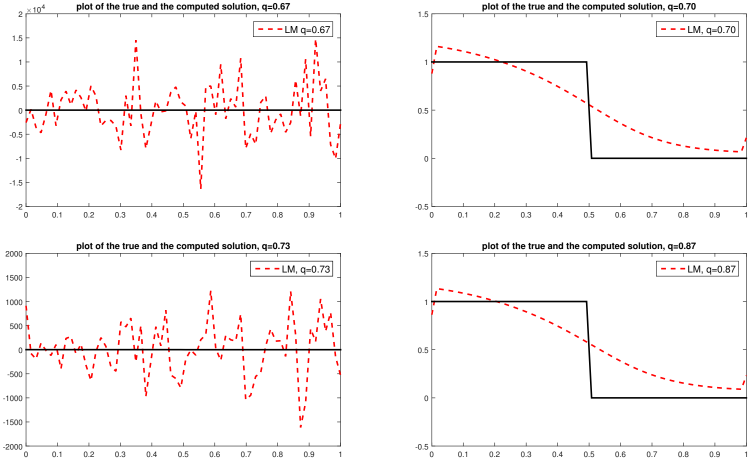
Comparison between regularizing TR e LM, $\delta = 10^{-2}$



Left: regularizing TR , **Right**: regularizing LM , **Solid line**: solution of the original problem.

The q-condition

The condition imposed by Hanke is strongly dependent on the choice of the value of free parameter q. Values of $q = 0.67, 0.70, 0.73, 0.87$, $\delta = 10^{-2}$.



1. **P1**: We want to reconstruct c in the 2D-elliptic problem

$$-\Delta u + cu = \hat{f} \text{ in } \Omega = (0,1) \times (0,1)$$
$$u = \hat{g} \text{ on } \partial\Omega$$

from the knowledge of u in Ω, $\hat{f} \in L^2(\Omega)$, $\hat{g}$ the trace of a function in $H^2(\Omega)$. If $F : D(F) \to L^2(\Omega)$ is the operator mapping parameter c to the solution u we solve

$$\min_c \frac{1}{2}\|F(c) - \tilde{u}\|^2$$

$\tilde{u}$ measured values of u.

2. **P2**: Reconstruct the conductivity x of the soil from measurements $b = (b_1, \ldots, b_m)^T$ at different heights $h_i, i = 1, \ldots, m$:

$$\min_x \frac{1}{2}\|m(x) - b\|^2.$$

Numerical tests on problem P1, $\delta = 10^{-2}$

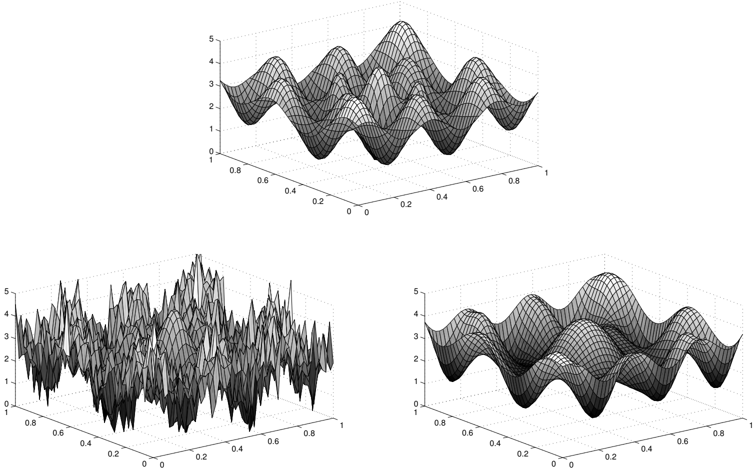


Figure: Solution approximations. Up: RTR with exact data. Lower part: standard trust-region (left) and RTR (right) for $\delta = 10^{-2}$.

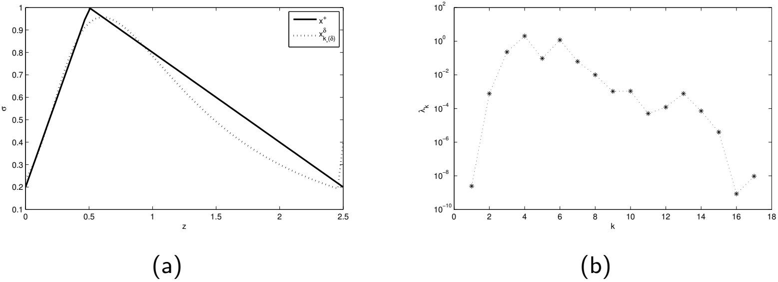


Figure: (a) plot of the true solution $x^\dagger$ and of the computed solution $x^\delta_{k^*(\delta)}$ for $\delta = 10^{-2}$, (b) regularization parameters λ_k.