Algorithms and application for special classes of nonlinear least squares problems 2023

Clément W. Royer

US-Mexico workshop on optimization and its applications

January 9, 2023

Not so long ago, not so far away...

BULL. AUSTRAL. MATH. SOC.
VOL. 31 (1985), 309-311.

90C30
(65K05)

ALGORITHMS AND APPLICATION FOR SPECIAL CLASSES OF NONLINEAR LEAST SQUARES PROBLEMS

STEPHEN J. WRIGHT

A Nonlinear Least Squares problem is an optimization problem for which the objective function to be minimized has the form

$$F(x) = \sum_{i=1}^{m} f_i^2(x) \ , \quad x \in R^n \ , \quad m \geq n \ .$$

- Nonlinear least squares + constraints/nonsmoothness;
- Efficient solve for sparse Jacobian;
- Cool application (geophysics).

What more can I do?

Rebooting the (least-squares) franchise

Today's goals

Nonlinear least squares

- A basic problem...
- ...with modern instances.

Revisit algorithms

- Complexity bounds for Gauss-Newton methods.
- Complexity metrics.

Go a bit further

- Inexactness and stochasticity.
- Probabilistic results.

Outline

Outline

General setup

Nonlinear least-squares

$$\min_{\boldsymbol{x}\in\mathbb{R}^n} f(\boldsymbol{x}) := \frac{1}{2}\|\boldsymbol{r}(\boldsymbol{x})\|^2, \qquad \boldsymbol{r}:\mathbb{R}^n \mapsto \mathbb{R}^m, \boldsymbol{r}\in\mathcal{C}^2.$$

Jacobian: $\boldsymbol{J}(\boldsymbol{x}) := \left[\nabla r_i(\boldsymbol{x})^\top\right] \in \mathbb{R}^{m\times n}$.

General setup

Nonlinear least-squares

$$\min_{\boldsymbol{x}\in\mathbb{R}^n} f(\boldsymbol{x}) := \frac{1}{2}\|\boldsymbol{r}(\boldsymbol{x})\|^2, \qquad \boldsymbol{r}:\mathbb{R}^n \mapsto \mathbb{R}^m, \boldsymbol{r}\in\mathcal{C}^2.$$

Jacobian: $\boldsymbol{J}(\boldsymbol{x}) := \left[\nabla r_i(\boldsymbol{x})^\top\right] \in \mathbb{R}^{m\times n}$.

Gauss-Newton techniques

- Gauss-Newton model $f(\boldsymbol{x}+\boldsymbol{s}) \approx \frac{1}{2}\|\boldsymbol{r}(\boldsymbol{x}) + \boldsymbol{J}(\boldsymbol{x})\boldsymbol{s}\|^2$;
- Steps computed via line search/trust region;
- Hessian approximated by $\boldsymbol{J}(\boldsymbol{x})^{\mathrm{T}}\boldsymbol{J}(\boldsymbol{x})$.

General setup

Nonlinear least-squares

$$\min_{\boldsymbol{x}\in\mathbb{R}^n} f(\boldsymbol{x}) := \frac{1}{2}\|\boldsymbol{r}(\boldsymbol{x})\|^2, \qquad \boldsymbol{r}:\mathbb{R}^n \mapsto \mathbb{R}^m, \boldsymbol{r}\in\mathcal{C}^2.$$

Jacobian: $\boldsymbol{J}(\boldsymbol{x}) := \left[\nabla r_i(\boldsymbol{x})^\top\right] \in \mathbb{R}^{m\times n}$.

Gauss-Newton techniques

- Gauss-Newton model $f(\boldsymbol{x}+\boldsymbol{s}) \approx \frac{1}{2}\|\boldsymbol{r}(\boldsymbol{x}) + \boldsymbol{J}(\boldsymbol{x})\boldsymbol{s}\|^2$;
- Steps computed via line search/trust region;
- Hessian approximated by $\boldsymbol{J}(\boldsymbol{x})^{\mathrm{T}}\boldsymbol{J}(\boldsymbol{x})$.

Levenberg(-Morrison)-Marquardt

- Regularized Gauss-Newton model:
  $f(\boldsymbol{x}+\boldsymbol{s}) \approx \frac{1}{2}\|\boldsymbol{r}(\boldsymbol{x}) + \boldsymbol{J}(\boldsymbol{x})\boldsymbol{s}\|^2 + \frac{\gamma}{2}\|\boldsymbol{s}\|^2$;
- **Regularization parameter** γ set adaptively.

Levenberg-Marquardt for $\min_{x \in \mathbb{R}^n} \frac{1}{2}\|r(x)\|^2$

Inputs: $x_0 \in \mathbb{R}^n, \gamma_0 \geq \gamma_{\min} > 0, \eta > 0$.
Iteration k: Given (x_k, γ_k),

- Compute

$$s_k \approx \operatorname*{argmin}_s m_k(s) := \frac{1}{2}\|r(x_k) + J(x_k)s\|^2 + \frac{\gamma_k}{2}\|s\|^2.$$

- If

$$\frac{\frac{1}{2}\|r(x_k)\|^2 - \frac{1}{2}\|r(x_k + s_k)\|^2}{m_k(0) - m_k(s_k)} \geq \eta,$$

  set $x_{k+1} = x_k + s_k$ and $\gamma_{k+1} = \max\{0.5\gamma_k, \gamma_{\min}\}$;
- Otherwise, set $x_{k+1} = x_k$ and $\gamma_{k+1} = 2\gamma_k$.

Levenberg-Marquardt for $\min_{x \in \mathbb{R}^n} \frac{1}{2}\|r(x)\|^2$

Inputs: $x_0 \in \mathbb{R}^n, \gamma_0 \geq \gamma_{\min} > 0, \eta > 0$.
Iteration k: Given (x_k, γ_k),

- Compute

$$s_k \approx \operatorname*{argmin}_s m_k(s) := \frac{1}{2}\|r(x_k) + J(x_k)s\|^2 + \frac{\gamma_k}{2}\|s\|^2.$$

- If

$$\frac{\frac{1}{2}\|r(x_k)\|^2 - \frac{1}{2}\|r(x_k + s_k)\|^2}{m_k(0) - m_k(s_k)} \geq \eta,$$

  set $x_{k+1} = x_k + s_k$ and $\gamma_{k+1} = \max\{0.5\gamma_k, \gamma_{\min}\}$;
- Otherwise, set $x_{k+1} = x_k$ and $\gamma_{k+1} = 2\gamma_k$.

Goal: Prove a complexity result for the method.

Complexity in nonconvex optimization

Setup: Sequence of points $\{\boldsymbol{x}_k\}$ generated by an algorithm applied to $\min_{\boldsymbol{x}\in\mathbb{R}^n} f(\boldsymbol{x})$.

Complexity in nonconvex optimization

Setup: Sequence of points $\{\boldsymbol{x}_k\}$ generated by an algorithm applied to $\min_{\boldsymbol{x}\in\mathbb{R}^n} f(\boldsymbol{x})$.

Standard complexity result

Given $\epsilon_g \in (0, 1)$:

- **Worst-case** number of iterations to obtain $\boldsymbol{x}_k$ such that $\|\nabla f(\boldsymbol{x}_k)\| \leq \epsilon_g$.
- Focus: **Dependency on** ϵ_g.

Complexity in nonconvex optimization

Setup: Sequence of points $\{\boldsymbol{x}_k\}$ generated by an algorithm applied to $\min_{\boldsymbol{x}\in\mathbb{R}^n} f(\boldsymbol{x})$.

Standard complexity result

Given $\epsilon_g \in (0, 1)$:

- **Worst-case** number of iterations to obtain $\boldsymbol{x}_k$ such that $\|\nabla f(\boldsymbol{x}_k)\| \leq \epsilon_g$.
- Focus: **Dependency on** ϵ_g.

Some examples

- Gradient descent: $\mathcal{O}(\epsilon_g^{-2})$ iterations.
- Newton: $\mathcal{O}(\epsilon_g^{-2})$ iterations.
- Cubic regularization/Modified Newton: $\mathcal{O}(\epsilon_g^{-3/2})$ iterations.

The least-squares setting

- Problem: $\min_{\boldsymbol{x} \in \mathbb{R}^n} \frac{1}{2}\|\boldsymbol{r}(\boldsymbol{x})\|^2$:
- Goal: Find $\boldsymbol{x}_k$ such that $\|\boldsymbol{J}(\boldsymbol{x}_k)^{\mathrm{T}}\boldsymbol{r}(\boldsymbol{x}_k)\| \leq \epsilon_g$.

The least-squares setting

- Problem: $\min_{\boldsymbol{x} \in \mathbb{R}^n} \frac{1}{2}\|\boldsymbol{r}(\boldsymbol{x})\|^2$:
- Goal: Find $\boldsymbol{x}_k$ such that $\|\boldsymbol{J}(\boldsymbol{x}_k)^{\mathrm{T}}\boldsymbol{r}(\boldsymbol{x}_k)\| \leq \epsilon_g$.

Some results

- Gradient descent: $\mathcal{O}(\epsilon_g^{-2})$ iterations.
- Gauss-Newton + line search/trust region: $\mathcal{O}(\epsilon_g^{-2})$ iterations.
- Levenberg-Marquardt: $\mathcal{O}(\epsilon_g^{-2})$ iterations.

Levenberg-Marquardt for $\min_{\boldsymbol{x}\in\mathbb{R}^n} \frac{1}{2}\|\boldsymbol{r}(\boldsymbol{x})\|^2$

Inputs: $\boldsymbol{x}_0 \in \mathbb{R}^n, \gamma_0 \geq \gamma_{\min} > 0, \eta > 0.$

Iteration k**:** Given $(\boldsymbol{x}_k, \gamma_k)$,

- Compute

$$\boldsymbol{s}_k \approx \operatorname*{argmin}_{\boldsymbol{s}} m_k(\boldsymbol{s}) := \frac{1}{2}\|\boldsymbol{r}(\boldsymbol{x}_k) + \boldsymbol{J}(\boldsymbol{x}_k)\boldsymbol{s}\|^2 + \frac{\gamma_k}{2}\|\boldsymbol{s}\|^2.$$

- If

$$\frac{\frac{1}{2}\|r(\boldsymbol{x}_k)\|^2 - \frac{1}{2}\|\boldsymbol{r}(\boldsymbol{x}_k + \boldsymbol{s}_k)\|^2}{m_k(0) - m_k(\boldsymbol{s}_k)} \geq \eta,$$

  set $\boldsymbol{x}_{k+1} = \boldsymbol{x}_k + \boldsymbol{s}_k$ and $\gamma_{k+1} = \max\{0.5\gamma_k, \gamma_{\min}\}$;

- Otherwise, set $\boldsymbol{x}_{k+1} = \boldsymbol{x}_k$ and $\gamma_{k+1} = 2\gamma_k$.

Levenberg-Marquardt for $\min_{x \in \mathbb{R}^n} \frac{1}{2}\|r(x)\|^2$

Inputs: $x_0 \in \mathbb{R}^n, \gamma_0 \geq \gamma_{\min} > 0, \eta > 0$.
Iteration k: Given (x_k, γ_k),

- Compute

$$s_k \approx \operatorname*{argmin}_s m_k(s) := \frac{1}{2}\|r(x_k) + J(x_k)s\|^2 + \frac{\gamma_k}{2}\|s\|^2.$$

- If

$$\frac{\frac{1}{2}\|r(x_k)\|^2 - \frac{1}{2}\|r(x_k + s_k)\|^2}{m_k(0) - m_k(s_k)} \geq \eta,$$

  set $x_{k+1} = x_k + s_k$ and $\gamma_{k+1} = \max\{0.5\gamma_k, \gamma_{\min}\}$;
- Otherwise, set $x_{k+1} = x_k$ and $\gamma_{k+1} = 2\gamma_k$.

Goal: Prove a complexity result for the method.

Main arguments

Decrease guarantee

For any successful iteration ($\boldsymbol{x}_{k+1} \neq \boldsymbol{x}_k$),

$$\|r(\boldsymbol{x}_k)\|^2 - \|r(\boldsymbol{x}_{k+1})\|^2 \geq \mathcal{O}\left(\frac{\|\boldsymbol{J}(\boldsymbol{x}_k)^{\mathrm{T}}\boldsymbol{r}(\boldsymbol{x}_k)\|^2}{\gamma_k}\right).$$

Main arguments

Decrease guarantee

For any successful iteration ($\boldsymbol{x}_{k+1} \neq \boldsymbol{x}_k$),

$$\|r(\boldsymbol{x}_k)\|^2 - \|r(\boldsymbol{x}_{k+1})\|^2 \geq \mathcal{O}\left(\frac{\|\boldsymbol{J}(\boldsymbol{x}_k)^{\mathrm{T}}\boldsymbol{r}(\boldsymbol{x}_k)\|^2}{\gamma_k}\right).$$

Regularization parameter

- If γ_k large enough, the iteration is successful.
- $\gamma_k \leq \gamma_{\max}$ for all k.

Main arguments

Decrease guarantee

For any successful iteration ($\boldsymbol{x}_{k+1} \neq \boldsymbol{x}_k$),

$$\|r(\boldsymbol{x}_k)\|^2 - \|r(\boldsymbol{x}_{k+1})\|^2 \geq \mathcal{O}\left(\frac{\|\boldsymbol{J}(\boldsymbol{x}_k)^{\mathrm{T}}\boldsymbol{r}(\boldsymbol{x}_k)\|^2}{\gamma_k}\right).$$

Regularization parameter

- If γ_k large enough, the iteration is successful.
- $\gamma_k \leq \gamma_{\max}$ for all k.

Complexity of standard Levenberg-Marquardt

The method reaches $\boldsymbol{x}_k$ such that $\|\boldsymbol{J}(\boldsymbol{x}_k)^{\mathrm{T}}\boldsymbol{r}(\boldsymbol{x}_k)\| \leq \epsilon_g$ in at most $\mathcal{O}(\epsilon_g^{-2})$ iterations.

Outline

An alternate complexity measure

Our problem: $\min_{\boldsymbol{x}\in\mathbb{R}^n} \frac{1}{2}\|\boldsymbol{r}(\boldsymbol{x})\|^2$

- Used $\|\boldsymbol{J}(\boldsymbol{x})^{\mathrm{T}}\boldsymbol{r}(\boldsymbol{x})\|$ as a complexity metric;
- Oblivious to the least-square structure;
- May want to stop when residuals are small.

An alternate complexity measure

Our problem: $\min_{x \in \mathbb{R}^n} \frac{1}{2}\|r(x)\|^2$

- Used $\|J(x)^T r(x)\|$ as a complexity metric;
- Oblivious to the least-square structure;
- May want to stop when residuals are small.

Scaled gradient (Cartis, Gould, Toint '13; Gould, Rees, Scott '19)

$$g(x) := \begin{cases} \frac{J(x)^T r(x)}{\|r(x)\|} & \text{if } \|r(x)\| > 0 \\ 0 & \text{otherwise.} \end{cases}$$

- Stopping criterion for complexity:

$$\|r(x)\| \leq \epsilon_r \quad \text{or} \quad \|g(x)\| \leq \epsilon_g.$$

An alternate complexity measure ('ed)

Goal: Find $\boldsymbol{x}_k$ such that

$$\|\boldsymbol{r}(\boldsymbol{x}_k)\| \leq \epsilon_r \quad \text{or} \quad \|\boldsymbol{g}(\boldsymbol{x}_k)\| \leq \epsilon_g, \quad \boldsymbol{g}(\boldsymbol{x}_k) := \begin{cases} \frac{\boldsymbol{J}(\boldsymbol{x}_k)^{\mathrm{T}} \boldsymbol{r}(\boldsymbol{x}_k)}{\|\boldsymbol{r}(\boldsymbol{x}_k)\|} & \text{if } \|\boldsymbol{r}(\boldsymbol{x}_k)\| > 0 \\ 0 & \text{otherwise.} \end{cases}$$

An alternate complexity measure ('ed)

Goal: Find $\boldsymbol{x}_k$ such that

$$\|\boldsymbol{r}(\boldsymbol{x}_k)\| \leq \epsilon_r \quad \text{or} \quad \|\boldsymbol{g}(\boldsymbol{x}_k)\| \leq \epsilon_g, \quad \boldsymbol{g}(\boldsymbol{x}_k) := \begin{cases} \frac{\boldsymbol{J}(\boldsymbol{x}_k)^{\mathrm{T}}\boldsymbol{r}(\boldsymbol{x}_k)}{\|\boldsymbol{r}(\boldsymbol{x}_k)\|} & \text{if } \|\boldsymbol{r}(\boldsymbol{x}_k)\| > 0 \\ 0 & \text{otherwise.} \end{cases}$$

Complexity of Levenberg-Marquardt (Gould, Rees, Scott '19)

For any $i \in \mathbb{N} \cup \{-1\}$, the method finds a suitable $\boldsymbol{x}_k$ in at most

$$\mathcal{O}(2^i \epsilon_g^{-2} \epsilon_r^{-1/2^i}) \quad \text{iterations.}$$

An alternate complexity measure ('ed)

Goal: Find $\boldsymbol{x}_k$ such that

$$\|\boldsymbol{r}(\boldsymbol{x}_k)\| \leq \epsilon_r \quad \text{or} \quad \|\boldsymbol{g}(\boldsymbol{x}_k)\| \leq \epsilon_g, \quad \boldsymbol{g}(\boldsymbol{x}_k) := \begin{cases} \frac{\boldsymbol{J}(\boldsymbol{x}_k)^{\mathrm{T}}\boldsymbol{r}(\boldsymbol{x}_k)}{\|\boldsymbol{r}(\boldsymbol{x}_k)\|} & \text{if } \|\boldsymbol{r}(\boldsymbol{x}_k)\| > 0 \\ 0 & \text{otherwise.} \end{cases}$$

Complexity of Levenberg-Marquardt (Gould, Rees, Scott '19)

For any $i \in \mathbb{N} \cup \{-1\}$, the method finds a suitable $\boldsymbol{x}_k$ in at most

$$\mathcal{O}(2^i \epsilon_g^{-2} \epsilon_r^{-1/2^i}) \quad \text{iterations.}$$

- Part of more results on high-order regularization methods.
- Asymptotically: $\epsilon_r^{-1/2^i} \to 1$ but $2^i \to \infty$.

Alternate metric (Bergou, Diouane, Kungurstev, R '22)

New scaled gradient

Given $\mathtt{i} \in \mathbb{N} \cup \{-1\}$,

$$\boldsymbol{g}^{\mathtt{i}}(\boldsymbol{x}) := \begin{cases} \frac{\|\boldsymbol{J}(\boldsymbol{x})^T \boldsymbol{r}(\boldsymbol{x})\|}{\|\boldsymbol{r}(\boldsymbol{x})\|^{2-2^{-\mathtt{i}}}} & \text{if } \|\boldsymbol{r}(\boldsymbol{x})\| \neq 0, \\ 0 & \text{otherwise.} \end{cases}$$

- Stopping criterion for complexity:

$$\|\boldsymbol{r}(\boldsymbol{x})\| \leq \epsilon_r \quad \text{or} \quad \|\boldsymbol{g}^{\mathtt{i}}(\boldsymbol{x})\| \leq \epsilon_g.$$

Alternate metric (Bergou, Diouane, Kungurstev, R '22)

New scaled gradient

Given $\mathtt{i} \in \mathbb{N} \cup \{-1\}$,

$$\boldsymbol{g}^{\mathtt{i}}(\boldsymbol{x}) := \begin{cases} \frac{\|\boldsymbol{J}(\boldsymbol{x})^T \boldsymbol{r}(\boldsymbol{x})\|}{\|\boldsymbol{r}(\boldsymbol{x})\|^{2-2^{-\mathtt{i}}}} & \text{if } \|\boldsymbol{r}(\boldsymbol{x})\| \neq 0, \\ 0 & \text{otherwise.} \end{cases}$$

- Stopping criterion for complexity:

$$\|\boldsymbol{r}(\boldsymbol{x})\| \leq \epsilon_r \quad \text{or} \quad \|\boldsymbol{g}^{\mathtt{i}}(\boldsymbol{x})\| \leq \epsilon_g.$$

- $\mathtt{i} = -1$: Classical gradient;
- $\mathtt{i} = 0$: CGT scaled gradient;
- $\mathtt{i} \to \infty$: $\|\boldsymbol{g}^{\mathtt{i}}(\boldsymbol{x})\| > \epsilon_g$ akin to gradient dominance.

Decrease guarantees

For any successful iteration ($\boldsymbol{x}_{k+1} \neq \boldsymbol{x}_k$),

$$\|\boldsymbol{r}(\boldsymbol{x}_k)\|^2 - \|\boldsymbol{r}(\boldsymbol{x}_{k+1})\|^2 \geq \mathcal{O}\left(\frac{\|\boldsymbol{J}(\boldsymbol{x}_k)^{\mathrm{T}}\boldsymbol{r}(\boldsymbol{x}_k)\|^2}{\gamma_k}\right)$$

and (if $\|\boldsymbol{r}(\boldsymbol{x}_k)\| \neq 0$)

$$\|\boldsymbol{r}(\boldsymbol{x}_k)\|^{\frac{1}{2^{\mathrm{i}}}} - \|\boldsymbol{r}(\boldsymbol{x}_{k+1})\|^{\frac{1}{2^{\mathrm{i}}}} \geq \mathcal{O}\left(\frac{\|\boldsymbol{g}^{\mathrm{i}}(\boldsymbol{x}_k)\|^2\|\boldsymbol{r}(\boldsymbol{x}_k)\|^{(4-2^{1-\mathrm{i}})}}{\gamma_k}\right).$$

Main arguments

Decrease guarantees

For any successful iteration ($\boldsymbol{x}_{k+1} \neq \boldsymbol{x}_k$),

$$\|\boldsymbol{r}(\boldsymbol{x}_k)\|^2 - \|\boldsymbol{r}(\boldsymbol{x}_{k+1})\|^2 \geq \mathcal{O}\left(\frac{\|\boldsymbol{J}(\boldsymbol{x}_k)^{\mathrm{T}}\boldsymbol{r}(\boldsymbol{x}_k)\|^2}{\gamma_k}\right)$$

and (if $\|\boldsymbol{r}(\boldsymbol{x}_k)\| \neq 0$)

$$\|\boldsymbol{r}(\boldsymbol{x}_k)\|^{\frac{1}{2^{\mathrm{i}}}} - \|\boldsymbol{r}(\boldsymbol{x}_{k+1})\|^{\frac{1}{2^{\mathrm{i}}}} \geq \mathcal{O}\left(\frac{\|\boldsymbol{g}^{\mathrm{i}}(\boldsymbol{x}_k)\|^2\|\boldsymbol{r}(\boldsymbol{x}_k)\|^{(4-2^{1-\mathrm{i}})}}{\gamma_k}\right).$$

Regularization parameter

- If γ_k large enough, the iteration is successful.
- $\gamma_k \leq \gamma_{\max}$ for all k.

Complexity table

Goal: Find $\boldsymbol{x}_k$ such that

$$\|\boldsymbol{r}(\boldsymbol{x}_k)\| \leq \epsilon_r \quad \text{or} \quad \|\boldsymbol{g}^{\mathrm{i}}(\boldsymbol{x}_k)\| \leq \epsilon_g.$$

Complexity results (BDKR '22)

i	Arbitrary	$\mathrm{i} = -1$	$\mathrm{i} = 0$
$g^{\mathrm{i}}(\boldsymbol{x})$	$\frac{\|\boldsymbol{J}(\boldsymbol{x})^{\mathrm{T}}\boldsymbol{r}(\boldsymbol{x})\|}{\|\boldsymbol{r}(\boldsymbol{x})\|^{2-2^{-\mathrm{i}}}}$	$\|\boldsymbol{J}(\boldsymbol{x})^{\mathrm{T}}\boldsymbol{r}(\boldsymbol{x})\|$	$\frac{\|\boldsymbol{J}(\boldsymbol{x})^{\mathrm{T}}\boldsymbol{r}(\boldsymbol{x})\|}{\|\boldsymbol{r}(\boldsymbol{x})\|}$
Order	$\epsilon_g^{-2}\epsilon_r^{-(4-2^{1-\mathrm{i}})}$	ϵ_g^{-2}	$\epsilon_g^{-2}\epsilon_r^{-2}$.

- Matches existing results for $\mathrm{i} = -1$.
- For $\mathrm{i} = 0$: previous results get better bounds in terms of ϵ_r **but** with very large constants (2^i).

Outline

A stochastic problem

Stochastic nonlinear least-squares

$$\min_{\boldsymbol{x}\in\mathbb{R}^n} f(\boldsymbol{x}) = \frac{1}{2}\|\boldsymbol{r}(\boldsymbol{x})\|^2$$

- Values of $\boldsymbol{r}$ and Jacobian $\boldsymbol{J}$ only accessed through **stochastic estimates**.

Challenges

- Every evaluation is replaced by a random estimate;
- Decrease no longer guaranteed;
- Accuracy of evaluation matters.

Using inexact values

Inputs: $\boldsymbol{x}_0 \in \mathbb{R}^n, \gamma_0 \geq \gamma_{\min} > 0$.
Iteration k: Given $(\boldsymbol{x}_k, \gamma_k)$,

- Compute $\boldsymbol{r}_{m_k} \approx \boldsymbol{r}(\boldsymbol{x}_k)$, $\boldsymbol{J}_{m_k} \approx \boldsymbol{J}(\boldsymbol{x}_k)$ and $\boldsymbol{s}_k \approx \operatorname{argmin}_{\boldsymbol{s}} m_k(\boldsymbol{s}) = \frac{1}{2}\|\boldsymbol{r}_{m_k} + \boldsymbol{J}_{m_k}\boldsymbol{s}\|^2 + \frac{\gamma_k}{2}\|\boldsymbol{s}\|^2$.
- Compute $\boldsymbol{r}_k^0 \approx \boldsymbol{r}(\boldsymbol{x}_k)$ and $\boldsymbol{r}_k^s \approx \boldsymbol{r}(\boldsymbol{x}_k + \boldsymbol{s}_k)$.
- If $\frac{\frac{1}{2}\|\boldsymbol{r}_k^0\|^2 - \frac{1}{2}\|\boldsymbol{r}_k^s\|^2}{m_k(0) - m_k(s)} \geq \eta$, set $\boldsymbol{x}_{k+1} = \boldsymbol{x}_k + \boldsymbol{s}_k$ and $\gamma_{k+1} = \max\{0.5\gamma_k, \gamma_{\min}\}$;
- Otherwise, set $\boldsymbol{x}_{k+1} = \boldsymbol{x}_k$ and $\gamma_{k+1} = 2\gamma_k$.

Using inexact values

Inputs: $\boldsymbol{x}_0 \in \mathbb{R}^n, \gamma_0 \geq \gamma_{\min} > 0$.
Iteration k**:** Given $(\boldsymbol{x}_k, \gamma_k)$,

- Compute $\boldsymbol{r}_{m_k} \approx \boldsymbol{r}(\boldsymbol{x}_k)$, $\boldsymbol{J}_{m_k} \approx \boldsymbol{J}(\boldsymbol{x}_k)$ and $\boldsymbol{s}_k \approx \operatorname{argmin}_{\boldsymbol{s}} m_k(\boldsymbol{s}) = \frac{1}{2}\|\boldsymbol{r}_{m_k} + \boldsymbol{J}_{m_k}\boldsymbol{s}\|^2 + \frac{\gamma_k}{2}\|\boldsymbol{s}\|^2$.
- Compute $\boldsymbol{r}_k^0 \approx \boldsymbol{r}(\boldsymbol{x}_k)$ and $\boldsymbol{r}_k^s \approx \boldsymbol{r}(\boldsymbol{x}_k + \boldsymbol{s}_k)$.
- If $\frac{\frac{1}{2}\|\boldsymbol{r}_k^0\|^2 - \frac{1}{2}\|\boldsymbol{r}_k^s\|^2}{m_k(0) - m_k(s)} \geq \eta$, set $\boldsymbol{x}_{k+1} = \boldsymbol{x}_k + \boldsymbol{s}_k$ and $\gamma_{k+1} = \max\{0.5\gamma_k, \gamma_{\min}\}$;
- Otherwise, set $\boldsymbol{x}_{k+1} = \boldsymbol{x}_k$ and $\gamma_{k+1} = 2\gamma_k$.

Goal: Prove a complexity result for this inexact method.
Key: Use γ_k to monitor the inexactness and the convergence.

Complexity analysis in an inexact setting

Accuracy requirements (model)

For every k,

$$\|\boldsymbol{J}(\boldsymbol{x}_k)^\top \boldsymbol{r}(\boldsymbol{x}_k) - \boldsymbol{J}_{m_k}^\top \boldsymbol{r}_{m_k}\| \leq \mathcal{O}\left(\frac{1}{\gamma_k}\right)$$

and

$$\left|\tfrac{1}{2}\|\boldsymbol{r}(x_k)\|^2 - \tfrac{1}{2}\|\boldsymbol{r}_{m_k}\|^2\right| \leq \mathcal{O}\left(\frac{1}{\gamma_k}\right).$$

Accuracy requirements (evaluations)

For every k,

$$\left|\tfrac{1}{2}\|\boldsymbol{r}_k^0\|^2 - \tfrac{1}{2}\|\boldsymbol{r}(\boldsymbol{x}_k)\|^2\right| \leq \mathcal{O}\left(\frac{1}{\gamma_k^2}\right)$$

and

$$\left|\tfrac{1}{2}\|\boldsymbol{r}_k^s\|^2 - \tfrac{1}{2}\|\boldsymbol{r}(\boldsymbol{x}_k + \boldsymbol{s}_k)\|^2\right| \leq \mathcal{O}\left(\frac{1}{\gamma_k^2}\right).$$

Complexity analysis

Using inexactness

- With the same theory as in the exact case, get $\mathcal{O}(\epsilon^{-3})$ instead of $\mathcal{O}(\epsilon^{-2})$ to obtain $\|\boldsymbol{J}(\boldsymbol{x}_k)^{\mathrm{T}}\boldsymbol{r}(\boldsymbol{x}_k)\| \leq \epsilon$!

Complexity analysis

Using inexactness

- With the same theory as in the exact case, get $\mathcal{O}(\epsilon^{-3})$ instead of $\mathcal{O}(\epsilon^{-2})$ to obtain $\|\boldsymbol{J}(\boldsymbol{x}_k)^{\mathrm{T}}\boldsymbol{r}(\boldsymbol{x}_k)\| \leq \epsilon$!
- Arguments:
  - Still decrease in $\mathcal{O}\left(\frac{\|\boldsymbol{J}(\boldsymbol{x}_k)^{\mathrm{T}}\boldsymbol{r}(\boldsymbol{x}_k)\|^2}{\gamma_k}\right)$;
  - But now γ_k grows as $\mathcal{O}(\epsilon^{-1})$!

Complexity analysis

Using inexactness

- With the same theory as in the exact case, get $\mathcal{O}(\epsilon^{-3})$ instead of $\mathcal{O}(\epsilon^{-2})$ to obtain $\|\boldsymbol{J}(\boldsymbol{x}_k)^{\mathrm{T}}\boldsymbol{r}(\boldsymbol{x}_k)\| \leq \epsilon$!
- Arguments:
  - Still decrease in $\mathcal{O}\left(\frac{\|\boldsymbol{J}(\boldsymbol{x}_k)^{\mathrm{T}}\boldsymbol{r}(\boldsymbol{x}_k)\|^2}{\gamma_k}\right)$;
  - But now γ_k grows as $\mathcal{O}(\epsilon^{-1})$!

A fix (BDKR '22)

- The analysis reveals $\gamma = \mathcal{O}(\|J(\boldsymbol{x})^{\top} r(\boldsymbol{x})\|/\|\boldsymbol{s}\|)$;
- By analogy with trust-region, we want $\gamma = 1/\|\boldsymbol{s}\|$;
- A **scaling** will help us achieve that.

A corrected Levenberg-Marquardt method

Inputs: $\boldsymbol{x}_0 \in \mathbb{R}^n, \gamma_0 \geq \gamma_{\min} > 0$.
Iteration k: Given $(\boldsymbol{x}_k, \gamma_k)$,

- Compute $\boldsymbol{r}_{m_k} \approx \boldsymbol{r}(\boldsymbol{x}_k)$, $\boldsymbol{J}_{m_k} \approx \boldsymbol{J}(\boldsymbol{x}_k)$ and $\boldsymbol{s}_k \approx \operatorname{argmin}_{\boldsymbol{s}} m_k(\boldsymbol{s}) = \frac{1}{2}\|\boldsymbol{r}_{m_k} + \boldsymbol{J}_{m_k}\boldsymbol{s}\|^2 + \frac{\gamma_k \|\boldsymbol{J}_{m_k}^\top \boldsymbol{r}_{m_k}\|}{2}\|\boldsymbol{s}\|^2$.
- Compute $\boldsymbol{r}_k^0 \approx \boldsymbol{r}(\boldsymbol{x}_k)$ and $\boldsymbol{r}_k^s \approx \boldsymbol{r}(\boldsymbol{x}_k + \boldsymbol{s}_k)$.
- If $\frac{\frac{1}{2}\|\boldsymbol{r}_k^0\|^2 - \frac{1}{2}\|\boldsymbol{r}_k^s\|^2}{m_k(0) - m_k(s)} \geq \eta$, set $\boldsymbol{x}_{k+1} = \boldsymbol{x}_k + \boldsymbol{s}_k$ and $\gamma_{k+1} = \max\{0.5\gamma_k, \gamma_{\min}\}$;
- Otherwise, set $\boldsymbol{x}_{k+1} = \boldsymbol{x}_k$ and $\gamma_{k+1} = 2\gamma_k$.

A corrected Levenberg-Marquardt method

Inputs: $\boldsymbol{x}_0 \in \mathbb{R}^n, \gamma_0 \geq \gamma_{\min} > 0$.
Iteration k: Given $(\boldsymbol{x}_k, \gamma_k)$,

- Compute $\boldsymbol{r}_{m_k} \approx \boldsymbol{r}(\boldsymbol{x}_k)$, $\boldsymbol{J}_{m_k} \approx \boldsymbol{J}(\boldsymbol{x}_k)$ and $\boldsymbol{s}_k \approx \operatorname{argmin}_{\boldsymbol{s}} m_k(\boldsymbol{s}) = \frac{1}{2}\|\boldsymbol{r}_{m_k} + \boldsymbol{J}_{m_k}\boldsymbol{s}\|^2 + \frac{\gamma_k \|\boldsymbol{J}_{m_k}^\top \boldsymbol{r}_{m_k}\|}{2}\|\boldsymbol{s}\|^2$.
- Compute $\boldsymbol{r}_k^0 \approx \boldsymbol{r}(\boldsymbol{x}_k)$ and $\boldsymbol{r}_k^s \approx \boldsymbol{r}(\boldsymbol{x}_k + \boldsymbol{s}_k)$.
- If $\frac{\frac{1}{2}\|\boldsymbol{r}_k^0\|^2 - \frac{1}{2}\|\boldsymbol{r}_k^s\|^2}{m_k(0) - m_k(s)} \geq \eta$, set $\boldsymbol{x}_{k+1} = \boldsymbol{x}_k + \boldsymbol{s}_k$ and $\gamma_{k+1} = \max\{0.5\gamma_k, \gamma_{\min}\}$;
- Otherwise, set $\boldsymbol{x}_{k+1} = \boldsymbol{x}_k$ and $\gamma_{k+1} = 2\gamma_k$.

- **Two sources of inexactness** (models/estimates);
- Analysis can be deterministic or **probabilistic**.

Accuracy property

For any realization, $(\boldsymbol{J}_{m_k}, \boldsymbol{r}_{m_k})$ is called accurate if

$$\|\boldsymbol{J}(\boldsymbol{x}_k)^\top \boldsymbol{r}(\boldsymbol{x}_k) - \boldsymbol{J}_{m_k}^\top \boldsymbol{r}_{m_k}\| \leq \mathcal{O}\left(\frac{1}{\gamma_k}\right) \quad \left|\tfrac{1}{2}\|\boldsymbol{r}(\boldsymbol{x}_k)\|^2 - \tfrac{1}{2}\|\boldsymbol{r}_{m_k}\|^2\right| \leq \mathcal{O}\left(\frac{1}{\gamma_k^2}\right).$$

Probabilistic accuracy property

The **random** model sequence $\{(\boldsymbol{J}_{m_k}, \boldsymbol{r}_{m_k})\}$ is called p-accurate if

$$\forall k, \quad \mathbb{P}\left((\boldsymbol{J}_{m_k}, \boldsymbol{r}_{m_k}) \text{ accurate } | \mathcal{F}_{k-1}\right) \geq p.$$

- $\mathcal{F}_{k-1} = \sigma(m_0, \ldots, m_{k-1}, \boldsymbol{r}_0^0, \boldsymbol{r}_0^s, \ldots, \boldsymbol{r}_{k-1}^0, \boldsymbol{r}_{k-1}^s)$ represents the history of the algorithm up to iteration k.

Probabilistic function estimates

Accurate function estimates

$$\begin{aligned} \left| \tfrac{1}{2}\|\boldsymbol{r}_k^0\|^2 - \tfrac{1}{2}\|\boldsymbol{r}(\boldsymbol{x}_k)\|^2 \right| &\leq \mathcal{O}\left(\tfrac{1}{\gamma_k^2}\right) \\ \left| \tfrac{1}{2}\|\boldsymbol{r}_k^s\|^2 - \tfrac{1}{2}\|\boldsymbol{r}(\boldsymbol{x}_k + \boldsymbol{s}_k)\|^2 \right| &\leq \mathcal{O}\left(\tfrac{1}{\gamma_k^2}\right). \end{aligned}$$

Probabilistically accurate estimates

The random estimate sequence $\{(\boldsymbol{r}_k^0, \boldsymbol{r}_k^s)\}$ is q-accurate if

$$\forall k, \quad \mathbb{P}\left((\boldsymbol{r}_k^0, \boldsymbol{r}_k^1) \text{ accurate } \middle| \mathcal{F}_{k-1/2}\right) \geq q.$$

- $\mathcal{F}_{k-1/2} = \sigma(m_0, \ldots, m_{k-1}, m_k, \boldsymbol{r}_0^0, \boldsymbol{r}_0^s, \ldots, \boldsymbol{r}_{k-1}^0, \boldsymbol{r}_{k-1}^s)$ represents the iteration of the algorithm up to the computation of $\boldsymbol{r}_k^0$ and $\boldsymbol{r}_k^s$.

Probabilistic complexity results

Goal: Bound the **stopping time**

$$K_\epsilon = \min\{k \ | \|\boldsymbol{r}(\boldsymbol{x})\| \leq \epsilon_r \quad \text{or} \quad \|\boldsymbol{g}^{\mathrm{i}}(\boldsymbol{x})\| \leq \epsilon_g\}.$$

Probabilistic complexity results

Goal: Bound the **stopping time**

$$K_\epsilon = \min\{k \ | \|\boldsymbol{r}(\boldsymbol{x})\| \leq \epsilon_r \quad \text{or} \quad \|\boldsymbol{g}^{\mathrm{i}}(\boldsymbol{x})\| \leq \epsilon_g\}.$$

Theorem (BDKR '22)

If $\{(\boldsymbol{J}_{m_k}, \boldsymbol{r}_{m_k})\}$ are p-accurate and $\{(\boldsymbol{r}_k^0, \boldsymbol{r}_k^s)\}$ are q-accurate, then

$$\mathbb{E}[K_\epsilon] \ \leq \ \mathcal{O}\left(\frac{pq}{pq - 1/2}\, \epsilon_g^{-2} \epsilon_r^{-(4-2^{1-\mathrm{i}})}\right).$$

Outline

Motivation: Learning ODE parameters

Problem

- **Data:** $\{\boldsymbol{z}(t_i)\}_{i=0}^m$ obtained from the solution $z(t)$ of an ODE

$$\frac{d\boldsymbol{z}}{dt}(t) = \phi_{\boldsymbol{A}}(\boldsymbol{z}(t)).$$

  with $\boldsymbol{z}(0) = \boldsymbol{z}_0$.
- Goal: Learn the parameters $\boldsymbol{A}$ of the dynamics ϕ.

Motivation: Learning ODE parameters

Problem

- **Data:** $\{\boldsymbol{z}(t_i)\}_{i=0}^{m}$ obtained from the solution $z(t)$ of an ODE

$$\frac{d\boldsymbol{z}}{dt}(t) = \phi_{\boldsymbol{A}}(\boldsymbol{z}(t)).$$

  with $\boldsymbol{z}(0) = \boldsymbol{z}_0$.
- Goal: Learn the parameters $\boldsymbol{A}$ of the dynamics ϕ.

Model: NeuralODE

- A neural network defined as the solution of an ODE: $\boldsymbol{z} \mapsto \boldsymbol{y}(1)$, where $\boldsymbol{y}$ solution of

$$\frac{d\boldsymbol{y}}{dt}(t) = \phi_{\boldsymbol{X}}(\boldsymbol{y}(t))$$

  with $\boldsymbol{y}(0) = \boldsymbol{y}_0$.
- Goal: Learn $\boldsymbol{X}$ close to $\boldsymbol{A}$.

Illustration: Linear ODE

Problem

- Noisy data $\{\boldsymbol{z}_i\}_{i=0}^m$ generated by a linear ODE $\frac{d\boldsymbol{z}}{dt}(t) = \boldsymbol{A}\boldsymbol{z}(t)$;
- Closed-form expression: $\boldsymbol{z}(t) = e^{\boldsymbol{A}t}\boldsymbol{z}(0)$.

Training problem

$$\underset{\boldsymbol{X}\in\mathbb{R}^{n\times n}}{\text{minimize}} \frac{1}{m}\sum_{i=1}^{m}\left\|\left(\boldsymbol{I}+\frac{\boldsymbol{X}}{\ell}\right)^{\ell}\boldsymbol{z}_i - \boldsymbol{z}_{i+1}\right\|^2.$$

- Euler's formula:

$$e^{\boldsymbol{X}} \approx \left(\boldsymbol{I}+\frac{\boldsymbol{X}}{\ell}\right)^{\ell}, \ell \geq 1.$$

- Nonconvex nonlinear least squares for $\ell \geq 2$ (strict, even high-order saddle points).

Applying Levenberg-Marquardt

Setup

- 100 trajectories on a spiral (2-dimensional linear ODE)
- Comparison: Levenberg-Marquardt with two complexity metrics as stopping criteria.

Applying Levenberg-Marquardt

Setup

- 100 trajectories on a spiral (2-dimensional linear ODE)
- Comparison: Levenberg-Marquardt with two complexity metrics as stopping criteria.

Criterion	Best error in $\boldsymbol{X}_*$
$\|\boldsymbol{J}(\boldsymbol{x}_k)^{\mathrm{T}}\boldsymbol{r}(\boldsymbol{x}_k)\| \leq 10^{-3}$	49
$\frac{\|\boldsymbol{J}(\boldsymbol{x}_k)^{\mathrm{T}}\boldsymbol{r}(\boldsymbol{x}_k)\|}{\|\boldsymbol{r}(\boldsymbol{x}_k)\|} \leq 10^{-3}$ or $\|\boldsymbol{r}(\boldsymbol{x}_k)\| \leq 10^{-6}$	56.

Summary

Complexity and nonlinear least squares

- A family of **complexity metrics** and results.
- Derived for Gauss-Newton methods (Levenberg-Marquardt type).
- Works with inexact/stochastic values and derivatives.

E. Bergou, Y. Diouane, V. Kungurstev and C. W. Royer. A stochastic Levenberg-Marquardt method using random models with complexity results. SIAM/ASA JUQ, 2022.

Summary

Complexity and nonlinear least squares

- A family of **complexity metrics** and results.
- Derived for Gauss-Newton methods (Levenberg-Marquardt type).
- Works with inexact/stochastic values and derivatives.

E. Bergou, Y. Diouane, V. Kungurstev and C. W. Royer. A stochastic Levenberg-Marquardt method using random models with complexity results. SIAM/ASA JUQ, 2022.

Next

- Gauss-Newton vs Newton steps?
- Application to NeuralODE training.
- Go beyond least squares (cross-entropy loss).

A. Allauzen, I. S. Legheraba and C. W. Royer. Optimization landscape of linear neural ordinary differential equations, in preparation.

The end

Thank you, and happy birthday Steve!

Harrison Ford	Steve Wright
Part of the Star Wars saga	Part of the IPM saga
Plays a professor/adventurer with a hat and a whip	Is a professor and from Australia