Stochastic Levenberg-Marquardt Methods for Noisy Derivative-Free Optimization with Complexity Results and Application to Data Assimilation

Clément Royer *(Université Paris Dauphine-PSL)*

Joint work with Elhoucine Bergou, Youssef Diouane, Vyacheslav Kungurtsev

SIAM CSE, March 4, 2021

From CSE 2019 to CSE 2021

Then

- Stochastic Levenberg-Marquardt;
- Stochastic estimates+probabilistic models.

Now

- A new version, with more complexity results;
  A stochastic Levenberg-Marquardt method using random models with complexity results and application to data assimilation.
- **Another work** on Levenberg-Marquardt methods with constraints (but derivatives...).
  A nonmonotone matrix-free algorithm for nonlinear equality-constrained least-squares problems.

Outline

Motivation: data assimilation problem

Strongly constrained 4DVAR formulation

$$\min_{\boldsymbol{z}_0,\boldsymbol{z}_1,\ldots,\boldsymbol{z}_T\in\mathbb{R}^n} \quad \overbrace{\frac{1}{2}\|\boldsymbol{z}_0-\boldsymbol{z}_b\|^2_{\boldsymbol{B}^{-1}}}^{\textit{Background state}} + \overbrace{\frac{1}{2}\sum_{i=0}^{T}\|\boldsymbol{y}_i-\mathcal{H}(\boldsymbol{z}_i)\|^2_{\boldsymbol{R}_i^{-1}}}^{\textit{Measurement fit}}$$

$$\text{s.t.} \quad \underbrace{\boldsymbol{z}_i=\mathcal{M}_i(\boldsymbol{z}_{i-1}),\ i=1,\ldots,T.}_{\textit{Governing equations}}$$

Motivation: data assimilation problem

Strongly constrained 4DVAR formulation

$$\min_{\boldsymbol{z}_0,\boldsymbol{z}_1,\ldots,\boldsymbol{z}_T\in\mathbb{R}^n} \quad \overbrace{\frac{1}{2}\|\boldsymbol{z}_0-\boldsymbol{z}_b\|^2_{\boldsymbol{B}^{-1}}}^{\textit{Background state}} + \overbrace{\frac{1}{2}\sum_{i=0}^{T}\|\boldsymbol{y}_i-\mathcal{H}(\boldsymbol{z}_i)\|^2_{\boldsymbol{R}_i^{-1}}}^{\textit{Measurement fit}}$$

$$\text{s.t.} \quad \underbrace{\boldsymbol{z}_i=\mathcal{M}_i(\boldsymbol{z}_{i-1}),\ i=1,\ldots,T.}_{\textit{Governing equations}}$$

Reformulation ($\boldsymbol{x}\leftrightarrow\boldsymbol{z}_0$)

$$\min_{\boldsymbol{x}\in\mathbb{R}^n} f(\boldsymbol{x}) := \frac{1}{2}\left(\|\boldsymbol{x}-\boldsymbol{z}_b\|^2_{\boldsymbol{B}^{-1}}+\|\boldsymbol{y}-\mathscr{H}(\boldsymbol{x})\|^2_{\boldsymbol{R}^{-1}}\right).$$

- Derivatives expensive to compute;
- $\boldsymbol{B}$ unknown $\Rightarrow$ Estimated via random samples. instead!

A more general setup

Derivative-free nonlinear least-squares

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

- $\boldsymbol{r} : \mathbb{R}^n \mapsto \mathbb{R}^m$, $\boldsymbol{r} \in \mathcal{C}^{1,1}$;
- $\boldsymbol{J}(\boldsymbol{x}) := \left[\nabla r_i(\boldsymbol{x})^\top\right] \in \mathbb{R}^{m\times n}$ **not available**;
- Values of $\boldsymbol{r}$ only accessed through **noisy estimates**.

A more general setup

Derivative-free nonlinear least-squares

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

- $\boldsymbol{r} : \mathbb{R}^n \mapsto \mathbb{R}^m$, $\boldsymbol{r} \in \mathcal{C}^{1,1}$;
- $\boldsymbol{J}(\boldsymbol{x}) := \left[\nabla r_i(\boldsymbol{x})^\top\right] \in \mathbb{R}^{m\times n}$ **not available**;
- Values of $\boldsymbol{r}$ only accessed through **noisy estimates**.

Classical Levenberg-Marquardt approach

- 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.

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

Inputs: $\boldsymbol{x}_0 \in \mathbb{R}^n, \gamma_0 > 0, \eta > 0$.
Iteration j: Given $(\boldsymbol{x}_j, \gamma_j)$,

- Compute $\boldsymbol{r}_{m_j} \approx \boldsymbol{r}(\boldsymbol{x}_j)$, $\boldsymbol{J}_{m_j} \approx \boldsymbol{J}(\boldsymbol{x}_j)$ and

  $\boldsymbol{s}_j \approx \operatorname{argmin}_{\boldsymbol{s}} \; m_j(\boldsymbol{s}) := \frac{1}{2}\|\boldsymbol{r}_{m_j} + \boldsymbol{J}_{m_j}\boldsymbol{s}\|^2 + \frac{\gamma_j\|\boldsymbol{J}_{m_j}^\top \boldsymbol{r}_{m_j}\|}{2}\|\boldsymbol{s}\|^2$.
- Compute $\boldsymbol{r}_j^0 \approx \boldsymbol{r}(\boldsymbol{x}_j)$ and $\boldsymbol{r}_j^s \approx \boldsymbol{r}(\boldsymbol{x}_j + \boldsymbol{s}_j)$.
- If $\frac{\frac{1}{2}\|\boldsymbol{r}_j^0\|^2 - \frac{1}{2}\|\boldsymbol{r}_j^s\|^2}{m_j(0) - m_j(s)} \geq \eta$, set $\boldsymbol{x}_{j+1} = \boldsymbol{x}_j + \boldsymbol{s}_j$ and $\gamma_{j+1} = 0.5\gamma_j$;
- Otherwise, set $\boldsymbol{x}_{j+1} = \boldsymbol{x}_j$ and $\gamma_{j+1} = 2\gamma_j$.

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

Inputs: $\boldsymbol{x}_0 \in \mathbb{R}^n, \gamma_0 > 0, \eta > 0$.
Iteration j: Given $(\boldsymbol{x}_j, \gamma_j)$,

- Compute $\boldsymbol{r}_{m_j} \approx \boldsymbol{r}(\boldsymbol{x}_j)$, $\boldsymbol{J}_{m_j} \approx \boldsymbol{J}(\boldsymbol{x}_j)$ and
  $\boldsymbol{s}_j \approx \operatorname{argmin}_{\boldsymbol{s}} \; m_j(\boldsymbol{s}) := \frac{1}{2}\|\boldsymbol{r}_{m_j} + \boldsymbol{J}_{m_j}\boldsymbol{s}\|^2 + \frac{\gamma_j\|\boldsymbol{J}_{m_j}^\top \boldsymbol{r}_{m_j}\|}{2}\|\boldsymbol{s}\|^2$.
- Compute $\boldsymbol{r}_j^0 \approx \boldsymbol{r}(\boldsymbol{x}_j)$ and $\boldsymbol{r}_j^s \approx \boldsymbol{r}(\boldsymbol{x}_j + \boldsymbol{s}_j)$.
- If $\frac{\frac{1}{2}\|\boldsymbol{r}_j^0\|^2 - \frac{1}{2}\|\boldsymbol{r}_j^s\|^2}{m_j(0) - m_j(s)} \geq \eta$, set $\boldsymbol{x}_{j+1} = \boldsymbol{x}_j + \boldsymbol{s}_j$ and $\gamma_{j+1} = 0.5\gamma_j$;
- Otherwise, set $\boldsymbol{x}_{j+1} = \boldsymbol{x}_j$ and $\gamma_{j+1} = 2\gamma_j$.

- Goal: **Complexity results**.
- Key: **Accuracy properties** for the models/estimates.

Model and estimates quality

First-order accuracy

A model m_j is called $(\kappa_{ef}, \kappa_{eg})$-accurate if

$$\|\boldsymbol{J}(\boldsymbol{x}_j)^\top \boldsymbol{r}(\boldsymbol{x}_j) - \boldsymbol{J}_{m_j}^\top \boldsymbol{r}_{m_j}\| \leq \frac{\kappa_{eg}}{\gamma_j} \quad \text{and} \quad \left|\tfrac{1}{2}\|\boldsymbol{r}(x_j)\|^2 - \tfrac{1}{2}\|\boldsymbol{r}_{m_j}\|^2\right| \leq \frac{\kappa_{ef}}{\gamma_j^2}.$$

Accurate function estimates

Estimates $(\boldsymbol{r}_j^0, \boldsymbol{r}_j^s)$ are ε_f-accurate if

$$\left|\tfrac{1}{2}\|\boldsymbol{r}_j^0\|^2 - \tfrac{1}{2}\|\boldsymbol{r}(\boldsymbol{x}_j)\|^2\right| \leq \frac{\varepsilon_f}{\gamma_j^2} \quad \text{and} \quad \left|\tfrac{1}{2}\|\boldsymbol{r}_j^s\|^2 - \tfrac{1}{2}\|\boldsymbol{r}(\boldsymbol{x}_j + \boldsymbol{s}_j)\|^2\right| \leq \frac{\varepsilon_f}{\gamma_j^2}$$

Complexity analysis without derivatives

Goal: Bound the number of iterations to reach an ϵ_d-point $\boldsymbol{x}$ such that

$$\|\nabla f(\boldsymbol{x})\| = \|\boldsymbol{J}(\boldsymbol{x})^\top \boldsymbol{r}(\boldsymbol{x})\| \leq \epsilon_d$$

as a function of ϵ_d.

Complexity analysis without derivatives

Goal: Bound the number of iterations to reach an ϵ_d-point $\boldsymbol{x}$ such that

$$\|\nabla f(\boldsymbol{x})\| = \|\boldsymbol{J}(\boldsymbol{x})^T \boldsymbol{r}(\boldsymbol{x})\| \leq \epsilon_d$$

as a function of ϵ_d.

Theorem (BDKR '18)

Suppose that

- Every model m_j is $(\kappa_{ef}, \kappa_{eg})$-accurate.
- Every estimate pair $(\boldsymbol{r}_j^0, \boldsymbol{r}_j^s)$ is ε_f-accurate.

Then, the method reaches an ϵ_d-point in at most

$$\mathcal{O}\left(\kappa^2 \epsilon_d^{-2}\right)$$

iterations with $\kappa = \max\{\kappa_{ef}, \kappa_{eg}, \varepsilon_f\}$.

- Derivative-based case: $\mathcal{O}(\epsilon_d^{-2})$.

Considering randomness

Recall : Our function estimates are noisy.

Inputs: $\boldsymbol{x}_0 \in \mathbb{R}^n, \gamma_0 > 0$.
Iteration j**:** Given $(\boldsymbol{x}_j, \gamma_j)$,

- Compute $\boldsymbol{r}_{m_j} \approx \boldsymbol{r}(\boldsymbol{x}_j)$, $\boldsymbol{J}_{m_j} \approx \boldsymbol{J}(\boldsymbol{x}_j)$ and
  $\boldsymbol{s}_j \approx \operatorname{argmin}_{\boldsymbol{s}} m_j(\boldsymbol{s}) = \frac{1}{2}\|\boldsymbol{r}_{m_j} + \boldsymbol{J}_{m_j}\boldsymbol{s}\|^2 + \frac{\gamma_j \|\boldsymbol{J}_{m_j}^\top \boldsymbol{r}(\boldsymbol{x}_j)\|}{2}\|\boldsymbol{s}\|^2$.
- Compute $\boldsymbol{r}_j^0 \approx \boldsymbol{r}(\boldsymbol{x}_j)$ and $\boldsymbol{r}_j^s \approx \boldsymbol{r}(\boldsymbol{x}_j + \boldsymbol{s}_j)$.
- If $\frac{\frac{1}{2}\|\boldsymbol{r}_j^0\|^2 - \frac{1}{2}\|\boldsymbol{r}_j^s\|^2}{m_j(0) - m_j(s)} \geq \eta$, set $\boldsymbol{x}_{j+1} = \boldsymbol{x}_j + \boldsymbol{s}_j$ and $\gamma_{j+1} = 0.5\gamma_j$;
- Otherwise, set $\boldsymbol{x}_{j+1} = \boldsymbol{x}_j$ and $\gamma_{j+1} = 2\gamma_j$.

- **Two sources of randomness** (models/estimates);
- Accounted for via **martingale-type properties**.

Probabilistic models

Accuracy property

For any realization, the model m_j is called $(\kappa_{ef}, \kappa_{eg})$-accurate if

$$\|\boldsymbol{J}(\boldsymbol{x}_j)^\top \boldsymbol{r}(\boldsymbol{x}_j) - \boldsymbol{J}_{m_j}^\top \boldsymbol{r}_{m_j}\| \leq \frac{\kappa_{eg}}{\gamma_j}, \quad \left|\frac{1}{2}\|\boldsymbol{r}(\boldsymbol{x}_j)\|^2 - \frac{1}{2}\|\boldsymbol{r}_{m_j}\|^2\right| \leq \frac{\kappa_{ef}}{\gamma_j^2}$$

Probabilistic accuracy property

The **random** model sequence $\{m_j\}$ is called $(p, \kappa_{ef}, \kappa_{eg})$-accurate if

$$\forall j, \quad \mathbb{P}\left(m_j \ (\kappa_{ef}, \kappa_{eg})\text{-accurate} \,\middle|\, \mathcal{F}_{j-1}\right) \geq p.$$

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

Probabilistic function estimates

Accurate function estimates

Estimates $(\boldsymbol{r}_j^0, \boldsymbol{r}_j^s)$ are ε_f-accurate if

$$\left|\tfrac{1}{2}\|\boldsymbol{r}_j^0\|^2 - \tfrac{1}{2}\|\boldsymbol{r}(\boldsymbol{x}_j)\|^2\right| \leq \frac{\varepsilon_f}{\gamma_j^2} \quad \text{and} \quad \left|\tfrac{1}{2}\|\boldsymbol{r}_j^s\|^2 - \tfrac{1}{2}\|\boldsymbol{r}(\boldsymbol{x}_j + \boldsymbol{s}_j)\|^2\right| \leq \frac{\varepsilon_f}{\gamma_j^2}$$

Probabilistically accurate estimates

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

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

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

Probabilistic complexity analysis

Goal: Bound the **stopping time** $T_\epsilon = \min\{j \mid \|\boldsymbol{J}(\boldsymbol{x}_j)^\top \boldsymbol{r}(\boldsymbol{x}_j)\| \leq \epsilon_d\}$.

Theorem (BDKR '18)

If $\{m_j\}$ is $(p, \kappa_{ef}, \kappa_{eg})$-accurate and $\{(\boldsymbol{r}_j^0, \boldsymbol{r}_j^s)\}$ is (q, ε_f)-accurate, then

$$\mathbb{E}[T_\epsilon] \leq \mathcal{O}\left(\frac{pq}{pq - 1/2}\,\kappa^2 \epsilon_d^{-2}\right),$$

$\kappa = \max\{\kappa_{ef}, \kappa_{eg}, \varepsilon_f\}$.

Probabilistic accuracy: An example

Theorem (BDKR '20)

If for every iteration j, B is approximated using an ensemble of

- $n_m = \mathcal{O}\left(\max\left\{\frac{\|\boldsymbol{x}_j - \boldsymbol{z}_b\|^2_{\boldsymbol{B}^{-1}}}{\kappa_{ef}}, \frac{\|\boldsymbol{B}^{-1}(\boldsymbol{x}_j - \boldsymbol{z}_b)\|}{\kappa_{eg}}\right\} n\right)$ samples to build m_j,
- $n_f = \mathcal{O}\left(\frac{\|\boldsymbol{x}_j - \boldsymbol{z}_b\|^2_{\boldsymbol{B}^{-1}}}{\varepsilon_f} n\right)$ samples to compute $(\boldsymbol{r}_j^0, \boldsymbol{r}_j^s)$,

then there exists (p, q) such that $\{m_j\}$ is $(p, \kappa_{ef}, \kappa_{eg})$-accurate, and $\{(\boldsymbol{r}_j^0, \boldsymbol{r}_j^s)\}$ is (q, ε_f)-accurate.

- p, q depend on variance terms, likely unknown;
- $p, q \to 1$ as $n_m, n_f \to \infty$.

About the complexity results

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.

About the complexity results

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.

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

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

- Stopping criterion for complexity:

$$\|r(\boldsymbol{x})\| \leq \epsilon_p \quad \text{or} \quad \|g(\boldsymbol{x})\| \leq \epsilon_d.$$

- Complexity of LM: for any $i \in \mathbb{N} \cup \{-1\}$, $\mathcal{O}(\epsilon_d^{-2}\epsilon_p^{-1/2^{\mathrm{i}}})$.

Our metric

New scaled gradient

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

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

- Stopping criterion for complexity:

$$\|\boldsymbol{r}(\boldsymbol{x})\| \leq \epsilon_p \quad \text{or} \quad \|\boldsymbol{g}^{\mathfrak{i}}(\boldsymbol{x})\| \leq \epsilon_d.$$

- $\mathfrak{i} = -1$: Classical gradient;
- $\mathfrak{i} = 0$: CGT scaled gradient;
- $\mathfrak{i} \to \infty$: Resembles gradient dominance.

Complexity table

Goal: Find an (ϵ_p, ϵ_d)-point $\boldsymbol{x}_k$ such that

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

Complexity results (BDKR '20)

i	Arbitrary	$\mathrm{i} = -1$	$\mathrm{i} = 0$	$\mathrm{i} \to \infty$
$g^{\mathrm{i}}(\boldsymbol{x})$	$\frac{\|\boldsymbol{J}(\boldsymbol{x})^{\mathrm{T}}\boldsymbol{r}(\boldsymbol{x})\|}{\|\boldsymbol{r}(\boldsymbol{x})\|^{2-2^{-\mathrm{i}}}}$	$\|\boldsymbol{J}(\boldsymbol{x})^{\mathrm{T}}\boldsymbol{r}(\boldsymbol{x})\|$	$\frac{\|\boldsymbol{J}(\boldsymbol{x})^{\mathrm{T}}\boldsymbol{r}(\boldsymbol{x})\|}{\|\boldsymbol{r}(\boldsymbol{x})\|}$	$\frac{\|\boldsymbol{J}(\boldsymbol{x})^{\mathrm{T}}\boldsymbol{r}(\boldsymbol{x})\|}{\|\boldsymbol{r}(\boldsymbol{x})\|^2}$
Order	$\epsilon_d^{-2}\epsilon_p^{-(4+2^{1-i})}$	ϵ_d^{-2}	$\epsilon_d^{-2}\epsilon_p^{-2}$	$\epsilon_d^{-2}\epsilon_p^{-4}$.

- Probabilistic counterparts (expected stopping time);
- For the connoisseur: other analyzes get better bounds in terms of ϵ_p/ϵ_d but have exponential dependencies in 2^{i}.

Summary

Our contributions

- Redefined probabilistic property for Levenberg-Marquardt schemes (analogy with trust region);
- Complexity analysis for stochastic function estimates;
- Application in a data assimilation setting.
- A family of **complexity metrics** and results.

Derivative-free Levenberg-Marquardt

A stochastic Levenberg-Marquardt method using random models with complexity results and application to data assimilation.

E. Bergou, Y. Diouane, V. Kungurstev and C. W. Royer.

`https://arXiv.org/abs/1807.02176v2`

Perspective

A nonmonotone matrix-free algorithm for nonlinear equality-constrained least-squares problems.

E. Bergou, Y. Diouane, V. Kungurstev and C. W. Royer.

`https://arXiv.org/abs/arXiv:2006.16340v2`.

- Addressed (among others) the original 4Dvar formulation:

$$\min_{\boldsymbol{z}_0,\boldsymbol{z}_1,\ldots,\boldsymbol{z}_T\in\mathbb{R}^n} \quad \frac{1}{2}\|\boldsymbol{z}_0-\boldsymbol{z}_b\|^2_{\boldsymbol{B}^{-1}}+\frac{1}{2}\sum_{i=0}^{T}\|\boldsymbol{y}_i-\mathcal{H}(\boldsymbol{z}_i)\|^2_{\boldsymbol{R}_i^{-1}}$$

$$\text{s.t.} \qquad \boldsymbol{z}_i=\mathcal{M}_i(\boldsymbol{z}_{i-1}),\ i=1,\ldots,T.$$

- Derivative-based approach but **inexact**.

Going forward (CSE 23?)

- Combine random models with constraints;
- Practical probabilistic approaches.