A decoupled first/second-order steps technique for nonconvex optimization

Clément Royer *(University of Wisconsin-Madison)*

Co-authors: Serge Gratton, Luis Nunes Vicente

MOPTA - August 16, 2019

Outline

1. Trust-region methods and complexity
2. Decoupled method
3. Derivative-free setting

Outline

Context

$$\min_{x \in \mathbb{R}^n} f(x) \quad \text{with } f \in \mathcal{C}^2 \text{ bounded below and nonconvex.}$$

Goals: Converge towards a **second-order stationary point.**

$$\|\nabla f(x)\| = 0, \qquad \nabla^2 f(x) \succeq 0.$$

Context

$$\min_{x \in \mathbb{R}^n} f(x) \quad \text{with } f \in \mathcal{C}^2 \text{ bounded below and nonconvex.}$$

Goals: Converge towards a **second-order stationary point.**

$$\|\nabla f(x)\| = 0, \qquad \nabla^2 f(x) \succeq 0.$$

Example: Factored matrix completion

$$\min_{U \in \mathbb{R}^{n \times r}, V \in \mathbb{R}^{m \times r}} \left\| \mathcal{P}_\Omega(UV^T - M) \right\|_F^2, \quad M \in \mathbb{R}^{n \times m}, \ \Omega \subset [n] \times [m].$$

- Nonconvex in U, V;
- Global optima are second-order stationary points.

Complexity guarantees

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

First-order complexity result

Given $\varepsilon_{\mathtt{C}} \in (0,1)$, bound the worst-case cost to obtain x_K such that $\|\nabla f(x_K)\| \leq \varepsilon_{\mathtt{C}}$.

Second-order complexity result

Given $\varepsilon_{\mathtt{C}}, \varepsilon_{\mathtt{E}} \in (0,1)$, bound the worst-case cost to obtain x_K such that

$$\|\nabla f(x_K)\| \leq \varepsilon_{\mathtt{C}}, \qquad \nabla^2 f(x_K) \succeq -\varepsilon_{\mathtt{E}} I.$$

Complexity guarantees

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

First-order complexity result

Given $\varepsilon_{\mathtt{C}} \in (0,1)$, bound the worst-case cost to obtain x_K such that $\|\nabla f(x_K)\| \le \varepsilon_{\mathtt{C}}$.

Second-order complexity result

Given $\varepsilon_{\mathtt{C}}, \varepsilon_{\mathtt{E}} \in (0,1)$, bound the worst-case cost to obtain x_K such that

$$\|\nabla f(x_K)\| \le \varepsilon_{\mathtt{C}}, \qquad \nabla^2 f(x_K) \succeq -\varepsilon_{\mathtt{E}} I.$$

- Here, cost $\Leftrightarrow$ number of iterations;
- Focus: Dependencies on $\varepsilon_{\mathtt{C}}, \varepsilon_{\mathtt{E}}$.

Trust-region algorithm

Inputs: $x_0 \in \mathbb{R}^n$, $\eta \in (0,1), 0 < \delta_0 \leq \delta_{\max}$.

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

- Compute a model $m_k(x_k + s) = m_k(x_k) + g_k^\top s + \frac{1}{2} s^\top H_k s$;
- Compute a step $s_k \approx \operatorname{argmin}_{\|s\| \leq \delta_k} m_k(x_k + s)$;
- Evaluate $\rho_k = \frac{f(x_k) - f(x_k + s_k)}{m_k(x_k) - m_k(x_k + s_k)}$.
- If $\rho_k \geq \eta$, set $x_{k+1} = x_k + s_k$ and $\delta_{k+1} = \max\{2\delta_k, \delta_{\max}\}$.
- Otherwise, set $x_{k+1} = x_k$ and $\delta_{k+1} = \delta_k / 2$.

Assumptions on models $m_k(x_k + s) = m_k(x_k) + g_k^\top s + \frac{1}{2} s^\top H_k s$

- Fraction of **Cauchy decrease**:

$$m_k(x_k) - m_k(x_k + s_k) \geq \tau \|g_k\| \min \left\{ \frac{\|g_k\|}{\|H_k\|}, \delta_k \right\}.$$

First-order analysis

Assumptions on models $m_k(x_k + s) = m_k(x_k) + g_k^\top s + \frac{1}{2} s^\top H_k s$

- Fraction of **Cauchy decrease**:

$$m_k(x_k) - m_k(x_k + s_k) \geq \tau \|g_k\| \min \left\{ \frac{\|g_k\|}{\|H_k\|}, \delta_k \right\}.$$

- Close to a first-order Taylor expansion:
  - With derivatives: $g_k = \nabla f(x_k)$;
  - Without derivatives: fully linear models.

First-order analysis

Assumptions on models $m_k(x_k + s) = m_k(x_k) + g_k^\top s + \frac{1}{2} s^\top H_k s$

- Fraction of **Cauchy decrease**:

$$m_k(x_k) - m_k(x_k + s_k) \geq \tau \|g_k\| \min\left\{ \frac{\|g_k\|}{\|H_k\|}, \delta_k \right\}.$$

- Close to a first-order Taylor expansion:
  - With derivatives: $g_k = \nabla f(x_k)$;
  - Without derivatives: fully linear models.

Complexity results

To reach a point for which $\|\nabla f(x_k)\| \leq \varepsilon_C$, the method requires at most $\mathcal{O}(\varepsilon_C^{-2})$ iterations.

Second-order analysis

Assumptions on models $m_k(x_k+s) = m_k(x_k) + g_k^\top s + \frac{1}{2} s^\top H_k s$

- Fraction of Cauchy decrease + Fraction of eigendecrease:

$$m_k(x_k) - m_k(x_k+s_k) \geq \tau \max\left\{ \|g_k\| \min\left\{ \frac{\|g_k\|}{\|H_k\|}, \delta_k \right\}, [-\lambda_k]_+ \delta_k^2 \right\},$$

with $\lambda_k = \lambda_{\min}(H_k)$.

Second-order analysis

Assumptions on models $m_k(x_k + s) = m_k(x_k) + g_k^\top s + \frac{1}{2} s^\top H_k s$

- Fraction of Cauchy decrease + Fraction of eigendecrease:

$$m_k(x_k) - m_k(x_k + s_k) \geq \tau \max \left\{ \|g_k\| \min \left\{ \frac{\|g_k\|}{\|H_k\|}, \delta_k \right\}, [-\lambda_k]_+ \delta_k^2 \right\},$$

  with $\lambda_k = \lambda_{\min}(H_k)$.
- Close to a second-order Taylor expansion:
  - With derivatives: $g_k = \nabla f(x_k), H_k = \nabla^2 f(x_k)$;
  - Without derivatives: fully quadratic models.

Second-order analysis

Assumptions on models $m_k(x_k + s) = m_k(x_k) + g_k^\top s + \frac{1}{2} s^\top H_k s$

- Fraction of Cauchy decrease + Fraction of eigendecrease:

$$m_k(x_k) - m_k(x_k + s_k) \geq \tau \max\left\{ \|g_k\| \min\left\{ \frac{\|g_k\|}{\|H_k\|}, \delta_k \right\}, [-\lambda_k]_+ \delta_k^2 \right\},$$

  with $\lambda_k = \lambda_{\min}(H_k)$.
- Close to a second-order Taylor expansion:
  - With derivatives: $g_k = \nabla f(x_k), H_k = \nabla^2 f(x_k)$;
  - Without derivatives: fully quadratic models.

Complexity results

To reach a point for which $\|\nabla f(x_k)\| \leq \varepsilon_{\mathtt{C}}$ and $\nabla^2 f(x_k)) \succeq -\varepsilon_{\mathtt{E}} I$, the method requires at most $\mathcal{O}(\max\{\varepsilon_{\mathtt{C}}^{-2}\varepsilon_{\mathtt{E}}^{-1}, \varepsilon_{\mathtt{E}}^{-3}\})$ iterations.

A matter of ε

Complexity of classical trust-region methods

(Cartis, Gould, Toint '10 - '12)

- First order ($\|\nabla f(x_k)\| \le \varepsilon_{\mathtt{C}}$): $\mathcal{O}(\varepsilon_{\mathtt{C}}^{-2})$;
- Second order ($\|\nabla f(x_k)\| \le \varepsilon_{\mathtt{C}}$ and $\lambda_{\min}(\nabla^2 f(x_k)) > -\varepsilon_{\mathtt{E}}$):

$$\mathcal{O}\left(\max\{\varepsilon_{\mathtt{C}}^{-2}\varepsilon_{\mathtt{E}}^{-1}, \varepsilon_{\mathtt{E}}^{-3}\}\right).$$

A matter of ε

Complexity of classical trust-region methods
(Cartis, Gould, Toint '10-'12)

- First order ($\|\nabla f(x_k)\| \le \varepsilon_{\mathtt{C}}$): $\mathcal{O}(\varepsilon_{\mathtt{C}}^{-2})$;
- Second order ($\|\nabla f(x_k)\| \le \varepsilon_{\mathtt{C}}$ and $\lambda_{\min}(\nabla^2 f(x_k)) > -\varepsilon_{\mathtt{E}}$):

$$\mathcal{O}\left(\max\{\varepsilon_{\mathtt{C}}^{-2}\varepsilon_{\mathtt{E}}^{-1}, \varepsilon_{\mathtt{E}}^{-3}\}\right).$$

My issue

- With simple methods (steepest descent+negative curvature), can get:

$$\mathcal{O}\left(\max\{\varepsilon_{\mathtt{C}}^{-2}, \varepsilon_{\mathtt{E}}^{-3}\}\right)$$

with and without derivatives.

- What about trust region?

Outline

Behind the complexity results

Two criteria...

- $\|\nabla f(x_k)\|$:
  - Cauchy decrease;
  - First-order Taylor-like model;
- $\lambda_{\min}(\nabla f(x_k))$:
  - Eigendecrease;
  - Second-order Taylor-like model.

Behind the complexity results

Two criteria...

- $\|\nabla f(x_k)\|$:
  - Cauchy decrease;
  - First-order Taylor-like model;
- $\lambda_{\min}(\nabla f(x_k))$:
  - Eigendecrease;
  - Second-order Taylor-like model.

...but one trust region

- Second order involves a coupled measure:

$$\sigma_k = \max\{\|\nabla f(x_k)\|, -\lambda_{\min}(\nabla^2 f(x_k))\};$$

- Reason: **single step size parameter** (δ_k);
- Theory requires $\delta_k \propto \sigma_k$.

The DESTRESS algorithm

- *DEcoupled Steps in a Trust-REgionS Strategy*
- **Idea:** One trust region/model/step for each criterion.

The DESTRESS algorithm

- *DEcoupled Steps in a Trust-REgionS Strategy*
- **Idea:** One trust region/model/step for each criterion.

Inputs: $x_0 \in \mathbb{R}^n$, $\eta \in (0,1), 0 < \delta_0 \leq \delta_{\max}$.

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

1. Compute a first-order model $m_k^{\mathtt{C}}$ and a step $s_k^{\mathtt{C}}$;
2. Compute a second-order model $m_k^{\mathtt{E}}$ and a step $s_k^{\mathtt{E}}$;
3. Choose $s_k \in \operatorname{argmin}_{s \in \{s_k^{\mathtt{C}}, s_k^{\mathtt{E}}\}} \{f(x_k + s)\}$;
4. Evaluate $\rho_k = \frac{f(x_k) - f(x_k + s_k)}{\max\left\{m_k^{\mathtt{C}}(x_k) - m_k^{\mathtt{C}}(x_k + s_k^{\mathtt{C}}), m_k^{\mathtt{E}}(x_k) - m_k^{\mathtt{E}}(x_k + s_k^{\mathtt{E}})\right\}}$.
5. If $\rho_k \geq \eta$, set $x_{k+1} = x_k + s_k$ and $\delta_{k+1} = \max\{2\delta_k, \delta_{\max}\}$.
6. Otherwise, set $x_{k+1} = x_k$ and $\delta_{k+1} = \delta_k / 2$.

First-order step

Step calculation

$$s_k^{\mathrm{C}} \approx \min_{\|s\| \le \delta_k^{\mathrm{C}}} m_k^{\mathrm{C}}(x_k + s).$$

- Model $m_k^{\mathrm{C}}(x_k + s) := m_k^{\mathrm{C}}(x_k) + [g_k^{\mathrm{C}}]^\top s + \frac{1}{2} s^\top H_k^{\mathrm{C}} s$;
- **First-order trust-region radius:** $\delta_k^{\mathrm{C}} = \|g_k^{\mathrm{C}}\| \delta_k$.

Fraction of Cauchy decrease

$$\begin{aligned} m_k^{\mathrm{C}}(x_k) - m_k^{\mathrm{C}}(x_k + s_k) &\ge \tau_{\mathrm{C}} \|g_k^{\mathrm{C}}\| \min\left\{ \frac{\|g_k^{\mathrm{C}}\|}{\|H_k^{\mathrm{C}}\|}, \delta_k^{\mathrm{C}} \right\} \\ &= \tau_{\mathrm{C}} \|g_k^{\mathrm{C}}\|^2 \min\left\{ \frac{1}{\|H_k^{\mathrm{C}}\|}, \delta_k \right\}. \end{aligned}$$

Second-order step

Step calculation

$$s_k^{\mathrm{E}} \approx \min_{\|s\| \le \delta_k^{\mathrm{E}}} m_k^{\mathrm{E}}(x_k + s).$$

- Model $m_k^{\mathrm{E}}(x_k + s) := m_k^{\mathrm{E}}(x_k) + [g_k^{\mathrm{E}}]^\top s + \frac{1}{2} s^\top H_k^{\mathrm{E}} s$;
- **Second-order trust-region radius:** $\delta_k^{\mathrm{E}} = [-\lambda_k^{\mathrm{E}}]_+ \delta_k$, with $\lambda_k^{\mathrm{E}} = \lambda_{\min}(H_k^{\mathrm{E}})$.

Fraction of eigendecrease

$$m_k^{\mathrm{E}}(x_k) - m_k^{\mathrm{E}}(x_k + s_k) \ \ge \ \tau_{\mathrm{E}} [-\lambda_k^{\mathrm{E}}]_+ (\delta_k^{\mathrm{E}})^2 \ = \ \tau_{\mathrm{E}} [-\lambda_k^{\mathrm{E}}]_+^3 \delta_k^2.$$

$\Rightarrow$ *No Cauchy decrease required here.*

Goal: Reach an $(\varepsilon_{\mathtt{C}}, \varepsilon_{\mathtt{E}})$-stationary point:

$$\|\nabla f(x_k)\| < \varepsilon_{\mathtt{C}} \quad \text{and} \quad \left[-\lambda_{\min}\left(\nabla^2 f(x_k)\right)\right] < \varepsilon_{\mathtt{E}}.$$

Theorem (derivative-based)

Suppose that $g_k^{\mathtt{C}} = \nabla f(x_k)$ and $H_k^{\mathtt{E}} = \nabla^2 f(x_k)$ for all k. Then DESTRESS reaches an $(\varepsilon_{\mathtt{C}}, \varepsilon_{\mathtt{E}})$-stationary point in at most

$$\mathcal{O}\left(\max\left\{\varepsilon_{\mathtt{C}}^{-2}, \varepsilon_{\mathtt{E}}^{-3}\right\}\right) \qquad \text{iterations.}$$

- Better than $\mathcal{O}(\max\left\{\varepsilon_{\mathtt{C}}^{-2}\varepsilon_{\mathtt{E}}^{-1}, \varepsilon_{\mathtt{E}}^{-3}\right\})$;
- Reason clearly identified.

Meanwhile...

Meanwhile...

An alternate proposal (Curtis, Lubberts, Robinson, 2018)

Compute $s_k \approx \mathrm{argmin}_{s \in \Delta_k} f(x_k) + \nabla f(x_k)^\top s + \frac{1}{2} s^\top \nabla^2 f(x_k) s$, with

$$\Delta_k = \begin{cases} \delta_k \|\nabla f(x_k)\| & \text{if } \|\nabla f(x_k)\|^2 \geq [-\lambda_{\min}(\nabla^2 f(x_k))]_+^3 \\ \delta_k \left[-\lambda_{\min}(\nabla^2 f(x_k))\right]_+ & \text{otherwise.} \end{cases}$$

- Exact derivatives;
- Same iteration complexity.

Meanwhile...

An alternate proposal (Curtis, Lubberts, Robinson, 2018)

Compute $s_k \approx \mathrm{argmin}_{s \in \Delta_k} f(x_k) + \nabla f(x_k)^\top s + \frac{1}{2} s^\top \nabla^2 f(x_k) s$, with

$$\Delta_k = \begin{cases} \delta_k \|\nabla f(x_k)\| & \text{if } \|\nabla f(x_k)\|^2 \geq [-\lambda_{\min}(\nabla^2 f(x_k))]_+^3 \\ \delta_k \left[-\lambda_{\min}(\nabla^2 f(x_k))\right]_+ & \text{otherwise.} \end{cases}$$

- Exact derivatives;
- Same iteration complexity.

Revised DESTRESS algorithm

- Criterion based on model derivatives (exact or not);
- Uses only one function value per iteration.

Numerical results - Derivative Based

Performance profiles

- Criterion: # of function evaluations to satisfy

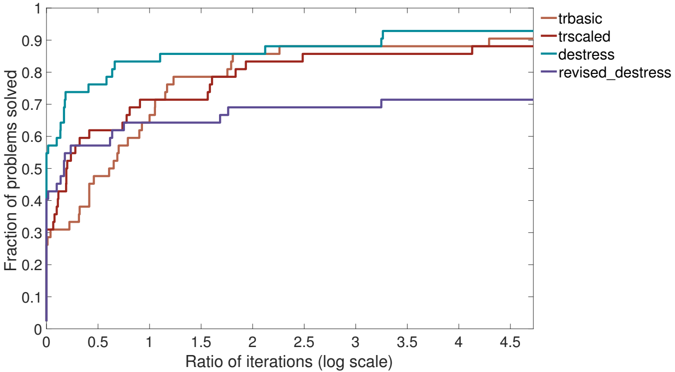
$$\|\nabla f(x_k)\| \le 10^{-4}, \quad \lambda_{\min}(\nabla^2 f(x_k)) \ge -10^{-3}.$$

- Budget: 500 iterations.

Algorithmic variants

Variant	Trust region(s)
trbasic	δ_k
trscaled	$\delta_k \max\left\{\|\nabla f(x_k)\|, [-\lambda_{\min}(\nabla^2 f(x_k))]_+\right\}$
destress	$\delta_k \|\nabla f(x_k)\|$ and $\delta_k \left[-\lambda_{\min}(\nabla^2 f(x_k))\right]_+$
revised-destress	$\delta_k \|\nabla f(x_k)\|$ xor $\delta_k \left[-\lambda_{\min}(\nabla^2 f(x_k))\right]_+$.

- Performance on 60 nonconvex problems from the CUTEst collection.



Outline

Trust-region algorithm - DFO algorithm

Inputs: $x_0 \in \mathbb{R}^n$, $\eta \in (0,1), 0 < \delta_0 \leq \delta_{\max}$.

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

- Compute a model $m_k(x_k + s) = m_k(x_k) + g_k^\top s + \frac{1}{2} s^\top H_k s$;
- Compute a step $s_k \approx \operatorname{argmin}_{\|s\| \leq \delta_k} m_k(x_k + s)$;
- Evaluate $\rho_k = \frac{f(x_k) - f(x_k + s_k)}{m_k(x_k) - m_k(x_k + s_k)}$.
- If $\rho_k \geq \eta$, set $x_{k+1} = x_k + s_k$ and $\delta_{k+1} = \max\{2\delta_k, \delta_{\max}\}$.
- Otherwise, set $x_{k+1} = x_k$ and $\delta_{k+1} = \delta_k / 2$.

Trust-region algorithm - DFO algorithm

Inputs: $x_0 \in \mathbb{R}^n$, $\eta \in (0,1), 0 < \delta_0 \leq \delta_{\max}$.

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

- Compute a model $m_k(x_k + s) = m_k(x_k) + g_k^\top s + \frac{1}{2} s^\top H_k s$;
- Compute a step $s_k \approx \operatorname{argmin}_{\|s\| \leq \delta_k} m_k(x_k + s)$;
- Evaluate $\rho_k = \frac{f(x_k) - f(x_k + s_k)}{m_k(x_k) - m_k(x_k + s_k)}$.
- If $\rho_k \geq \eta$, set $x_{k+1} = x_k + s_k$ and $\delta_{k+1} = \max\{2\delta_k, \delta_{\max}\}$.
- Otherwise, set $x_{k+1} = x_k$ and $\delta_{k+1} = \delta_k / 2$.

Derivatives of f not available

- Models based on function values (interpolation, regression);
- Fully linear for first-order results;
- Fully quadratic for second-order results.

Classical criticality step

- Adjusts δ_k to obtain a good fully linear (resp. fully quadratic) model on $B(x_k, \delta_k)$;
- Updates involve comparing δ_k and $\|g_k\|$ (or $\max\{\|g_k\|, -\lambda_k\}$)!
- Bounded number of steps, accounted for in complexity (Garmanjani, Júdice, Vicente '18).

Criticality step

Classical criticality step

- Adjusts δ_k to obtain a good fully linear (resp. fully quadratic) model on $B(x_k, \delta_k)$;
- Updates involve comparing δ_k and $\|g_k\|$ (or $\max\{\|g_k\|, -\lambda_k\}$)!
- Bounded number of steps, accounted for in complexity (Garmanjani, Júdice, Vicente '18).

Coupled criticality step ($x_k, m_k, \gamma > \beta > 0$)

1. If $\delta_k > \gamma \|g_k\|$ stop, otherwise set $\ell = 0$.
2. Compute $m_{k+\ell+1}$ fully linear on $B(x_k, \delta_{k+\ell}/2)$ and set $\delta_{k+\ell+1} = \max\{\delta_{k+\ell}/2, \beta\|g_{k+\ell+1}\|\}$.
3. Set $\ell \leftarrow \ell + 1$; if $\delta_{k+\ell} > \gamma \|g_{k+\ell}\|$, stop. Otherwise, go back to 2.

Inputs: $x_0 \in \mathbb{R}^n$, $\eta \in (0,1), 0 < \delta_0 \leq \delta_{\max}$.

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

1. Compute a first-order model $m_k^{\mathtt{C}}$ and a step $s_k^{\mathtt{C}}$;
2. Compute a second-order model $m_k^{\mathtt{E}}$ and a step $s_k^{\mathtt{E}}$;
3. Choose $s_k \in \operatorname{argmin}_{s \in \{s_k^{\mathtt{C}}, s_k^{\mathtt{E}}\}} \{f(x_k + s)\}$;
4. Evaluate $\rho_k = \frac{f(x_k) - f(x_k + s_k)}{\max\left\{m_k^{\mathtt{C}}(x_k) - m_k^{\mathtt{C}}(x_k + s_k^{\mathtt{C}}), m_k^{\mathtt{E}}(x_k) - m_k^{\mathtt{E}}(x_k + s_k^{\mathtt{E}})\right\}}$.
5. If $\rho_k \geq \eta$, set $x_{k+1} = x_k + s_k$ and $\delta_{k+1} = \max\{2\delta_k, \delta_{\max}\}$.
6. Otherwise, set $x_{k+1} = x_k$ and $\delta_{k+1} = \delta_k / 2$.

Derivatives of f not available

- Fully linear first-order model $m_k^{\mathtt{C}}$;
- Fully quadratic second-order model $m_k^{\mathtt{E}}$.

Decoupled variant

Decoupled criticality step $(x_k, \varepsilon_{\mathtt{C}}, \varepsilon_{\mathtt{E}})$

1. Generate $m_k^{\mathtt{C}}$ fully linear on $B(x_k, \delta_k \varepsilon_{\mathtt{C}})$. If $\|g_k^{\mathtt{C}}\| > \varepsilon_{\mathtt{C}}$ stop. Otherwise, perform ℓ_1 iterations to obtain $\|g_{k+\ell_1}^{\mathtt{C}}\| \le \varepsilon_{\mathtt{C}}$ or $m_{k+\ell_1}$ fully linear in $B(x_k, \frac{\delta_k}{2^{\ell_1}} \|g_{\ell_1}^{\mathtt{C}}\|)$.
2. Generate $m_k^{\mathtt{E}}$ fully quadratic on $B(x_k, \delta_k \varepsilon_{\mathtt{E}})$. If $[-\lambda_k^{\mathtt{E}}]_+ > \varepsilon_{\mathtt{E}}$ stop. Otherwise, perform ℓ_2 iterations to obtain $[-\lambda_{\ell_2}^{\mathtt{E}}]_+ \le \varepsilon_{\mathtt{E}}$ or m_{ℓ_2} fully quadratic in $B(x_k, \frac{\delta_k}{2^{\ell_2}} [-\lambda_{\ell_2}^{\mathtt{E}}]_+)$.
3. Set $\delta_k \leftarrow \delta_k / 2^{\min\{\ell_1, \ell_2\}}$, $m_k^{\mathtt{C}} \leftarrow m_{k+\ell_1}^{\mathtt{C}}$, $m_k^{\mathtt{E}} \leftarrow m_{k+\ell_2}^{\mathtt{E}}$.

Decoupled variant

Decoupled criticality step $(x_k, \varepsilon_{\mathtt{C}}, \varepsilon_{\mathtt{E}})$

1. Generate $m_k^{\mathtt{C}}$ fully linear on $B(x_k, \delta_k \varepsilon_{\mathtt{C}})$. If $\|g_k^{\mathtt{C}}\| > \varepsilon_{\mathtt{C}}$ stop. Otherwise, perform ℓ_1 iterations to obtain $\|g_{k+\ell_1}^{\mathtt{C}}\| \le \varepsilon_{\mathtt{C}}$ or $m_{k+\ell_1}$ fully linear in $B(x_k, \frac{\delta_k}{2^{\ell_1}} \|g_{\ell_1}^{\mathtt{C}}\|)$.
2. Generate $m_k^{\mathtt{E}}$ fully quadratic on $B(x_k, \delta_k \varepsilon_{\mathtt{E}})$. If $[-\lambda_k^{\mathtt{E}}]_+ > \varepsilon_{\mathtt{E}}$ stop. Otherwise, perform ℓ_2 iterations to obtain $[-\lambda_{\ell_2}^{\mathtt{E}}]_+ \le \varepsilon_{\mathtt{E}}$ or m_{ℓ_2} fully quadratic in $B(x_k, \frac{\delta_k}{2^{\ell_2}} [-\lambda_{\ell_2}^{\mathtt{E}}]_+)$.
3. Set $\delta_k \leftarrow \delta_k / 2^{\min\{\ell_1, \ell_2\}}$, $m_k^{\mathtt{C}} \leftarrow m_{k+\ell_1}^{\mathtt{C}}$, $m_k^{\mathtt{E}} \leftarrow m_{k+\ell_2}^{\mathtt{E}}$.

- Well defined, guaranteed to terminate;
- Exploits our decoupled scaling;
- Complexity bound in $\mathcal{O}\left(\max\left\{\varepsilon_{\mathtt{C}}^{-2}, \varepsilon_{\mathtt{E}}^{-3}\right\}\right)$.

Numerical results - Derivative Free

Performance profiles

- Criterion: # of function evaluations to satisfy

$$f(x_k) - f_{best} < 10^{-6}(f(x_0) - f_{best}).$$

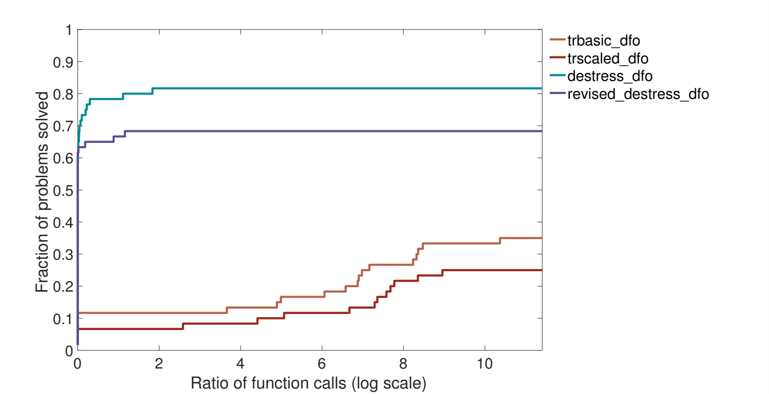
- Budget: 500 iterations.

Algorithmic variants

Variant	Trust region(s)
trbasic	δ_k
trscaled	$\delta_k \max\left\{\|\nabla m_k(x_k)\|, [-\lambda_{\min}(\nabla^2 m_k(x_k))]_+\right\}$
destress	$\delta_k\|\nabla m_k^{\mathtt{C}}(x_k)\|$ and $\delta_k\left[-\lambda_{\min}(\nabla^2 m_k^{\mathtt{E}}(x_k))\right]_+$
revised-destress	$\delta_k\|\nabla m_k^{\mathtt{C}}(x_k)\|$ xor $\delta_k\left[-\lambda_{\min}(\nabla^2 m_k^{\mathtt{E}}(x_k))\right]_+$.

Numerical experiments - Derivative Free

- Performance on 60 nonconvex problems, $2 \leq n \leq 50$.



Summary

Decoupling

- Classical second-order analysis done using a coupled measure;
- Decoupling can lead to improvement;
- In DFO, connected to criticality steps.

Going further

- Results apply to inexact derivatives too;
- Inexact function values much harder (Bellavia et al., '19);
- Better criticality steps?

For more details

- Paper
  A decoupled first/second-order steps technique for nonconvex nonlinear unconstrained optimization with improved complexity bounds. S. Gratton, C. W. Royer and L. N. Vicente, Mathematical Programming, 2018.
- Code
  `DEcoupled Steps in a Trust-REgionS Strategy (DESTRESS)`
  Available online

For more details

- Paper
  A decoupled first/second-order steps technique for nonconvex nonlinear unconstrained optimization with improved complexity bounds. S. Gratton, C. W. Royer and L. N. Vicente, Mathematical Programming, 2018.
- Code
  `DEcoupled Steps in a Trust-REgionS Strategy (DESTRESS)`
  Available online

Thank you!

`croyer2@wisc.edu`