Mesures de criticalité d'ordres 1 et 2 en recherche directe

From first to second-order criticality measures in direct search

Clément Royer
ENSEEIHT-IRIT, Toulouse, France

Co-auteurs: S. Gratton, L. N. Vicente

Journées du GDR MOA - 02/12/15

Outline

Introduction

We are interested in solving an unconstrained optimization problem:

$$\min_{x \in \mathbb{R}^n} f(x).$$

The objective function f

- f bounded from below, $\mathcal{C}^2$;
- $\nabla f, \nabla^2 f$ Lipschitz continuous;
- f nonconvex $\Rightarrow$ the Hessian matrix is not always positive semidefinite.

Caring about second order

Our definition of a second-order method

An optimization algorithm that exploits the (negative) curvature information contained in the Hessian matrix, to ensure second-order convergence.

Second-order tools for the analysis

- Taylor expansion :

$$f(x+s) - f(x) \leq \nabla f(x)^\top s + \frac{1}{2} s^\top \nabla^2 f(x)\, s + L_{\nabla^2 f} \|s\|^3,$$

- Directional derivative estimate

$$f(x+s) - 2\,f(x) + f(x-s) = s^\top \nabla^2 f(x)\, s + \mathcal{O}\left(\|s\|^3\right).$$

Second-order derivative-based optimization

- Early treatment in Trust-Region and (Curvilinear) Line Search Methods;
- Negative curvature is seldom handled to provide second-order convergence guarantees;
- Regain of interest, with the outbreak of cubic models: Curtis et al '13,'14,'15, **Wong ISMP '15**.

Main issues

- Cost of computing negative curvature directions;
- Dissociate the contributions from orders 1 and 2;
- No natural scaling between $\|\nabla f(x)\|$ and $\left|\lambda_{\min}\left(\nabla^2 f(x)\right)\right|$.

Outline

Solving the problem without using the derivatives

We consider a setting in which derivatives of f are unavailable or too expensive for computation.

Derivative-Free Optimization (DFO) methods

- Do not use the derivatives within the algorithm;
- Two main classes:
  - Model-based methods;
  - Direct-search methods.

Introduction to Derivative-Free Optimization
A.R. Conn, K. Scheinberg, L.N. Vicente. (2009)

Solving the problem without using the derivatives

We consider a setting in which derivatives of f are unavailable or too expensive for computation.

Derivative-Free Optimization (DFO) methods

- Do not use the derivatives within the algorithm;
- Two main classes:
  - Model-based methods;
  - Direct-search methods.

Introduction to Derivative-Free Optimization
A.R. Conn, K. Scheinberg, L.N. Vicente. (2009)

A simple direct-search framework

1. **Initialization** Set $x_0, \alpha_0 > 0, \theta < 1 \leq \gamma$.
   Set $k = 0$.
2. **Poll Step**
   - Choose a polling/direction set of (unitary) vectors.
   - If it exists d_k within the set such that

$$f(x_k + \alpha_k\, d_k) - f(x_k) \;<\; -\alpha_k^3,$$

     then set $x_{k+1} := x_k + \alpha_k\, d_k$ and $\alpha_{k+1} := \gamma\, \alpha_k$.
   - Otherwise, set $x_{k+1} := x_k$ and $\alpha_{k+1} := \theta\, \alpha_k$.
3. Set $k = k + 1$ and go back to the poll step.

A simple direct-search framework

1. **Initialization** Set $x_0, \alpha_0 > 0, \theta < 1 \leq \gamma$.
   Set $k = 0$.
2. **Poll Step**
   - Choose a polling/direction set of (unitary) vectors.
   - If it exists d_k within the set such that

$$f(x_k + \alpha_k\, d_k) - f(x_k) \;<\; -\alpha_k^3,$$

     then set $x_{k+1} := x_k + \alpha_k\, d_k$ and $\alpha_{k+1} := \gamma\, \alpha_k$.
   - Otherwise, set $x_{k+1} := x_k$ and $\alpha_{k+1} := \theta\, \alpha_k$.
3. Set $k = k + 1$ and go back to the poll step.

Remarks

- Performance criterion : # of evaluations of f;
- Theoretical properties mainly depend on polling choices.

Order 2 in derivative-free methods

- Few practical methods that explicitly deal with nonconvexity;
- For direct search, most results due to Abramson et al ('05,'06,'14).

Issues with the existing direct-search approaches

- Study properties of (unknown) convergent subsequences;
- Rely on density assumptions and on direction sets dependent from an iteration to another.

Our objective is to develop a method that exploits second-order properties at the iteration level.

Outline

Back to the direct-search method

1. **Initialization** Set $x_0, \alpha_0 > 0, \theta < 1 \leq \gamma$.
   Set $k = 0$.
2. **Poll Step**
   - Choose a polling set of (unitary) vectors.
   - If it exists d_k within the set such that

$$f(x_k + \alpha_k\, d_k) - f(x_k) \;<\; -\alpha_k^3,$$

     then set $x_{k+1} := x_k + \alpha_k\, d_k$ and $\alpha_{k+1} := \gamma\, \alpha_k$.
   - Otherwise, set $x_{k+1} := x_k$ and $\alpha_{k+1} := \theta\, \alpha_k$.
3. Set $k = k + 1$ and go back to the poll step.

How can we define rules to choose the polling sets ?

First-order polling quality

- Typical direct-search methods ensure first-order convergence;
- The polling sets must provide good approximations of the negative gradient.

First-order polling quality

- Typical direct-search methods ensure first-order convergence;
- The polling sets must provide good approximations of the negative gradient.

A measure of first-order quality

Let D be a set of unitary vectors and $v \in \mathbb{R}^n \setminus \{0\}$. Then

$$\mathrm{cm}(D, v) = \max_{d \in D} \frac{d^\top v}{\|v\|}$$

is called the cosine measure of D at v.

First-order polling quality

- Typical direct-search methods ensure first-order convergence;
- The polling sets must provide good approximations of the negative gradient.

A measure of first-order quality

Let D be a set of unitary vectors and $v \in \mathbb{R}^n \setminus \{0\}$. Then

$$\mathrm{cm}(D, v) = \max_{d \in D} \frac{d^\top v}{\|v\|}$$

is called the cosine measure of D at v.
If $\mathrm{cm}(D, -\nabla f(x)) > 0$, it means that D contains a descent direction of f at x.

Usual polling choice

Positive Spanning Sets (PSS)

D is a PSS if it generates $\mathbb{R}^n$ by nonnegative linear combinations.

- D is a PSS **iff** $\forall v \neq 0, \mathrm{cm}(D, v) > 0$;
- a PSS contains at least $n + 1$ vectors;
- Ex) The coordinate set $D_{\oplus} = [I \quad -I]$.

Usual polling choice

Positive Spanning Sets (PSS)

D is a PSS if it generates $\mathbb{R}^n$ by nonnegative linear combinations.

- D is a PSS **iff** $\forall v \neq 0, \mathrm{cm}(D, v) > 0$;
- a PSS contains at least $n + 1$ vectors;
- Ex) The coordinate set $D_{\oplus} = [I \quad -I]$.

PSS and first-order convergence

Two main ideas :

- Use the Taylor expansion

$$f(x + \alpha d) - f(x) \leq \alpha \nabla f(x)^{\top} d + L_{\nabla f} \alpha^2.$$

- Assume that for every iteration,

$$\mathrm{cm}(D_k, -\nabla f(x_k)) \geq \kappa, \quad \kappa \in (0, 1).$$

First-order results

First-order polling strategy

1. Poll along a Positive Spanning Set D_k.

First-order results

First-order polling strategy

1. Poll along a Positive Spanning Set D_k.

Convergence arguments

- Independently of D_k, $\alpha_k \to 0$;
- On unsuccessful iterations,

$$\alpha_k \geq \mathcal{O}\left(\kappa \|\nabla f(x_k)\|\right).$$

Theorem (First-order convergence)

$$\liminf_{k\to\infty} \|\nabla f(x_k)\| = 0.$$

A second-order criticality measure

Definition

Given a set of unitary vectors D and a symmetric matrix A, the Rayleigh measure of D with respect to A is defined by

$$\mathrm{rm}\,(D, A) \;=\; \min_{d \in V(D)} d^{\top} A\, d,$$

where

$$V(D) \;=\; \{d \in D \mid -d \in D\}$$

is the symmetric part of D.

- The Rayleigh measure is an approximation of the minimum eigenvalue;
- We want this approximation to be sufficiently good.

Rayleigh measure and negative curvature

In derivative-based methods, if $\lambda_{\min}(\nabla^2 f(x_k)) < 0$, one uses a sufficient negative curvature direction:

$$d^\top \nabla^2 f(x_k)\, d \quad \leq \quad \beta\, \lambda_{\min}(\nabla^2 f(x_k)),$$

with $\beta \in (0, 1]$.

In a direct-search environment

- Derivative-free: Hessian eigenvalues cannot be computed;
- Direct search: The step size goes to zero;

We will be ensuring

$$\mathrm{rm}\left(D_k, \nabla^2 f(x_k)\right) \quad \leq \quad \beta\, \lambda_{\min}(\nabla^2 f(x_k)) + \mathcal{O}(\alpha_k).$$

A second-order polling strategy for Direct Search

Second-order polling rules

1. Poll along a PSS D_k (First-order rule);

A second-order polling strategy for Direct Search

Second-order polling rules

1. Poll along a PSS D_k (First-order rule);
2. Poll along -D_k;
3. Select a basis $B_k \subset D_k$ and build an approximated Hessian $H_k \approx B_k^\top \nabla^2 f(x_k) B_k$, using function values;
4. Compute a unitary vector such that $H_k v_k = \lambda_{\min}(H_k) v_k$; poll along v_k and -v_k.

A second-order polling strategy for Direct Search

Second-order polling rules

1. Poll along a PSS D_k (First-order rule);
2. Poll along -D_k;
3. Select a basis $B_k \subset D_k$ and build an approximated Hessian $H_k \approx B_k^\top \nabla^2 f(x_k)\, B_k$, using function values;
4. Compute a unitary vector such that $H_k\, v_k = \lambda_{\min}(H_k)\, v_k$; poll along v_k and -v_k.

- The cost of an iteration is at most $\mathcal{O}(n^2)$ evaluations.
- The polling stops as soon as it encounters a direction d such that

$$f(x_k + \alpha_k\, d) - f(x_k) \;<\; -\alpha_k^3.$$

Outline

Second-order convergence

Assumptions

- The D_k's are PSS with $\forall k,\ \mathrm{cm}(D_k, -\nabla f(x_k)) \geq \kappa > 0$;
- It exists $\sigma \in (0, 1]$ such that

$$\forall k, \quad \sigma_{\min}(B_k)^2 \geq \sigma > 0.$$

Minimum eigenvalue estimate

Let k be an unsuccessful iteration, and P_k the corresponding polling set.

$$\mathrm{rm}\left(P_k, \nabla^2 f(x_k)\right) \ \leq\ v_k^\top \nabla^2 f(x_k)\, v_k \ \leq\ \sigma\, \lambda_{\min}(\nabla^2 f(x_k)) + \mathcal{O}\left(n\, \alpha_k\right).$$

The factors σ and n are due to the approximation error.

Second-order convergence (2)

Convergence arguments

- As before, $\alpha_k \to 0$;
- On an unsuccessful iteration k, one has:

$$\alpha_k \geq \max\left\{\mathcal{O}\left(\kappa \|\nabla f(x_k)\|\right), \mathcal{O}\left(-\sigma n^{-1} \lambda_{\min}\left(\nabla^2 f(x_k)\right)\right)\right\}.$$

Second-order convergence (2)

Convergence arguments

- As before, $\alpha_k \to 0$;
- On an unsuccessful iteration k, one has:

$$\alpha_k \;\geq\; \max\left\{\mathcal{O}\left(\kappa\,\|\nabla f(x_k)\|\right), \mathcal{O}\left(-\sigma\, n^{-1}\,\lambda_{\min}\left(\nabla^2 f(x_k)\right)\right)\right\}.$$

Theorem (Second-order convergence)

$$\liminf_{k\to\infty} \max\left\{\|\nabla f(x_k)\|, -\lambda_{\min}(\nabla^2 f(x_k))\right\} \;=\; 0.$$

Second-order worst-case complexity

We aim to reach an (ϵ_g, ϵ_H)-second-order critical point, i.e.

$$\|\nabla f(x_k)\| < \epsilon_g \quad \text{and} \quad \lambda_{\min}(\nabla^2 f(x_k)) > -\epsilon_H.$$

Theorem

Let $N_{\epsilon_{gH}}$ the number of evaluations of f needed to reach a (ϵ_g, ϵ_H)-second-order critical point; then

$$N_{\epsilon_{gH}} \leq \mathcal{O}\left(n^2 \max\left\{\kappa^{-3}\,\epsilon_g^{-3}, \sigma^{-3}\,n^3\,\epsilon_H^{-3}\right\}\right).$$

Corollary

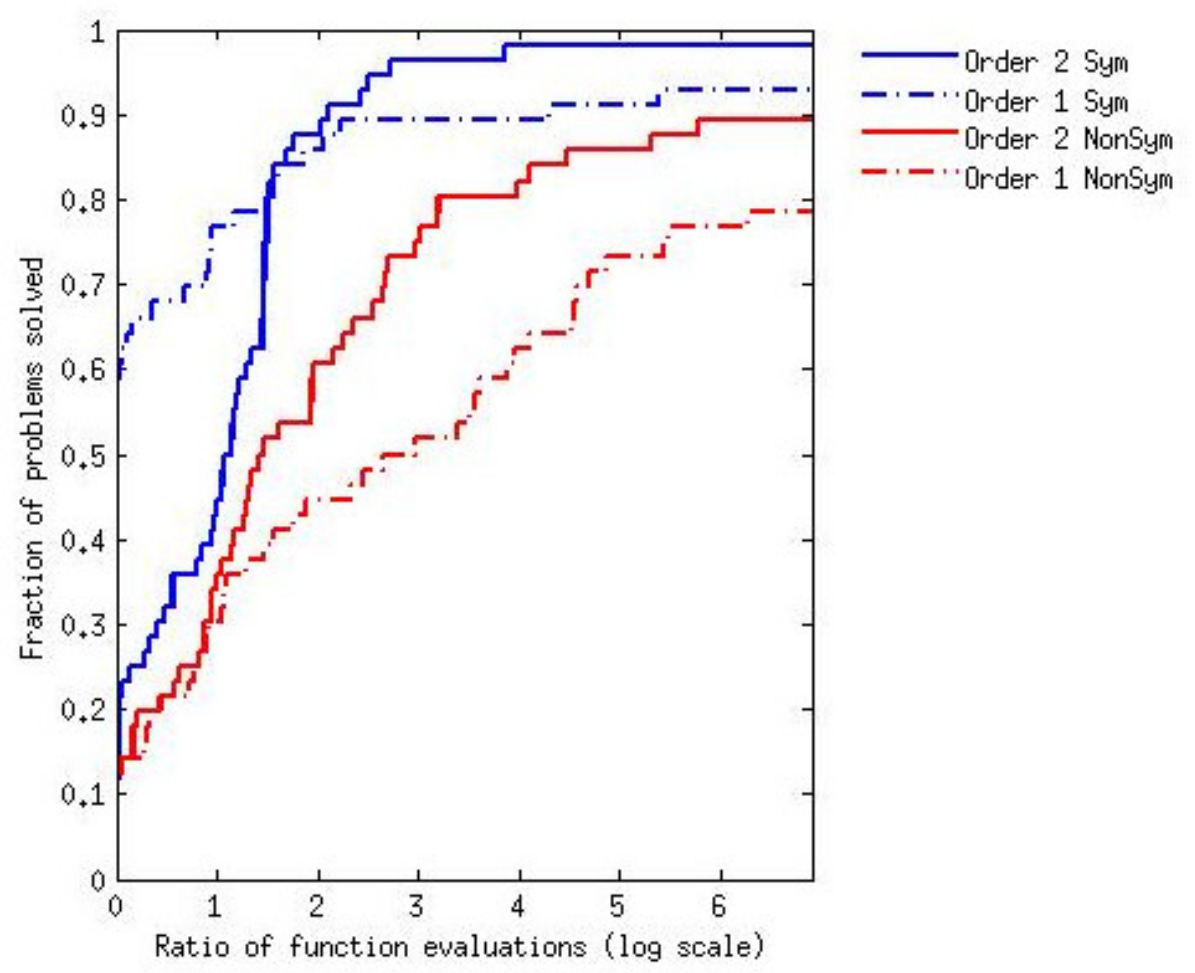
Choosing $D_k = [I \ -I]$ *yields* $\kappa = 1/\sqrt{n}, \sigma = 1$, *and the complexity bound is*

$$\mathcal{O}\left(n^5 \max\left\{\epsilon_g^{-3}, \epsilon_H^{-3}\right\}\right).$$

Practical insights

On 60 CUTEst problems with negative curvature:

- Using symmetric sets generally improves the performance;
- Second-order rules (plain lines) allow to solve more problems.



Conclusion

Our contributions

- The definition of a second-order criticality measure;
- A second-order direct-search method that converges w.r.t. this measure and its associated complexity;
- Numerical confirmation of the theoretical findings.

Conclusion

Our contributions

- The definition of a second-order criticality measure;
- A second-order direct-search method that converges w.r.t. this measure and its associated complexity;
- Numerical confirmation of the theoretical findings.

For more information

- **A second-order globally convergent direct-search method and its worst-case complexity**.
  S. Gratton, C. W. Royer, L. N. Vicente.
  To appear in *Optimization*.

Towards randomization

- Guaranteeing

$$\mathbb{P}\left(\mathrm{cm}(D_k, -\nabla f(x_k)) > \kappa\right) \geq p > 0$$

  is sufficient for first-order convergence, and we can do it in practice (Gratton, R., Vicente and Zhang '14);
- Can we do the same with second-order properties ?

Merci !

clement.royer@enseeiht.fr