Second-order convergence in direct-search methods

Clément Royer
IRIT & Université de Toulouse, France

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

CIMI Workshop on Optimization and Data Assimilation, Toulouse
January 14th, 2016

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.

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)

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:

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

We will use the Taylor expansions:

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

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

Second-order derivative-based optimization

- Early treatment in Trust-Region and Line Search Methods;
- Latest works: Curtis et al ('13,'14,'15), **Wong ISMP ('15)**,etc;
- Negative curvature not always handled to provide second-order convergence guarantees.

Main issues

- Cost of computing negative curvature directions;
- Identify 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

A simple direct-search framework

1. **Initialization** Set $x_0, \alpha_0 > 0, \theta < 1 \leq \gamma, k = 0$.
2. **Poll Step**
   - Let P_k be a polling set of (unitary) vectors.
   - If it exists $d_k \in P_k$ 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 $x_0, \alpha_0 > 0, \theta < 1 \leq \gamma, k = 0$.
2. **Poll Step**
   - Let P_k be a polling set of (unitary) vectors.
   - If it exists $d_k \in P_k$ 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.

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.

A classical, first-order 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) $D_{\oplus} = [I \quad -I]$ *is a PSS such that*

$$\forall v \neq 0, \ \mathrm{cm}(D_{\oplus}, v) \geq \frac{1}{\sqrt{n}}.$$

A classical, first-order 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) $D_{\oplus} = [I \quad -I]$ *is a PSS such that*

$$\forall v \neq 0, \ \mathrm{cm}(D_{\oplus}, v) \geq \frac{1}{\sqrt{n}}.$$

A classical, first-order 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) $D_{\oplus} = [I \quad -I]$ *is a PSS such that*

$$\forall v \neq 0,\ \mathrm{cm}(D_{\oplus}, v) \geq \frac{1}{\sqrt{n}}.$$

First-order polling choice

At every iteration k, the polling set P_k contains a PSS D_k such that

$$\mathrm{cm}\left(D_k, -\nabla f(x_k)\right) \geq \kappa,$$

with $\kappa \in (0, 1)$.

First-order results

Two convergence arguments

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

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

Theorem (First-order convergence)

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

Outline

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 or reason on successive iterations.

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

Towards second-order aspects

- The first-order polling works because it is easy to approximate a vector with a set of vectors;

Towards second-order aspects

- The first-order polling works because it is easy to approximate a vector with a set of vectors;
- For second-order, what we want to estimate is

$$\lambda_{\min}\left(\nabla^2 f(x_k)\right) \;=\; \min_{\|v\|=1} v^\top \nabla^2 f(x_k)\, v,$$

A minimization problem in itself !

Towards second-order aspects

- The first-order polling works because it is easy to approximate a vector with a set of vectors;
- For second-order, what we want to estimate is

$$\lambda_{\min}\left(\nabla^2 f(x_k)\right) = \min_{\|v\|=1} v^\top \nabla^2 f(x_k)\, v,$$

  A minimization problem in itself !
- We may need the equivalent of a matrix.

Towards second-order aspects

- The first-order polling works because it is easy to approximate a vector with a set of vectors;
- For second-order, what we want to estimate is

$$\lambda_{\min}\left(\nabla^2 f(x_k)\right) = \min_{\|v\|=1} v^\top \nabla^2 f(x_k)\, v,$$

  A minimization problem in itself !
- We may need the equivalent of a matrix.

What we propose

- A polling strategy that is good in a second-order sense...
- ...but only estimates the Hessian as a last resort.

A second-order polling strategy for Direct Search

Second-order polling choice

The polling set P_k is made of the following elements:

1. A positive spanning set D_k (First-order choice);
2. Its opposite -D_k;
3. Two unitary vectors v_k and -v_k obtained by:
   - Selecting a basis $B_k \subset D_k$;
   - Building a matrix $H_k \approx B_k^\top \nabla^2 f(x_k)\, B_k$ using function values;
   - Solving $H_k\, v_k = \lambda_{\min}(H_k)\, v_k$.

A second-order polling strategy for Direct Search

Second-order polling choice

The polling set P_k is made of the following elements:

1. A positive spanning set D_k (First-order choice);
2. Its opposite -D_k;
3. Two unitary vectors v_k and -v_k obtained by:
   - Selecting a basis $B_k \subset D_k$;
   - Building a matrix $H_k \approx B_k^\top \nabla^2 f(x_k)\, B_k$ using function values;
   - Solving $H_k\, v_k = \lambda_{\min}(H_k)\, v_k$.

- The cost of an iteration is at most $\mathcal{O}(n^2)$ evaluations.
- The polling and the construction of P_k stop as soon as a direction d satisfying

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

is encountered.

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

If $\lambda_{\min}\left(\nabla^2 f(x_k)\right) < 0$,

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

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

Second-order convergence (2)

Convergence arguments

- We still have $\alpha_k \to 0$;
- On an unsuccessful iteration k, one has:

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

Second-order convergence (2)

Convergence arguments

- We still have $\alpha_k \to 0$;
- 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.

$$\inf_{0 \leq l \leq k} \|\nabla f(x_l)\| < \epsilon_g \quad \text{and} \quad \sup_{0 \leq l \leq k} \lambda_{\min}(\nabla^2 f(x_l)) > -\epsilon_H.$$

Theorem

Let $N_{\epsilon_{gH}}$ the number of evaluations of f needed to reach an (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*

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

A testing framework

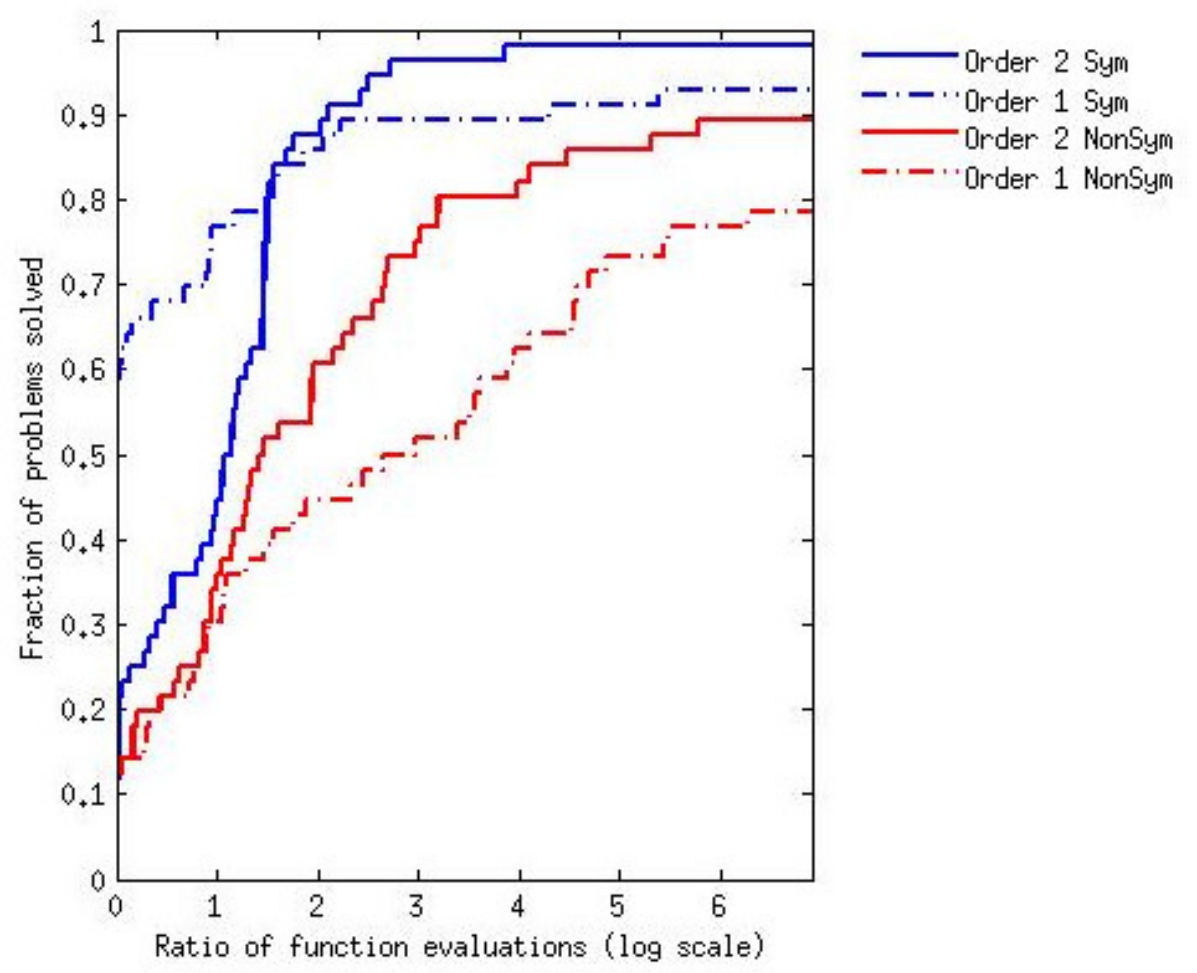
- Benchmark : 60 CUTEst problems with negative curvature;
- Four polling choices :
  - Two types of PSS: $D_{\oplus}$ (symmetric) and V_{n+1} (non symmetric);
  - Two rules: First/Second-order polling.
- A method has converged if

$$f(x_k) - f_* \;<\; 10^{-3}\,(f(x_0) - f_*)\,,$$

where f_* is the best value obtained by the methods in $2000\,n$ evaluations of f.

Practical insights

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



Outline

Summary

Our contributions

- The design of a second-order globally convergent direct-search framework;
- The associated complexity results;
- Numerical interest in the proposed approach.

Summary

Our contributions

- The design of a second-order globally convergent direct-search framework;
- The associated complexity results;
- Numerical interest in the proposed approach.

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 *Optimization*.

Two perspectives

Using two step sizes

- In the previous method, a single step size accounts for two optimality measures;
- Objective: Decouple the first and second-order aspects to handle scaling issues.

Two perspectives

Using two step sizes

- In the previous method, a single step size accounts for two optimality measures;
- Objective: Decouple the first and second-order aspects to handle scaling issues.

Towards randomization

- Guaranteeing

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

  is sufficient for first-order convergence, and we can do it in practice (Gratton, R., Vicente and Zhang '15).
- Can we do the same with second-order properties ?

Merci beaucoup !

clement.royer@enseeiht.fr