From First to Second-Order Quality Measures in Direct-Search Methods

Clément Royer
ENSEEIHT-IRIT, Toulouse, France

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

Journée des doctorants APO - 19/11/15

A little warning regarding the presentation

My thesis

- Official title: Stochastic methods for derivative-free optimization.
- A better title: Probabilistic tools in derivative-free optimization.

The idea is to randomize existing derivative-free algorithms.

This talk will not address randomization

- The deterministic material was not appropriate for a randomized analysis;
- We worked out our own method, and obtained new results.

Outline

Introductory assumptions and definitions

We consider an unconstrained smooth nonconvex optimization problem:

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

Assumptions on f

- f bounded from below, twice continuously differentiable;
- $\nabla f, \nabla^2 f$ are Lipschitz continuous;
- f nonconvex $\Rightarrow$ for some x, the matrix $\nabla^2 f(x)$ has negative eigenvalues.

Second-order optimality

Optimality conditions

If x^* is a local minimum of f,

$$\begin{cases} \|\nabla f(x^*)\| & = 0 \quad (1^{\text{st}} \text{ order}) \\ \lambda_{\min}\left(\nabla^2 f(x^*)\right) & \geq 0 \quad (2^{\text{nd}} \text{ order}) \end{cases}$$

Second-order optimality

Optimality conditions

If x^* is a local minimum of f,

$$\forall d \in \mathbb{R}^n, \begin{cases} d^\top \nabla f(x^*) & \geq & 0 \\ d^\top \nabla^2 f(x^*)\, d & \geq & 0. \end{cases}$$

Second-order optimality

Optimality conditions

If x^* is a local minimum of f,

$$\forall d \in \mathbb{R}^n, \begin{cases} d^\top \nabla f(x^*) & \geq & 0 \\ d^\top \nabla^2 f(x^*)\, d & \geq & 0. \end{cases}$$

Progress towards a minimum

If x is not a minimum, we can move in a direction satisfying either

- $d^\top \nabla f(x) < 0$ (descent direction) or
- $d^\top \nabla^2 f(x)\, d < 0$ (negative curvature direction),

possibly decreasing the function value.

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. **Polling 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 polling step.

A simple direct-search framework

1. **Initialization** Set $x_0, \alpha_0 > 0, \theta < 1 \leq \gamma$.
   Set $k = 0$.
2. **Polling 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 polling step.

What are the rules to choose the polling sets ?

Outline

Polling choice in direct search

- Typical direct-search methods ensure convergence of a subsequence of $\{\|\nabla f(x_k)\|\}$ to 0.
- This is possible if the polling sets are good in a first-order sense.

Polling choice in direct search

- Typical direct-search methods ensure convergence of a subsequence of $\{\|\nabla f(x_k)\|\}$ to 0.
- This is possible if the polling sets are good in a first-order sense.

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.

Polling choice in direct search

- Typical direct-search methods ensure convergence of a subsequence of $\{\|\nabla f(x_k)\|\}$ to 0.
- This is possible if the polling sets are good in a first-order sense.

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.

Ensuring (first-order) set quality

We do not know ∇f, thus we would like to have

$$\mathrm{cm}(D, v) > 0 \quad \forall v \neq 0.$$

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;

First-order polling rule

1. Poll along a PSS D_k.

First-order convergence

Lemma

Independently of the polling rule,

$$\lim_{k \to \infty} \alpha_k = 0.$$

Proposition

If the k-th iteration is unsuccessful and $\mathrm{cm}(D_k, -\nabla f(x_k)) \geq \kappa > 0$, *then*

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

First-order convergence

Lemma

Independently of the polling rule,

$$\lim_{k\to\infty} \alpha_k = 0.$$

Proposition

If the k-th iteration is unsuccessful and $\mathrm{cm}(D_k, -\nabla f(x_k)) \geq \kappa > 0$, *then*

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

Theorem (First-order convergence)

If $\forall k,\ \mathrm{cm}(D_k, -\nabla f(x_k)) \geq \kappa$, we have

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

Outline

Direct search and second-order properties

- Using PSS already ensures first-order convergence.
- To take second-order/curvature aspects into consideration, one needs additional properties on the polling sets.

Direct search and second-order properties

- Using PSS already ensures first-order convergence.
- To take second-order/curvature aspects into consideration, one needs additional properties on the polling sets.

Few results, mostly due to Abramson et al (2005,2006,2014)

- Second-order aspects are observed at limits of convergent subsequences;
- Consecutive polling sets are not independent.

Our approach

- Keep the first-order reasoning;
- Exploit second order at the iteration level.

Towards second-order aspects

- We aim to capture curvature information with the function values. A natural tool is

$$f(x+d) - f(x) + f(x-d) - f(x) \approx d^\top \nabla^2 f(x)\, d.$$

  Opposite directions may help.

- Our goal is to relate $\lambda_{\min}(\nabla^2 f(x))$ and the polling directions that we use.

Our concerns

- Can we define a second-order quality measure ?
- Can we characterize polling strategy that are good in a second-order sense ?

A second-order criticality measure

The symmetric part of a set of vectors D is defined by

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

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

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

It is an approximation of the minimum eigenvalue of A.

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 are not known;
- Direct search: The step size goes to zero;

We will be ensuring

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

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 $H_k \approx B_k^\top \nabla^2 f(x_k)\, B_k$;
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 $H_k \approx B_k^\top \nabla^2 f(x_k)\, B_k$;
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 such that $\forall v \neq 0,\ \mathrm{cm}(D_k, v) \geq \kappa > 0$;
- 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)

Lemma

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)

Lemma

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)

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

Is the second-order necessary ?

- Apply the classic direct-search framework to the following function:

$$f_1(x, y) = (9x - y)(11\,x - y) + \frac{x^4}{2},$$

  which is minimal at $\pm(1\,,10)$.
- If the method starts from $x_0 = (0\,,0)$ and uses $D_k = [I\ \text{-}I]$, it will never escape the origin.
- The second-order polling rules allow to escape this point and to converge.

Is the second-order necessary ?

- Apply the classic direct-search framework to the following function:

$$f_1(x, y) = (9x - y)(11\,x - y) + \frac{x^4}{2},$$

  which is minimal at $\pm(1\,, 10)$.
- If the method starts from $x_0 = (0\,, 0)$ and uses $D_k = [I\ \text{-}I]$, it will never escape the origin.
- The second-order polling rules allow to escape this point and to converge.

What happens on non-pathological examples ?

Practical insights

On 60 `CUTEr/st` problems with negative curvature:

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

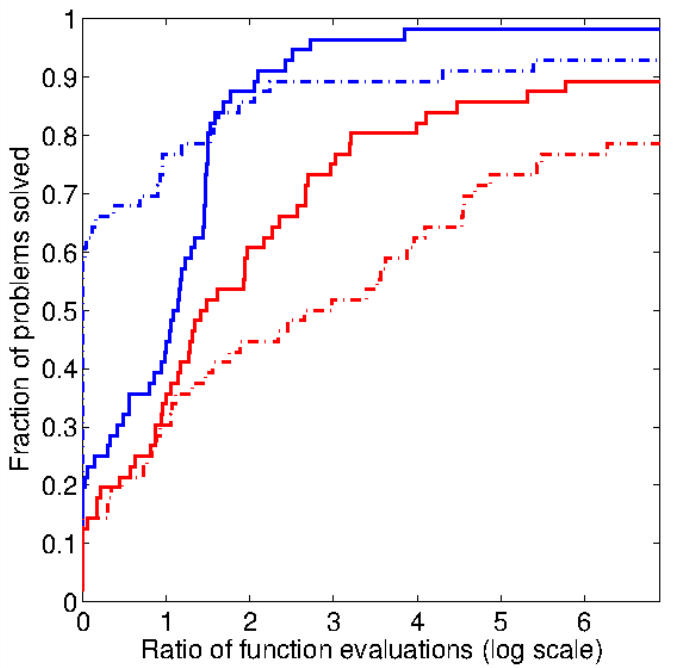

Figure : Performance of methods based on symmetric and nonsymmetric PSS, given a budget of $2000\,n$ evaluations.

Conclusion

Our contributions

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

Perspectives

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*.

...and for the future

- We know how to guarantee $\mathrm{cm}(D_k, -\nabla f(x_k)) > \kappa$ with a certain probability while maintaining the convergence.
- Can we do the same with second-order properties ?

Thank you for your attention !

Parenthèse

- L'IRIT a une commission des doctorants, avec un représentant par équipe proposé par les doctorants de l'équipe.
- Représentant APO : Clément ROYER (2014 - ?).
- Remplissez le questionnaire des doctorants !