Derivative-Free Optimization Methods based on Probabilistic and Deterministic Properties: Complexity Analysis and Numerical Relevance

Clément W. Royer

Thesis defence - Soutenance de doctorat
Université de Toulouse

4 novembre 2016

The thesis

Supervisors	S. Gratton (INPT-ENSEEIHT, Univ. Toulouse) L. N. Vicente (Univ. Coimbra, Portugal).
Laboratory	Institut de Recherche en Informatique de Toulouse.
Funding	Université Toulouse 3 Paul Sabatier.

A thesis in numerical optimization

Main topics:

- Introduction of random elements in derivative-free optimization.
- Complexity as a designing tool of optimization methods.

Introduction: Numerical optimization

An Optimization Problem

- An objective function $f(x)$ to be minimized or maximized.
- A set of values for x.

Goal: find the value(s) of x giving the best value of f.

Numerical Optimization

Obj: Develop algorithms to solve optimization problems.

- Theoretical analysis.
- Practical implementation.

Introduction: Randomness and optimization

Randomness has triggered significant recent advances in numerical optimization.

Multiple reasons:

- *Large-scale setting:* Classical methods too expensive.
- *Distributed computing:* Data not stored on a single computer/processor.
- *Applications:* Machine learning.

Introduction: Randomness and optimization

Randomness has triggered significant recent advances in numerical optimization.

Multiple reasons:

- *Large-scale setting:* Classical methods too expensive.
- *Distributed computing:* Data not stored on a single computer/processor.
- *Applications:* Machine learning.

Concerning randomness

- How does it affect the analysis of a method ?
- Improvement over deterministic ?
- Randomness in derivative-free methods ?

Complexity of optimization algorithms

Complexity Analysis

- Estimate the convergence rate of a given criterion.
- Provide worst-case bounds on algorithmic behavior.
- In presence of randomness: results in expectation.

Using complexity

- Guidance provided by complexity ?
- Practical relevance ?
- Importance for derivative-free methods ?

Objectives pursued in the thesis

Main track

1. Introduce random aspects in derivative-free frameworks.
2. Provide theoretical guarantees (especially complexity).
3. Compare complexity results with numerical behavior.
4. Treat first-order and second-order aspects.

Objectives pursued in the thesis

Main track

1. Introduce random aspects in derivative-free frameworks.
2. Provide theoretical guarantees (especially complexity).
3. Compare complexity results with numerical behavior.
4. Treat first-order and second-order aspects.

- In this talk: focus on direct-search methods;
- In the thesis: direct-search and trust-region algorithms.

Outline

1. Deterministic direct search
2. Direct search based on probabilistic descent
3. Deterministic and probabilistic second-order methods
4. Summary and conclusions

Outline

1. Deterministic direct search
   - Derivative-free optimization
   - Direct search
2. Direct search based on probabilistic descent
3. Deterministic and probabilistic second-order methods
4. Summary and conclusions

Introductory assumptions and definitions

We consider an unconstrained smooth problem:

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

Assumptions on f

- f bounded from below.
- f continuously differentiable, ∇f Lipschitz continuous.

Introductory assumptions and definitions

We consider an unconstrained smooth problem:

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

Assumptions on f

- f bounded from below.
- f continuously differentiable, ∇f Lipschitz continuous.

Solving the problem using the derivative

At $x \in \mathbb{R}^n$, moving along $-\nabla f(x)$ can decrease the function value !

- Basic paradigm of *gradient-based* methods.
- Goal: convergence towards a first-order stationary point

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

A derivative-free hypothesis

The gradient exists but cannot be used in an algorithm.

- *Simulation code:* gradient too expensive to be computed.
- *Black-box objective function:* no derivative code available.
- *Automatic differentiation:* inapplicable.

Examples: Weather forecasting, oil industry, biology,...

A derivative-free hypothesis

The gradient exists but cannot be used in an algorithm.

- *Simulation code:* gradient too expensive to be computed.
- *Black-box objective function:* no derivative code available.
- *Automatic differentiation:* inapplicable.

Examples: Weather forecasting, oil industry, biology,...

Performance indicator: Number of function evaluations.

Derivative-Free Optimization (DFO) algorithms

Deterministic DFO methods

- Model-based methods, e.g. Trust Region.
- Directional methods, e.g. Direct Search.

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

- Well-established: convergence theory (to local optima).
- Recent advances: complexity bounds/convergence rates.

Derivative-Free Optimization (DFO) 'ed

Stochastic DFO

- Typically global optimization methods:
  Ex) Evolution Strategies, Genetic Algorithms.
- Often use heuristics $\Rightarrow$ No general proof of convergence.
- No deterministic variant.

- This thesis did NOT address those methods.
- Distinction: stochastic VS using probabilistic elements.

DFO methods based on probabilistic properties

- Developed from deterministic algorithms.
- Keep theoretical guarantees from deterministic.
- Improve performance with randomness.

Outline

Direct search (DS)

- Directional methods $\sim$ Steepest/Gradient Descent.
- Early appearance: 1960s, convergence theory: 1990s.
- Attractive: simplicity, parallel potential.

- **Optimization by direct search: new perspectives on some classical and modern methods.**
  Kolda, Lewis and Torczon (*SIAM Review*, 2003).

A basic framework for direct-search algorithms

1. **Initialization:** Set $x_0 \in \mathbb{R}^n, \alpha_0 > 0, 0 < \theta < 1 \leq \gamma$.
2. **For** $k = 0, 1, 2, \ldots$
   - Choose a set D_k of m vectors.
   - If it exists $d_k \in D_k$ so that

     $$f(x_k + \alpha_k\, d_k) < f(x_k) - \alpha_k^2,$$

     then declare k *successful*, set $x_{k+1} := x_k + \alpha_k\, d_k$ and update $\alpha_{k+1} := \gamma\, \alpha_k$.
   - Otherwise declare k *unsuccessful*, set $x_{k+1} := x_k$ and update $\alpha_{k+1} := \theta\, \alpha_k$.

A basic framework for direct-search algorithms

1. **Initialization:** Set $x_0 \in \mathbb{R}^n, \alpha_0 > 0, 0 < \theta < 1 \leq \gamma$.
2. **For** $k = 0, 1, 2, \ldots$
   - Choose a set D_k of m vectors.
   - If it exists $d_k \in D_k$ so that

     $$f(x_k + \alpha_k d_k) < f(x_k) - \alpha_k^2,$$

     then declare k *successful*, set $x_{k+1} := x_k + \alpha_k d_k$ and update $\alpha_{k+1} := \gamma \alpha_k$.
   - Otherwise declare k *unsuccessful*, set $x_{k+1} := x_k$ and update $\alpha_{k+1} := \theta \alpha_k$.

Polling choice in deterministic direct search

We would like to choose directions/polling sets D_k sufficiently good to ensure convergence.

Polling choice in deterministic direct search

We would like to choose directions/polling sets D_k sufficiently good to ensure convergence.

A measure of set quality

For a set of vectors D, the cosine measure of D is

$$\mathrm{cm}(D) = \min_{v \in \mathbb{R}^n \setminus \{0\}} \max_{d \in D} \frac{d^\top v}{\|d\| \, \|v\|}.$$

Polling choice in deterministic direct search

We would like to choose directions/polling sets D_k sufficiently good to ensure convergence.

A measure of set quality

For a set of vectors D, the cosine measure of D is

$$\mathrm{cm}(D) = \min_{v \in \mathbb{R}^n \setminus \{0\}} \max_{d \in D} \frac{d^\top v}{\|d\| \, \|v\|}.$$

- When $\mathrm{cm}(D) > 0$, any v makes an acute angle with some $d \in D$.
- If $v = -\nabla f(x) \neq 0$, D contains a descent direction for f at x.

Set quality

We would like to have $\text{cm}(D) > 0$.

Positive Spanning Sets (PSS)

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

- D is a PSS **iff** $\text{cm}(D) > 0$.
- A PSS contains at least $n+1$ vectors.

We would like to have $\mathrm{cm}(D) > 0$.

Positive Spanning Sets (PSS)

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

- D is a PSS **iff** $\mathrm{cm}(D) > 0$.
- A PSS contains at least $n + 1$ vectors.

Example

$D_\oplus = \{e_1, \dots, e_n, \text{-}e_1, \dots, \text{-}e_n\}$ is a PSS with

$$\mathrm{cm}\left(D_\oplus\right) = \frac{1}{\sqrt{n}}.$$

Convergence for deterministic direct search

Lemma

Independently of $\{D_k\}$,

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

Lemma

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

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

Convergence for deterministic direct search

Lemma

Independently of $\{D_k\}$,

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

Lemma

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

$$\kappa \|\nabla f(x_k)\| \;\leq\; \mathcal{O}(\alpha_k).$$

Convergence Theorem

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

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

Worst-case complexity in deterministic direct search

Theorem (Vicente 2013)

Let $\epsilon \in (0,1)$ and N_ϵ be the number of function evaluations needed to reach an point such that $\inf_{0 \le l \le k} \|\nabla f(x_l)\| < \epsilon$. Then,

$$N_\epsilon \;\le\; \mathcal{O}\left(m\,(\kappa\,\epsilon)^{-2}\right).$$

Choosing $D_k = D_\oplus$, one has $\kappa = 1/\sqrt{n}, m = 2n$, and the bound becomes

$$N_\epsilon \;\le\; \mathcal{O}\left(n^2\,\epsilon^{-2}\right).$$

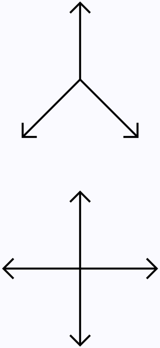
Outline

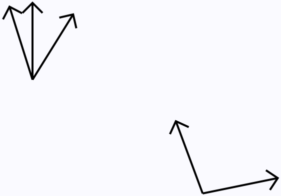
Introducing randomness

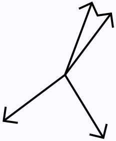
Idea from Gratton and Vicente (2013)

Randomly independently generate polling sets, possibly of
less than $n + 1$ vectors!

From PSS...







...to random sets

Numerical motivations

- Convergence test: $f(x_k) < f_{\mathrm{low}} + 10^{-3}\,(f(x_0) - f_{\mathrm{low}})$;
- Budget: $2000\,n$ evaluations.

Problem	$D_\oplus$	$Q\,D_\oplus$	$2\,n$	$n+1$	$n/2$	2	1
	Deterministic		Probabilistic				
arglina	3.42	16.67	10.30	6.01	3.21	1.00	–
arglinb	20.50	11.38	7.38	2.81	2.35	1.00	2.04
broydn3d	4.33	11.22	6.54	3.59	2.04	1.00	–
dqrtic	7.16	19.50	9.10	4.56	2.77	1.00	–
engval1	10.53	23.96	11.90	6.48	3.55	1.00	2.08
freuroth	56.00	1.33	1.00	1.67	1.33	1.00	4.00
integreq	16.04	18.85	12.44	6.76	3.52	1.00	–
nondquar	6.90	17.36	7.56	4.23	2.76	1.00	–
sinquad	–	2.12	1.31	1.00	1.60	1.23	–
vardim	1.00	3.30	1.80	2.40	2.30	1.80	4.30

Table : Relative number of function evaluations for different types of polling (mean on 10 runs, $n = 40$)

A probabilistic direct-search algorithm

From deterministic to probabilistic notations

- Polling sets/directions: $D_k = \mathfrak{D}_k(\omega)$, $d_k = \mathfrak{d}_k(\omega)$;
- Iterates: $x_k = X_k(\omega)$;
- Step sizes: $\alpha_k = \mathcal{A}_k(\omega)$.

1. **Initialization:** Set $x_0 \in \mathbb{R}^n, \alpha_0 > 0, 0 < \theta < 1 \leq \gamma$.
2. **For** $k = 0, 1, 2, ...,$
   - Choose a set $\mathfrak{D}_k$ of m **independent random** vectors.
   - If it exists $\mathfrak{d}_k \in \mathfrak{D}_k$ so that

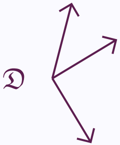
$$f(X_k + \mathcal{A}_k \, \mathfrak{d}_k) < f(X_k) - \mathcal{A}_k^2,$$

     then declare k successful, set $X_{k+1} := X_k + \mathcal{A}_k \, \mathfrak{d}_k$ and update $\mathcal{A}_{k+1} := \gamma \, \mathcal{A}_k$.
   - Otherwise, declare k unsuccessful, set $X_{k+1} := X_k$ and update $\mathcal{A}_{k+1} := \theta \, \mathcal{A}_k$.

Outline

First step: What is a good random polling set ?

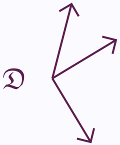
$\mathfrak{D}$ is not a PSS...

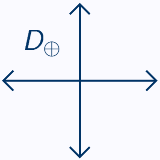


First step: What is a good random polling set ?

$\mathfrak{D}$ is not a PSS...

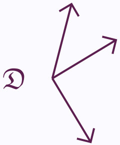
...$D_\oplus$ is...

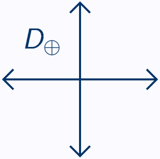


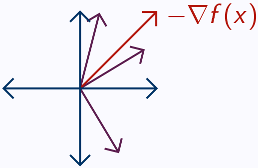


First step: What is a good random polling set ?

$\mathfrak{D}$ is not a PSS... ...$D_\oplus$ is... ...but here $-\nabla f(x)$ is closer to $\mathfrak{D}$!







Is being close to the negative gradient a sign of quality ?

A new measure of set quality

Set assumption in the deterministic case

- We required

$$\mathrm{cm}(D_k) = \min_{v \neq 0} \max_{d \in D_k} \frac{d^\top v}{\|d\| \, \|v\|} \geq \kappa.$$

- What we really need is

$$\mathrm{cm}\left(D_k, -\nabla f(x_k)\right) = \max_{d \in D_k} \frac{d^\top (-\nabla f(x_k))}{\|d\| \|\nabla f(x_k)\|} \geq \kappa.$$

- In the random case, the second one might happen with some probability.
- Can we find adequate probabilistic tools to express this fact ?

Probabilistic analysis

Several types of results

Deterministic/For all realizations

$\Downarrow$

With probability 1/Almost-sure

$\Downarrow$

With a given probability.

Submartingale

A submartingale is a sequence of random variables $\{V_k\}$ such that $\mathbb{E}[|V_k|] < \infty$ and

$$\mathbb{E}\left(V_k | \sigma\left(V_0, V_1, \dots, V_{k-1}\right)\right) \geq V_{k-1}.$$

(p, κ)-descent sets

- We want to look at

$$\mathbb{P}\left(\mathrm{cm}\left(\mathfrak{D}_k, -\nabla f(X_k)\right) \geq \kappa\right).$$

where X_k depends on $\mathfrak{D}_0, \ldots, \mathfrak{D}_{k-1}$ but not on $\mathfrak{D}_k$.

- A solution is to use conditional probabilities/conditioning to the past.

(p, κ)-descent sets

- We want to look at

$$\mathbb{P}\left(\mathrm{cm}\left(\mathfrak{D}_k, -\nabla f(X_k)\right) \geq \kappa\right).$$

where X_k depends on $\mathfrak{D}_0, \dots, \mathfrak{D}_{k-1}$ but not on $\mathfrak{D}_k$.

- A solution is to use conditional probabilities/conditioning to the past.

Probabilistic descent property

A random set sequence $\{\mathfrak{D}_k\}$ is said to be (p, κ)-descent if:

$$\begin{aligned} \mathbb{P}\left(\mathrm{cm}\left(\mathfrak{D}_0, -\nabla f(x_0)\right) \geq \kappa\right) &\geq p \\ \forall k \geq 1, \quad \mathbb{P}\left(\mathrm{cm}\left(\mathfrak{D}_k, -\nabla f(X_k)\right) \geq \kappa \,\middle|\, \mathfrak{S}_{k-1}^{\mathfrak{D}}\right) &\geq p, \end{aligned}$$

where $\mathfrak{S}_{k-1}^{\mathfrak{D}} = \sigma(\mathfrak{D}_0, \dots, \mathfrak{D}_{k-1})$.

Convergence results

Lemma

For all realizations $\{\alpha_k\}$ *of* $\{\mathcal{A}_k\}$, *independently of* $\{\mathfrak{D}_k\}$,

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

Lemma

If k *is an unsuccessful iteration; then*

$$\{\mathrm{cm}\,(\mathfrak{D}_k, -\nabla f(X_k)) \geq \kappa\} \subset \{\kappa \|\nabla f(X_k)\| \leq \mathcal{O}(\mathcal{A}_k)\}.$$

We need to show that $\{\mathrm{cm}\,(\mathfrak{D}_k, -\nabla f(X_k)) \geq \kappa\}$ happens sufficiently often.

Convergence results (2)

Let $\{\mathfrak{D}_k\}$ (p, κ)-descent and $Z_k = \mathbf{1}\left(\mathrm{cm}\left(\mathfrak{D}_k, -\nabla f(X_k)\right) \geq \kappa\right)$.

Proposition

Consider

$$S_k = \sum_{i=0}^{k-1} \left[Z_i - p_0\right], \quad p_0 = \frac{\ln \theta}{\ln(\theta/\gamma)}.$$

1. If $\liminf_k \|\nabla f(X_k)\| > 0$, then $S_k \to -\infty$.
2. If $p > p_0$, $\{S_k\}$ is a submartingale and $\mathbb{P}\left(\limsup S_k = \infty\right) = 1$.

Convergence results (2)

Let $\{\mathfrak{D}_k\}$ (p,κ)-descent and $Z_k = \mathbf{1}\left(\mathrm{cm}\left(\mathfrak{D}_k, -\nabla f(X_k)\right) \geq \kappa\right)$.

Proposition

Consider

$$S_k = \sum_{i=0}^{k-1} \left[Z_i - p_0\right], \quad p_0 = \frac{\ln \theta}{\ln(\theta/\gamma)}.$$

1. If $\liminf_k \|\nabla f(X_k)\| > 0$, then $S_k \to -\infty$.
2. If $p > p_0$, $\{S_k\}$ is a submartingale and $\mathbb{P}\left(\limsup S_k = \infty\right) = 1$.

Almost-sure Convergence Theorem

If $\{\mathfrak{D}_k\}$ is (p,κ)-descent with $p > p_0$, then

$$\mathbb{P}\left(\liminf_{k\to\infty} \|\nabla f(X_k)\| = 0 \right) = 1.$$

Probabilistic complexity bound

Probabilistic worst-case complexity

Let $\{\mathfrak{D}_k\}$ be (p,κ)-descent, $\epsilon \in (0,1)$ and N_ϵ the number of function evaluations needed to have $\inf_{0 \le l \le k} \|\nabla f(X_l)\| \le \epsilon$. Then

$$\mathbb{P}\left(N_\epsilon \le \mathcal{O}\left(\frac{m\,(\kappa\epsilon)^{-2}}{p-p_0}\right)\right) \ge 1 - \exp\left(-\mathcal{O}\left(\frac{p-p_0}{p}(\kappa\,\epsilon)^{-2}\right)\right).$$

Probabilistic complexity bound

Probabilistic worst-case complexity

Let $\{\mathfrak{D}_k\}$ be (p,κ)-descent, $\epsilon \in (0,1)$ and N_ϵ the number of function evaluations needed to have $\inf_{0 \le l \le k} \|\nabla f(X_l)\| \le \epsilon$. Then

$$\mathbb{P}\left(N_\epsilon \le \mathcal{O}\left(\frac{m\,(\kappa\epsilon)^{-2}}{p-p_0}\right)\right) \ge 1 - \exp\left(-\mathcal{O}\left(\frac{p-p_0}{p}(\kappa\,\epsilon)^{-2}\right)\right).$$

- Deterministic: $\mathcal{O}(n^2\,\epsilon^{-2})$.
- Probabilistic: $\mathcal{O}(m\,n\,\epsilon^{-2})$ in probability
  $\Rightarrow \mathcal{O}(n\,\epsilon^{-2})$ when $m = 2$!
- Improvement with high probability using few directions ?

Outline

A practical (p, κ)-descent sequence

We must ensure

$$p \;>\; p_0 \;=\; \frac{\ln(\theta)}{\ln(\theta/\gamma)}$$

with the minimum $m = |\mathfrak{D}_k|$ possible.

A practical example: uniform distribution over the unit sphere

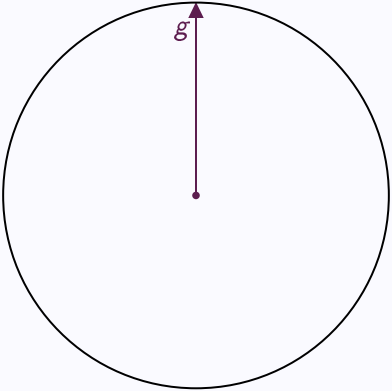
If

$$m > \log_2 \left(1 - \frac{\ln \theta}{\ln \gamma}\right),$$

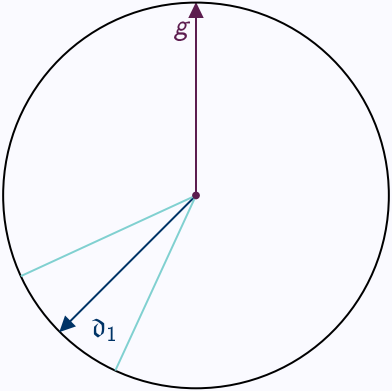
then there exist p and τ independent of n such that the sequence $\mathfrak{D}_k$ is $(p, \tau/\sqrt{n})$-descent, with $p > p_0$.

If $\gamma = \theta^{-1} = 2$, it suffices to choose $m \geq 2$ to have $p > \frac{1}{2}$.

Two uniform directions are enough, one is not

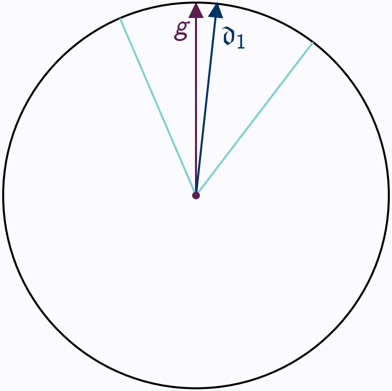


Two uniform directions are enough, one is not



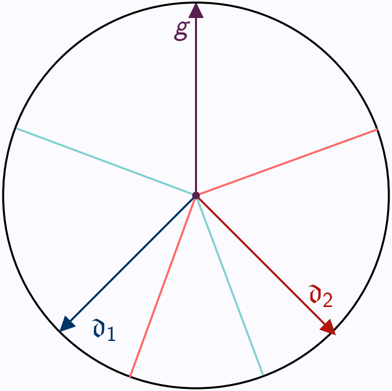
$$\mathfrak{d}_1 \sim \mathcal{U}(\mathbb{S}^1) \Rightarrow \forall \kappa \in (0,1), \quad \mathbb{P}\left(\mathrm{cm}\left(\mathfrak{d}_1, g\right) = \mathfrak{d}_1^\top g \geq \kappa\right) < 1/2.$$

Two uniform directions are enough, one is not



$$\mathfrak{d}_1 \sim \mathcal{U}(\mathbb{S}^1) \Rightarrow \forall \kappa \in (0,1), \quad \mathbb{P}\left(\mathrm{cm}\left(\mathfrak{d}_1, g\right) = \mathfrak{d}_1^\top g \geq \kappa\right) < 1/2.$$

Two uniform directions are enough, one is not



$$\mathfrak{d}_1 \sim \mathcal{U}(\mathbb{S}^1) \Rightarrow \forall \kappa \in (0,1), \quad \mathbb{P}\left(\mathrm{cm}\left(\mathfrak{d}_1, g\right) = \mathfrak{d}_1^\top g \geq \kappa\right) < 1/2.$$

$$\mathfrak{d}_1, \mathfrak{d}_2 \sim \mathcal{U}(\mathbb{S}^1) \Rightarrow \exists \kappa^* \in (0,1), \quad \mathbb{P}\left(\mathrm{cm}\left(\{\mathfrak{d}_1, \mathfrak{d}_2\}, g\right) \geq \kappa^*\right) > 1/2.$$

Outline

Exploiting second-order derivatives

- Previous analysis was concerned with first-order aspects.
- We improved the deterministic case and saved function values.

- Second-order considerations can come into play.
- Usually at a higher expense in evaluations, especially in DFO.

Second-order optimality

Assumption

- f twice continuously differentiable, ∇f and $\nabla^2 f$ Lipschitz continuous.
- f typically nonconvex.

Second-order methods

- Exploit (negative) curvature information given by $\nabla^2 f$.
- Converge towards second-order stationary points:

$$\liminf_k \max \left\{ \|\nabla f(x)\|, -\lambda_{\min}\left(\nabla^2 f(x)\right) \right\} = 0.$$

A new deterministic second-order direct search

Objective

- Introduce second order in our framework.
- Guarantees at the iteration level.
- Complexity analysis.

Key features

- A PSS D_k, as before.
- A linear basis B_k used to gather curvature information.
- Polling sets are of size $\mathcal{O}(n^2)$.
- Function decrease: α_k^3.

Second-order convergence

Arguments

- We still have $\alpha_k \to 0$.
- On unsuccessful iterations:
  - D_k is a PSS $\Rightarrow \quad \|\nabla f(x_k)\| \leq \mathcal{O}(\alpha_k)$
  - B_k well conditioned $\Rightarrow -\lambda_{\min}\left(\nabla^2 f(x_k)\right) \leq \mathcal{O}\left(\alpha_k\right)$.

Theorem

If there exist $\kappa, \sigma \in (0,1)$ such that

$$\forall k, \quad \mathrm{cm}(D_k) \geq \kappa \quad \& \quad \sigma_{\min}(B_k) \geq \sigma,$$

then

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

Complexity of second-order direct search

Theorem

For $(\epsilon_g, \epsilon_H) \in (0,1)^2$, the number of evaluations of f needed to achieve

$$\begin{cases} \inf_{0 \le l \le k} \|\nabla f(x_k)\| < \epsilon_g \\ \sup_{0 \le l \le k} \lambda_{\min}\left(\nabla^2 f(x_k)\right) > -\epsilon_H \end{cases}$$

is of order

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

Complexity of second-order direct search

Theorem

For $(\epsilon_g, \epsilon_H) \in (0,1)^2$, the number of evaluations of f needed to achieve

$$\begin{cases} \inf_{0 \le l \le k} \|\nabla f(x_k)\| < \epsilon_g \\ \sup_{0 \le l \le k} \lambda_{\min}\left(\nabla^2 f(x_k)\right) > -\epsilon_H \end{cases}$$

is of order

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

- Second-order expense (much) higher than first-order: Power of tolerances + $\mathcal{O}(n^2)$ evaluations per iteration.
- Reflects on practice:
  - Second order more robust...
  - ...but more expensive.

From deterministic to probabilistic second-order methods

What we have

- Second-order convergent deterministic method.
- First-order convergent probabilistic method.

What we would like

- Incorporate randomness in the second-order method.
- Improve its worst-case cost.

Two ways of introducing randomness

On the "first-order" directions

- We can satisfy $\mathrm{cm}(D_k, -\nabla f(x_k)) \geq \kappa$ in probability...
- ...with deterministic B_k !

On the "second-order" directions

- Focus on ensuring $\mathbb{P}\left(\sigma_{\min}(B_k) \geq \sigma\right)$;
- Use results from random linear algebra.

Two ways of introducing randomness

On the “first-order" directions

- We can satisfy $\mathrm{cm}(D_k, -\nabla f(x_k)) \geq \kappa$ in probability...
- ...with deterministic B_k !

On the “second-order" directions

- Focus on ensuring $\mathbb{P}\left(\sigma_{\min}(B_k) \geq \sigma\right)$;
- Use results from random linear algebra.

- Both converge almost surely.
- Still $\mathcal{O}(n^2)$ evaluations per iteration.
- Challenge: Get rid of B_k in probability.

Outline

Main conclusions and contributions

- Derivative-free optimization can be combined with probabilistic tools.
- Convergence can be maintained.
- Practical performance is enhanced in the direct-search case.
- Complexity confirms the numerical observations.

Direct search based on probabilistic descent
Gratton, Royer, Vicente and Zhang, *SIAM J. Optim.*, 2015.

Main conclusions and contributions

- Derivative-free optimization can be combined with probabilistic tools.
- Convergence can be maintained.
- Practical performance is enhanced in the direct-search case.
- Complexity confirms the numerical observations.

Direct search based on probabilistic descent
Gratton, Royer, Vicente and Zhang, *SIAM J. Optim.*, 2015.

- Second-order convergence can be ensured in the deterministic case.
- First complexity result for second order in DFO.
- Reveals worst-case cost of such guarantees.

A second-order globally convergent direct-search method and its worst-case complexity
Gratton, Royer and Vicente, *Optimization*, 2016.

Current and future work

Short-term perspectives of the manuscript

- MATLAB implementation of direct search using probabilistic descent
  Probabilistic treatment of bounds and linear constraints.
- De-coupled techniques for second-order convergent methods
  Ease the introduction of random aspects.

Challenges

- Probabilistic second-order properties in DFO.
- Probabilistic second-order derivative-based methods.

Thank you for your attention !

Thank you for your attention !

WCC for probabilistic descent

Intuitive idea

Let $G_k = \nabla f(X_k)$, so $Z_k = \mathbf{1}\left(\mathrm{cm}(\mathfrak{D}_k, -G_k) \geq \kappa\right)$.

- If $Z_k = 1$ and k unsuccessful, then $\kappa \|G_k\| < \mathcal{O}(\mathcal{A}_k)$...

WCC for probabilistic descent

Intuitive idea

Let $G_k = \nabla f(X_k)$, so $Z_k = \mathbf{1}\left(\mathrm{cm}(\mathfrak{D}_k, -G_k) \geq \kappa\right)$.

- If $Z_k = 1$ and k unsuccessful, then $\kappa \|G_k\| < \mathcal{O}(\mathcal{A}_k)$...
- ...so if $\inf_{0 \leq l \leq k} \|G_l\|$ has not decreased much, $\sum_{l=0}^{k} Z_l$ should not be too high.

WCC for probabilistic descent

Intuitive idea

Let $G_k = \nabla f(X_k)$, so $Z_k = \mathbf{1}\left(\mathrm{cm}(\mathfrak{D}_k, -G_k) \geq \kappa\right)$.

- If $Z_k = 1$ and k unsuccessful, then $\kappa \|G_k\| < \mathcal{O}(\mathcal{A}_k)$...
- ...so if $\inf_{0 \leq l \leq k} \|G_l\|$ has not decreased much, $\sum_{l=0}^{k} Z_l$ should not be too high.

A useful bound

For all realizations of the algorithm, one has

$$\sum_{l=0}^{k} z_l \;\leq\; \mathcal{O}\left(\frac{1}{\kappa^2 \, \|\tilde{g}_k\|^2}\right) + p_0 \, k,$$

with $\|\tilde{g}_k\| = \inf_{0 \leq l \leq k} \|g_l\|$.

WCC for probabilistic descent

We use again $Z_l = \mathbf{1}\left(\mathrm{cm}(\mathfrak{D}_l, -\nabla f(X_l) \geq \kappa\right)$.

An inclusion argument

$$\left\{ \inf_{0 \leq l \leq k} \|\nabla f(X_k)\| \geq \epsilon \right\} \subset \left\{ \sum_{l=0}^{k} Z_l \;\leq\; \lambda\, k \right\}$$

with $\lambda = \mathcal{O}\left(\frac{1}{k\,\kappa^2\,\epsilon^{-2}}\right) + p_0$.

A Chernoff-type probability result

For any $\lambda \in (0, p)$,

$$\mathbb{P}\left(\sum_{l=0}^{k-1} Z_l \;\leq\; \lambda\, k \right) \;\leq\; \exp\left[-\frac{(p-\lambda)^2}{2\,p} k \right].$$

WCC for probabilistic descent (3)

Probabilistic worst-case complexity

Let $\{\mathfrak{D}_k\}$ be (p,κ)-descent, $\epsilon \in (0,1)$ and N_ϵ the number of function evaluations needed to have $\inf_{0\le l\le k} \|\nabla f(X_l)\| \le \epsilon$. Then

$$\mathbb{P}\left(N_\epsilon \le \mathcal{O}\left(\frac{m\,(\kappa\epsilon)^{-2}}{p-p_0}\right)\right) \ge 1 - \exp\left(-\mathcal{O}\left(\frac{p-p_0}{p}\kappa^{-2}\epsilon^{-2}\right)\right).$$

Corollary

Using 2 uniformly distributed directions at every iteration, with $\gamma = \theta^{-1} = 2$, one has

$$\mathbb{P}\left(N_\epsilon \le \tfrac{32}{3}\left(f(x_0) - f_{\text{low}} + \tfrac{\alpha_0^2}{2}\right)\tfrac{(2+\nu)^2}{(2p-1)\tau^2}\,n\epsilon^{-2}\right)$$
$$\ge 1 - \exp\left[-\tfrac{1}{6}\left(f(x_0) - f_{\text{low}} + \tfrac{\alpha_0^2}{2}\right)\tfrac{(2p-1)(2+\nu)^2}{p\tau^2}\,n\epsilon^{-2}\right].$$

Looking at the second-order Taylor model

Let x such that $\|\nabla f(x)\| \neq 0, \lambda_{\min}\left(\nabla^2 f(x)\right) < 0$, and $\alpha > 0$.

Problem

Characterize the directions $d \in \mathbb{R}^n, \|d\| = 1$ for which the quadratic Taylor expansion

$$\alpha \nabla f(x)^\top d + \frac{\alpha^2}{2} d^\top \nabla^2 f(x)\, d$$

gives information on $\lambda = \lambda_{\min}\left(\nabla^2 f(x)\right)$.

A generic good direction when $\lambda < 0$

Looking for d satisfying:

$$\mathbb{P}\left(c_1\,\alpha\,\nabla f(x)^\top d + \frac{\alpha^2}{2} d^\top \nabla^2 f(x)\,d \;\leq\; c_2 \frac{\alpha^2}{2}\,\lambda + c_3\,\alpha^3 \right) \geq p.$$

A generic good direction when $\lambda < 0$

Looking for d satisfying:

$$\mathbb{P}\left(\frac{\alpha^2}{2} d^\top \nabla^2 f(x)\, d \;\leq\; c_2 \frac{\alpha^2}{2} \lambda \right) \geq p.$$

- $c_1 = c_3 = 0, c_2 \in (0,1)$ (Negative curvature direction)
  Gets harder as $\lambda \nearrow 0$.

A generic good direction when $\lambda < 0$

Looking for d satisfying:

$$\mathbb{P}\left(\qquad \frac{\alpha^2}{2} d^\top \nabla^2 f(x)\, d \;\leq\; c_2 \frac{\alpha^2}{2} \lambda + c_3\, \alpha^3 \right) \geq p.$$

- $c_1 = c_3 = 0, c_2 \in (0,1)$ (Negative curvature direction)
  Gets harder as $\lambda \nearrow 0$.
- $c_1 = 0, c_2 \in (0,1), c_3 > 0$ (Approx. Negative curvature direction)
  Ok but expensive.

A generic good direction when $\lambda < 0$

Looking for d satisfying:

$$\mathbb{P}\left(c_1\,\alpha\,\nabla f(x)^\top d + \frac{\alpha^2}{2} d^\top \nabla^2 f(x)\,d \;\leq\; c_2 \frac{\alpha^2}{2}\lambda + c_3\,\alpha^3\right) \geq p.$$

- $c_1 = c_3 = 0, c_2 \in (0,1)$ (Negative curvature direction)
  Gets harder as $\lambda \nearrow 0$.
- $c_1 = 0, c_2 \in (0,1), c_3 > 0$ (Approx. Negative curvature direction)
  Ok but expensive.
- $c_1, c_2 \in (0,1), c_3 > 0$ *(Approx. second-order direction)*
  Cheap but depends on α.

A generic good direction when $\lambda < 0$

Looking for d satisfying:

$$\mathbb{P}\left(c_1 \, \alpha \, \nabla f(x)^\top d + \frac{\alpha^2}{2} d^\top \nabla^2 f(x) \, d \; \leq \; c_2 \frac{\alpha^2}{2} \lambda + c_3 \, \alpha^3 \right) \geq p.$$

- $c_1 = c_3 = 0, c_2 \in (0,1)$ (Negative curvature direction)
  Gets harder as $\lambda \nearrow 0$.
- $c_1 = 0, c_2 \in (0,1), c_3 > 0$ (Approx. Negative curvature direction)
  Ok but expensive.
- $c_1, c_2 \in (0,1), c_3 > 0$ *(Approx. second-order direction)*
  Cheap but depends on α.