Direct search using probabilistic descent

Clément W. Royer
ENSEEIHT-IRIT, Toulouse, France

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

October 2nd, 2014
APO PhD student Day

Outline

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 exists and is 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 exists and is Lipschitz continuous.

Solving the problem using the derivative

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

- Steepest descent method;
- Gradient-related methods.

Introductory assumptions and definitions

Derivative-Free Optimization (DFO)

- Assumes that the gradient is unavailable (Ex: simulation code);
- Two main classes:
  - Model-based methods;
  - Direct-search methods.

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

Introductory assumptions and definitions

Derivative-Free Optimization (DFO)

- Assumes that the gradient is unavailable (Ex: simulation code);
- Two main classes:
  - Model-based methods;
  - Direct-search methods.

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

In this talk, we look at directional direct-search methods.

Optimization by Direct Search: New Perspectives on Some Classical and Modern Methods
T.G. Kolda, R.M. Lewis, V. Torczon (2003).

Outline

A basic framework for direct-search algorithms

1. **Initialization:** Set $x_0, \alpha_0, \theta < 1 \leq \gamma$.
2. **For** $k = 0, 1, 2, \ldots$
   - Choose a set D_k of m unitary 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 $x_0, \alpha_0, \theta < 1 \leq \gamma$.
2. **For** $k = 0, 1, 2, \ldots$
   - Choose a set D_k of m unitary 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 direction sets/polling sets D_k sufficiently good to ensure convergence :

- How do we know that a set is good ?
- How to relate set quality and convergence ?

Polling choice in deterministic direct search

We would like to choose direction sets/polling sets D_k sufficiently good to ensure convergence :

- How do we know that a set is good ?
- How to relate set quality and convergence ?

A measure of set quality

Let D be a set of unitary vectors. Then

$$\mathrm{cm}(D) = \min_{\|v\|=1} \max_{d \in D} d^T v$$

is the cosine measure of D.

Polling choice in deterministic direct search

We would like to choose direction sets/polling sets D_k sufficiently good to ensure convergence :

- How do we know that a set is good ?
- How to relate set quality and convergence ?

A measure of set quality

Let D be a set of unitary vectors. Then

$$\mathrm{cm}(D) = \min_{\|v\|=1} \max_{d \in D} d^T v$$

is the cosine measure of D.

Assumption

It exists $\kappa > 0$ *such that* $\forall k,\ \mathrm{cm}(D_k) \geq \kappa$.
Any vector (e.g. $-\nabla f(x_k)$*) is then close to an element of* D_k.

Set quality

A common choice is to use positive spanning sets.

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.

Set quality

A common choice is to use positive spanning sets.

Positive Spanning Sets (PSS)

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

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

Example

$D_{\oplus} = [I \ -I]$ is a PSS with

$$\operatorname{cm}(D_{\oplus}) = \frac{1}{\sqrt{n}}.$$

Convergence for deterministic direct search

Lemma

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

Proposition

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

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

Convergence result

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

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

From global convergence to worst-case complexity

Convergence analysis tells us that $\liminf \|\nabla f(x_k)\| \to 0$, but not how.

Worst-case complexity

Estimating the number of evaluations of f needed to reach

$$\inf_{0 \leq l \leq k} \|\nabla f(x_l)\| \leq \epsilon.$$

Worst-case complexity of direct search
L. N. Vicente (2013)

Worst-case complexity in deterministic direct-search

Theorem (Vicente - 2013)

Let N_ϵ be the number of function evaluations needed to reduce the gradient norm below $\epsilon \in (0,1)$; then

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

with $m \geq n+1$.

Corollary

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

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

Outline

Introducing randomness

Idea (Gratton, Vicente - 2013)

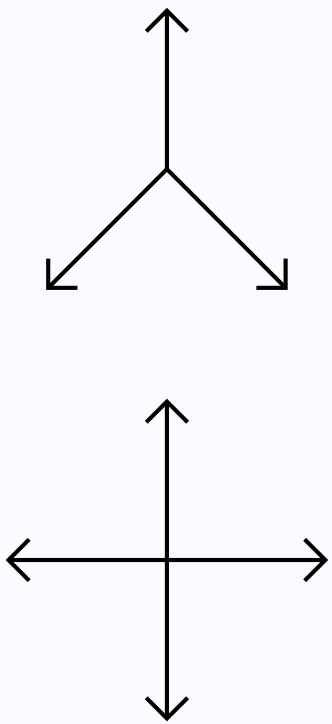
Randomly independently generate polling sets, possibly of less than $n + 1$ vectors!

Introducing randomness

Idea (Gratton, Vicente - 2013)

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

From PSS...

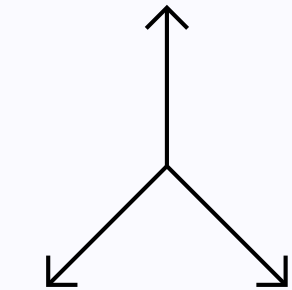


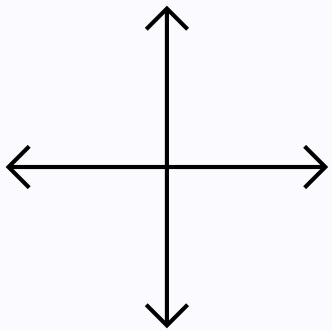
Introducing randomness

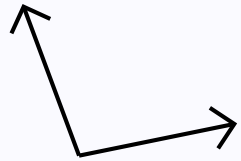
Idea (Gratton, Vicente - 2013)

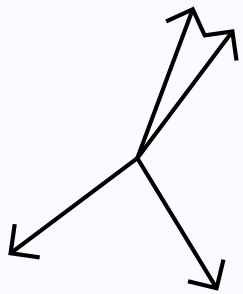
Randomly independently generate polling sets, possibly of less than $n+1$ vectors!

From PSS...









...to random sets

Numerical motivations

- Some results for $n = 40$ and $\epsilon = 10^{-3}$:

Problem	$[I\ -I]$	$[Q\ -Q]$	$2\,n$	$n+1$	$n/2$	2	1
arglina	3.42	8.44	10.30	6.01	3.21	1.00	–
arglinb	20.50	10.35	7.38	2.81	2.35	1.00	2.04
broydn3d	4.33	6.55	6.54	3.59	2.04	1.00	–
dqrtic	7.16	9.37	9.10	4.56	2.77	1.00	–
engval1	10.53	20.89	11.90	6.48	3.55	1.00	2.08
freuroth	56.00	6.33	1.00	1.67	1.33	1.00	4.00
integreq	16.04	16.29	12.44	6.76	3.52	1.00	–
nondquar	6.90	30.23	7.56	4.23	2.76	1.00	–
sinquad	–	–	1.31	1.00	1.60	1.23	–
vardim	1.00	3.80	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)

A probabilistic direct-search algorithm

From deterministic to probabilistic notations

- Polling sets : $D_k \to \mathfrak{D}_k$;
- Iterates : $x_k \to X_k$;
- Step sizes : $\alpha_k \to \mathcal{A}_k$.

A probabilistic direct-search algorithm

From deterministic to probabilistic notations

- Polling sets : $D_k \to \mathfrak{D}_k$;
- Iterates : $x_k \to X_k$;
- Step sizes : $\alpha_k \to \mathcal{A}_k$.

1. **Initialization:** Set $x_0, \alpha_0, \theta < 1 \leq \gamma$.
2. **For** $k = 0, 1, 2, ...,$
   - Choose a set $\mathfrak{D}_k$ of m unitary 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

Study of the probabilistic framework

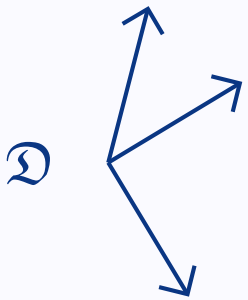
Theoretical questions

- Can we prove that it converges ?
  Global Convergence
- Can we bound the number of evaluations of f needed to reach a tolerance ϵ ?
  Worst-Case Complexity

The main issue is to find the adequate probabilistic tools to obtain such results.

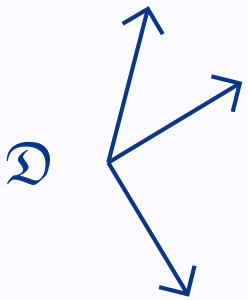
A new measure of set quality

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

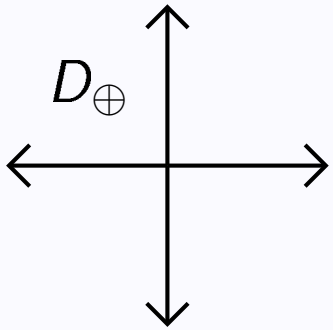


A new measure of set quality

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

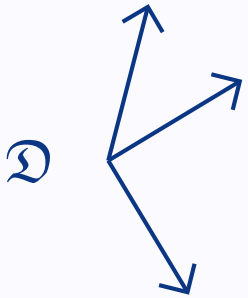


...$D_\oplus$ is...

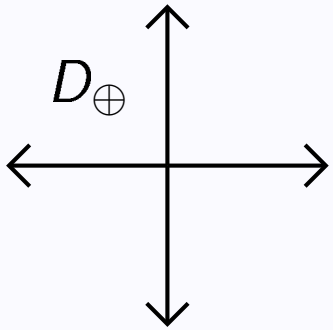


A new measure of set quality

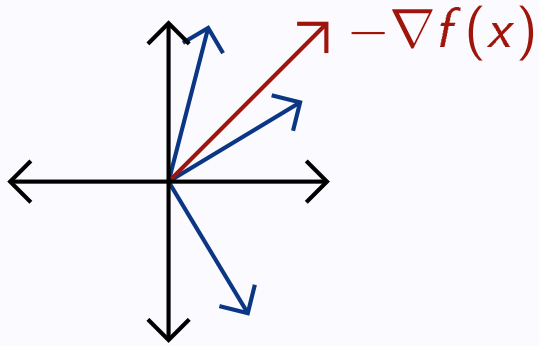
$\mathfrak{D}$ is not a PSS...



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



...but here $-\nabla f(x)$ is closer to $\mathfrak{D}$!



A new measure of set quality

Set assumption in the deterministic case

- We required:

$$\mathrm{cm}(D_k) = \min_{\|v\|=1} \max_{d \in D_k} d^T v \geq \kappa.$$

- Yet we only used:

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

In the random case, the second one might happen with some probability.

(p, κ)-descent sets

- We want to look at $\mathbb{P}\left(\mathrm{cm}\left(\mathfrak{D}_k, -\nabla f(X_k)\right) \geq \kappa\right)$, but X_k depends on $\mathfrak{D}_0, \ldots, \mathfrak{D}_{k-1}$.

Convergence on trust-region methods based on probabilistic models
A.S. Bandeira, K. Scheinberg, L.N. Vicente. (2014)

(p, κ)-descent sets

- We want to look at $\mathbb{P}\left(\mathrm{cm}\left(\mathfrak{D}_k, -\nabla f(X_k)\right) \geq \kappa\right)$, but X_k depends on $\mathfrak{D}_0, \ldots, \mathfrak{D}_{k-1}$.

Convergence on trust-region methods based on probabilistic models
A.S. Bandeira, K. Scheinberg, L.N. Vicente. (2014)

Probabilistic descent property

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

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

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

Convergence of probabilistic descent

Lemma

For all realizations $\{\alpha_k\}$ of $\{\mathcal{A}_k\}$:

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

Convergence Theorem

If $\{\mathfrak{D}_k\}$ is (p, κ)-descent with $p \geq \ln(\theta)\,\ln(\theta/\gamma)^{-1}$, then

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

Sketch of the proof

Two main ideas:

Lemma

If k is an unsuccessful iteration; then

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

Lemma

Let $Z_k = \mathbf{1}\,(\mathrm{cm}\,(\mathfrak{D}_k, -\nabla f(X_k)) \geq \kappa)$*; then*

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

is a submartingale and $\mathbb{P}\,(\limsup S_k = \infty) = 1$.

Concerning the probability p

In order to ensure global convergence, one must assume:

$$p \geq p_0 = \frac{\ln(\theta)}{\ln(\theta/\gamma)}.$$

This induces a lower bound on $m = |\mathfrak{D}_k|$.

A practical example: uniform distribution over the unit sphere

In that case, $\mathfrak{D}_k$ is (p_0, κ)-descent if

$$m \geq \ln\left(1 - \frac{\ln\theta}{\ln(\theta/\gamma)}\right) \ln\left(1 - \frac{1}{2} B_{1-\kappa^2}\left(\frac{n-1}{2}, \frac{1}{2}\right)\right)^{-1}.$$

where $B_x(a, b)$ is the incomplete Beta function.

WCC for probabilistic descent

Intuitive idea

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

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

WCC for probabilistic descent

Intuitive idea

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

- If $Z_k = 1$ and k unsuccessful, then $\|G_k\| < \mathcal{O}(\mathcal{A}_k)$...
- ...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 $G_k \overset{n}{=} \nabla f(X_k)$, so $Z_k = \mathbf{1}\left(\mathrm{cm}(\mathfrak{D}_k, -G_k) \geq \kappa\right)$.

- If $Z_k = 1$ and k unsuccessful, then $\|G_k\| < \mathcal{O}(\mathcal{A}_k)$...
- ...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 (2)

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 $\|\tilde{G}_k\| \leq \epsilon$. Then

$$\mathbb{P}\left(N_\epsilon \leq \mathcal{O}(m\,(\kappa\epsilon)^{-2})\right) \geq 1 - \exp\left(-\mathcal{O}\left(\epsilon^{-2}\right)\right).$$

WCC for probabilistic descent (2)

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 $\|\tilde{G}_k\| \leq \epsilon$. Then

$$\mathbb{P}\left(N_\epsilon \leq \mathcal{O}(m\,(\kappa\epsilon)^{-2})\right) \geq 1 - \exp\left(-\mathcal{O}\left(\epsilon^{-2}\right)\right).$$

- By taking $\mathfrak{D}_k = D_\oplus$, one has $\kappa = 1/\sqrt{n}$, $m = 2\,n$ and $p = 1$, we recover:
$$\mathcal{O}(n^2\,\epsilon^{-2}).$$
- With uniform generation, one can decrease this rate to $\mathcal{O}(m\,n\,\epsilon^{-2})$, with possibly $m \ll n+1$!

Summary

What comes out from our study ?

- A new method that converges without using PSS;

Summary

What comes out from our study ?

- A new method that converges without using PSS;
- A new probabilistic worst-case complexity argument, adaptable to other DFO methods (ex: Trust-Region);

Summary

What comes out from our study ?

- A new method that converges without using PSS;
- A new probabilistic worst-case complexity argument, adaptable to other DFO methods (ex: Trust-Region);
- Improved numerical performance.

To be continued

The paper

Direct Search based on Probabilistic Descent.
S. Gratton, C. W. Royer, L. N. Vicente, Z. Zhang.
Submitted, available on www.optimization-online.org.

What is next ?

- Extension to nonsmooth and constrained problems;
- Second-order results and probabilistic assumptions.

Thank you for your attention !