Direct search using probabilistic descent

Clément Royer

ENSEEIHT-IRIT

Journée des doctorants APO
13/09/2013

Outline

Introductory definitions and assumptions

We consider an unconstrained smooth problem:

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

Assumptions on f

- f smooth;
- ∇f Lipschitz continuous of constant ν.

Derivative-Free Optimization

Reasons not to use the derivatives

- Unavailable (ex: blackbox functions) ;
- Too expensive for computation.

Derivative-Free Optimization

Reasons not to use the derivatives

- Unavailable (ex: blackbox functions) ;
- Too expensive for computation.

In Derivative-Free Methods, only the information of order 0 is used.

The number of function evaluations is a measure of cost.

Direct search

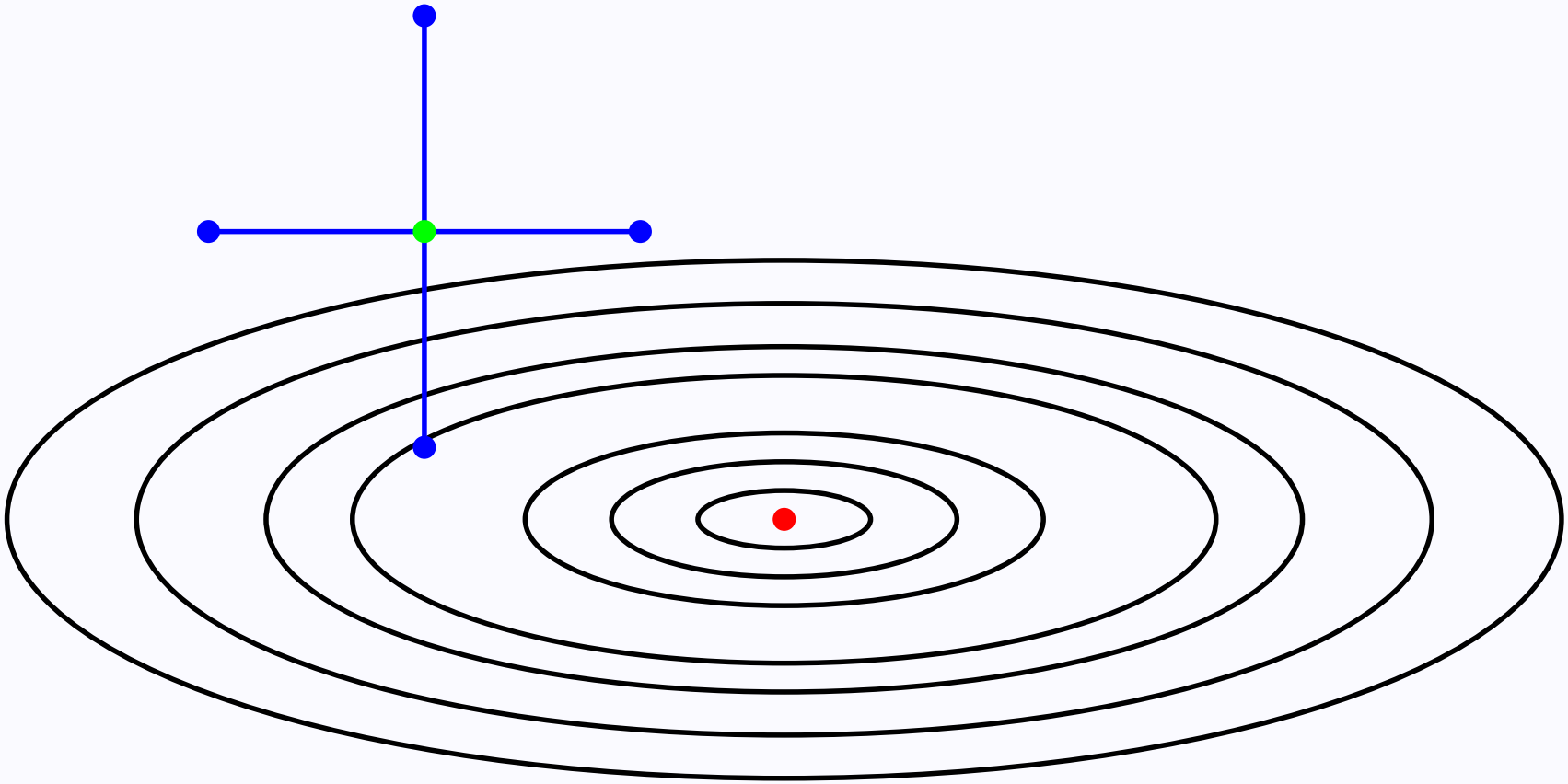
Two main types of DFO algorithms

- Model-based methods : build an approximation of f (ex : trust regions) ;
- Directional methods : evaluate f in some directions from the current point (ex : direct search).

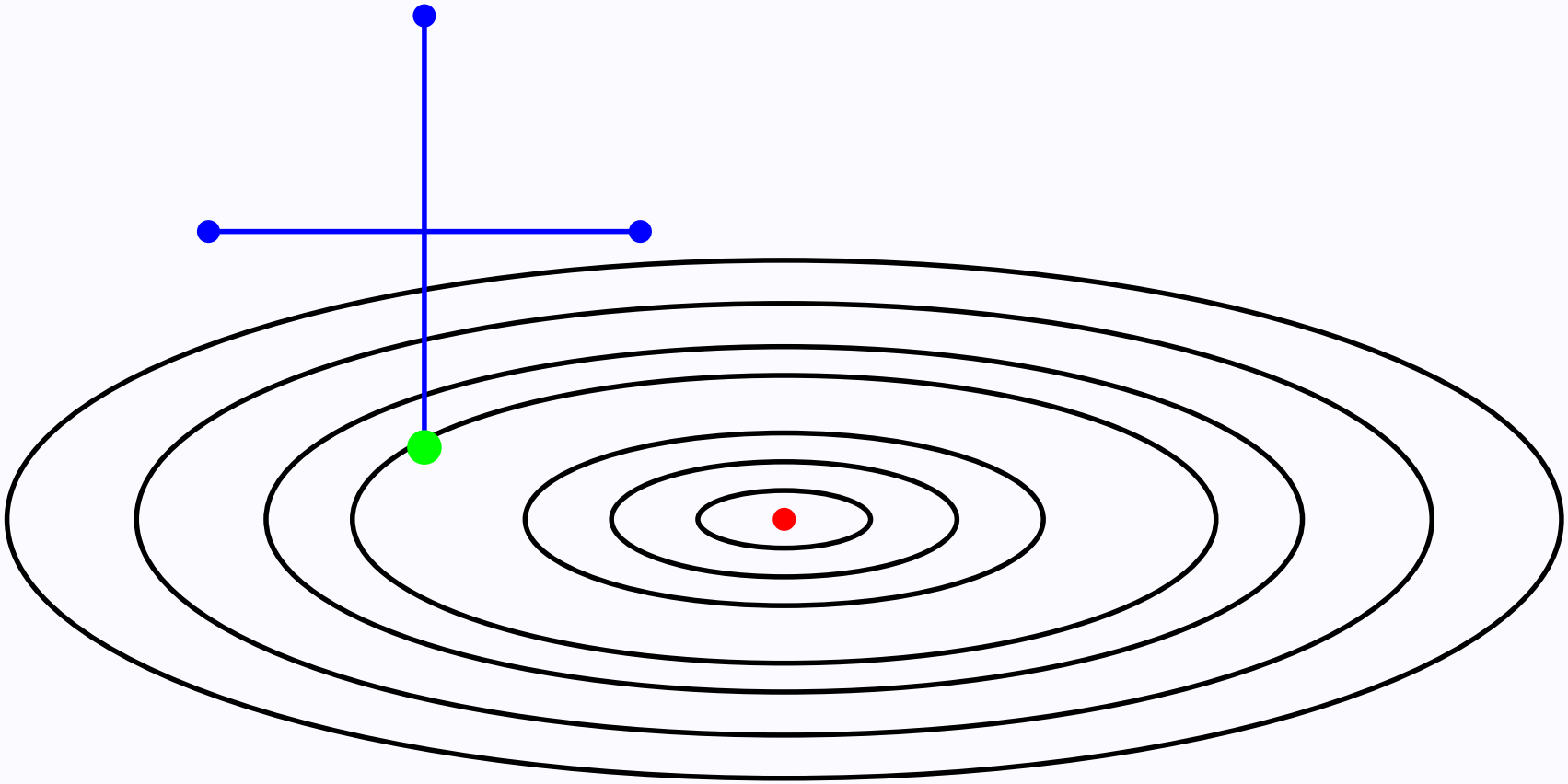
We will focus on Direct Search methods.

Reference : *Kolda, Lewis, Torczon - SIAM Review, 2003*

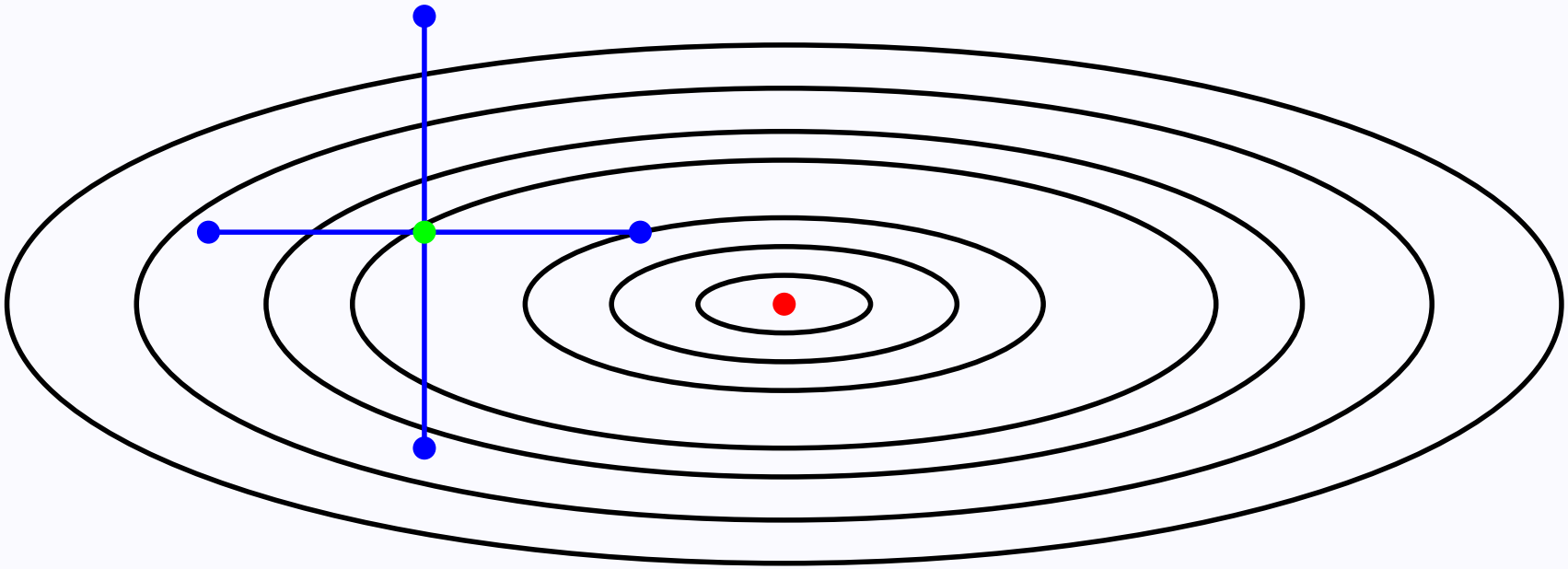
An example of DS : Coordinate Search



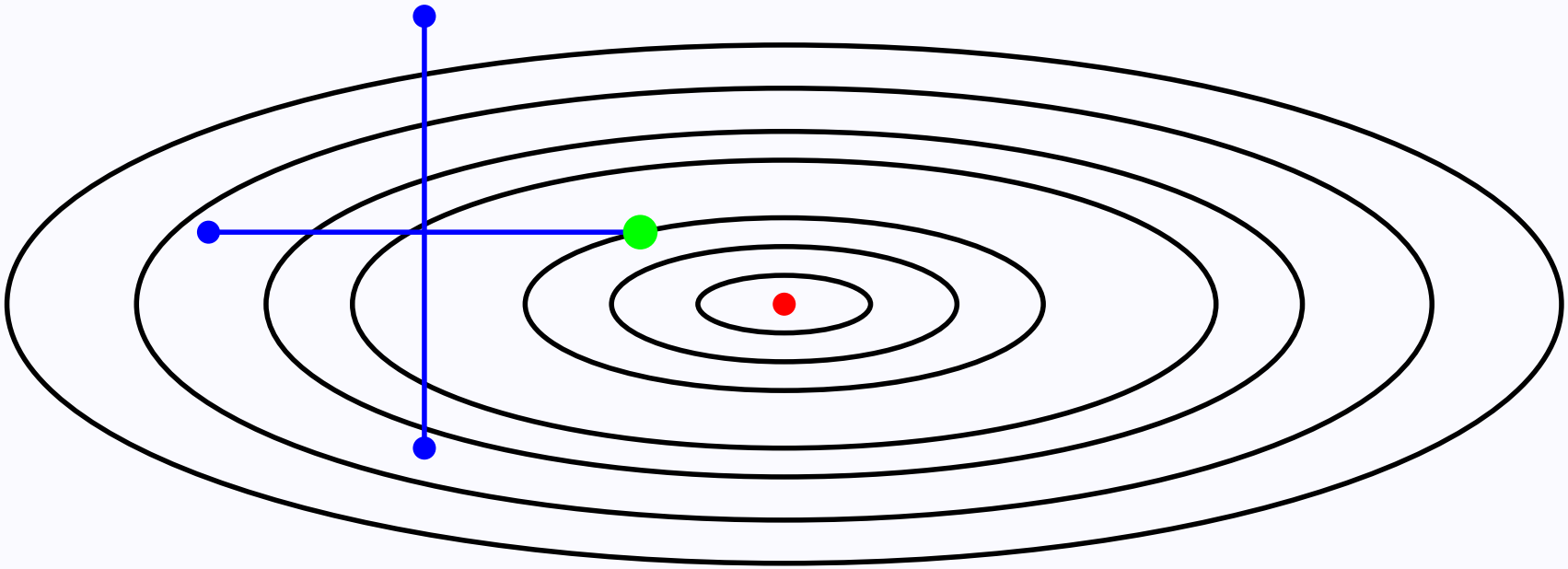
An example of DS : Coordinate Search



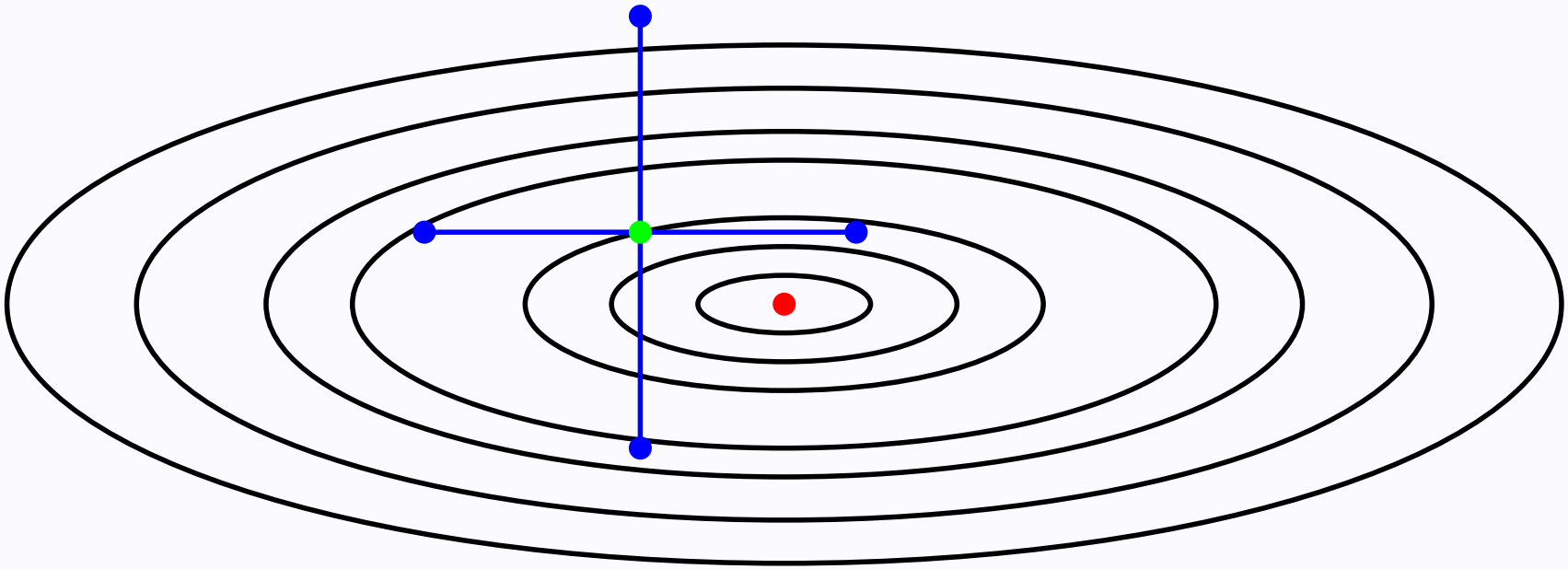
An example of DS : Coordinate Search



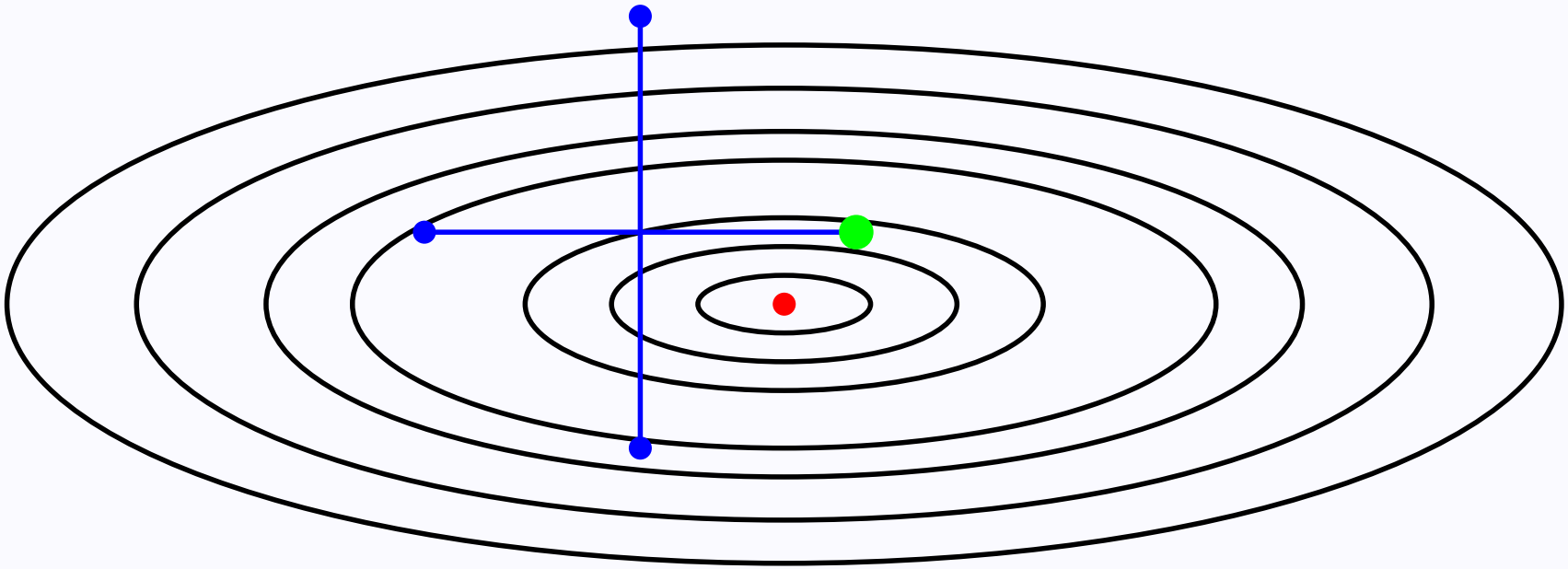
An example of DS : Coordinate Search



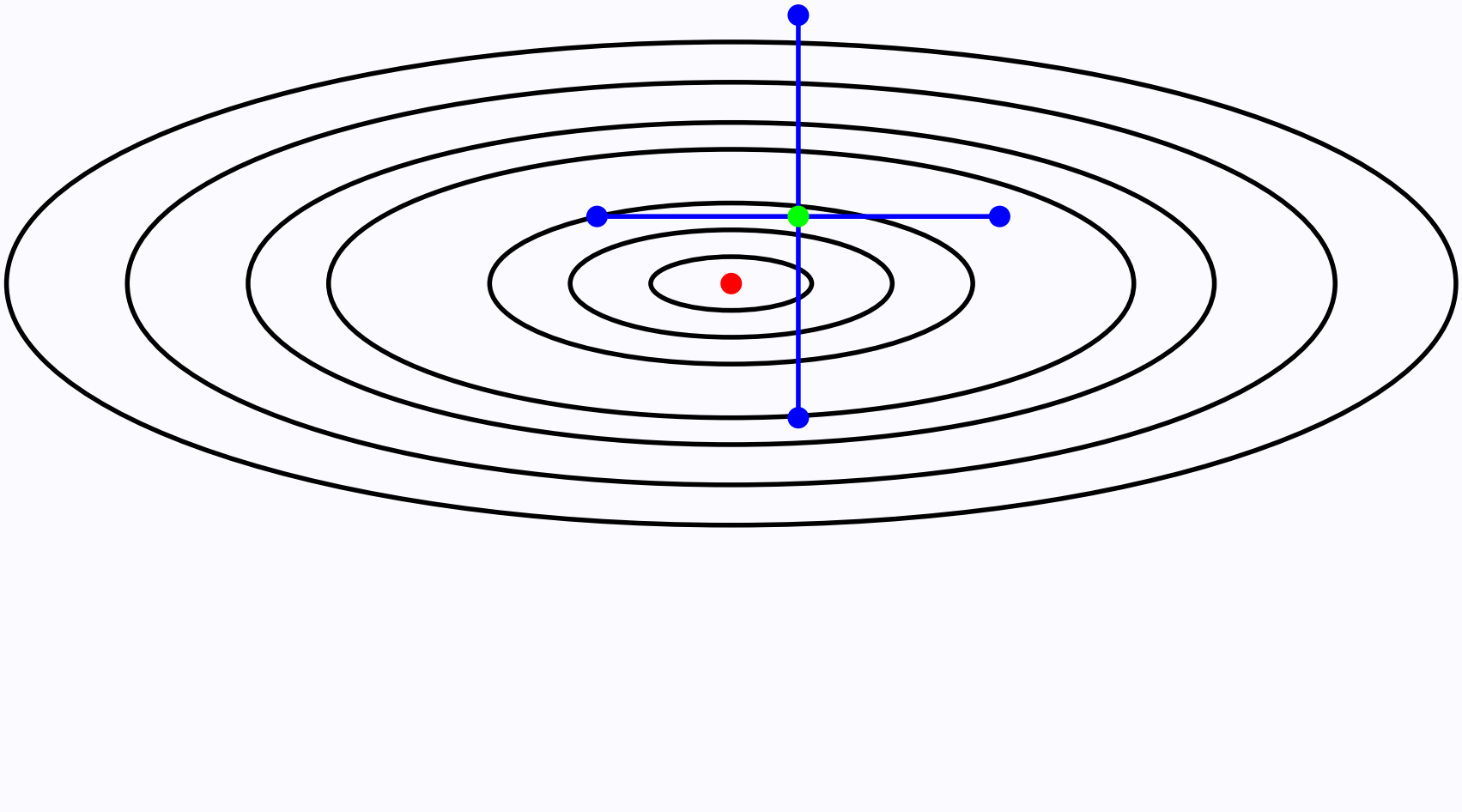
An example of DS : Coordinate Search



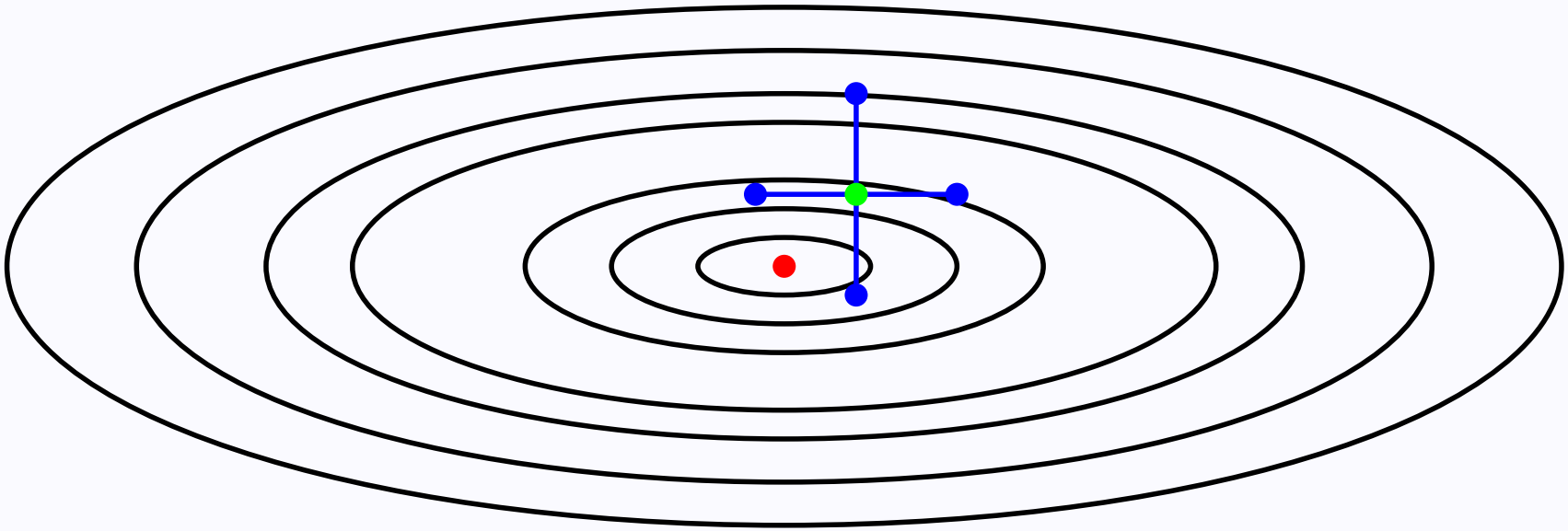
An example of DS : Coordinate Search



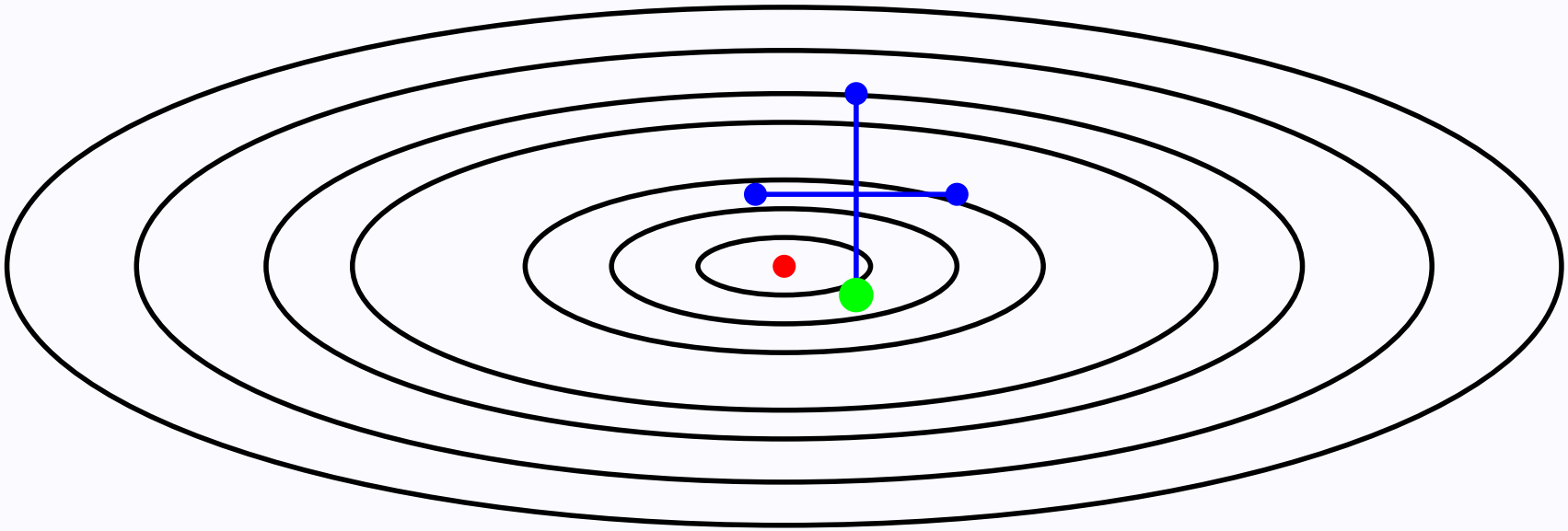
An example of DS : Coordinate Search



An example of DS : Coordinate Search



An example of DS : Coordinate Search



A basic framework for direct-search algorithms

Forcing function

- $\rho : [0, \infty[\rightarrow [0, \infty[$ monotonically nondecreasing
- $\rho(t) = o(t)$ when $t \rightarrow 0$.

A basic framework for direct-search algorithms

Forcing function

- $\rho : [0, \infty[\to [0, \infty[$ monotonically nondecreasing
- $\rho(t) = o(t)$ when $t \to 0$.

1. **Initialization:** Set $x_0, \alpha_0, \gamma \geq 1, \theta < 1$.

A basic framework for direct-search algorithms

Forcing function

- $\rho : [0, \infty[\to [0, \infty[$ monotonically nondecreasing
- $\rho(t) = o(t)$ when $t \to 0$.

1. **Initialization:** Set $x_0, \alpha_0, \gamma \geq 1, \theta < 1$.
2. **For** $k = 0, 1, 2, \ldots$, while α_k is not too small:
   - Choose a set D_k of m *unitary* vectors.
   - If it exists $d_k \in D_k$ so that

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

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

Main concern in deterministic direct search

We aim to choose good direction sets/polling sets D_k...but how ?

Main concern in deterministic direct search

We aim to choose good direction sets/polling sets D_k...but how ?

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.

Main concern in deterministic direct search

We aim to choose good direction sets/polling sets D_k...but how ?

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*

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

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 positive linear combinations.

- a PSS contains at least $n+1$ vectors ;
- if D is a PSS, then $\text{cm}(D) > 0$.

The following analysis will be driven considering PSS.

Convergence for deterministic direct search

Assumption

- $\mathcal{L}(x_0) = \{x : f(x) \leq f(x_0)\}$ *is bounded from below ;*
- $\rho \neq 0$*:* *sufficient decrease*. *For simplicity, assume* $\rho(\alpha) = \alpha^2$.

Convergence for deterministic direct search

Assumption

- $\mathcal{L}(x_0) = \{x : f(x) \leq f(x_0)\}$ *is bounded from below ;*
- $\rho \neq 0$*:* *sufficient decrease*. *For simplicity, assume* $\rho(\alpha) = \alpha^2$.

Lemma

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

Proposition

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

$$\|\nabla f(x_k)\| \leq \frac{\mathcal{C}(\nu)\, \alpha_k}{\mathrm{cm}(D_k)}.$$

Convergence for deterministic direct search

Theorem

Consider the sequence $\{x_k\}$ generated by our algorithm.

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

Convergence for deterministic direct search

Theorem

Consider the sequence $\{x_k\}$ generated by our algorithm.

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

In some cases, it is possible to ensure:

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

Global rates in deterministic direct search

For simplicity, we still suppose $\rho(t) = t^2$.

Theorem (Vicente - 2012)

To reduce the gradient norm below $\epsilon \in (0, 1)$, the algorithm takes at most

$$\mathcal{O}(\nu\, n\, \epsilon^{-2})$$

iterations.

Corollary

To reduce the gradient norm below ϵ, the algorithm costs at most $\mathcal{O}(\nu\, n^2\, \epsilon^{-2})$ function evaluations.

Previous randomizing approaches

Objective

Introducing randomness in direct-search methods *through the direction sets*

Previous randomizing approaches

Objective

Introducing randomness in direct-search methods *through the direction sets*

Several ideas

- Generate directions asymptotically dense in the unit sphere
  `MADS` *Audet, Dennis - 2006*
  Discontinuous functions: *Vicente, Custódio - 2012*
- Use random oracles (one gradient-like direction per iteration)
  Nesterov - 2011

Our randomizing framework

Idea (Gratton, Vicente - 2013)

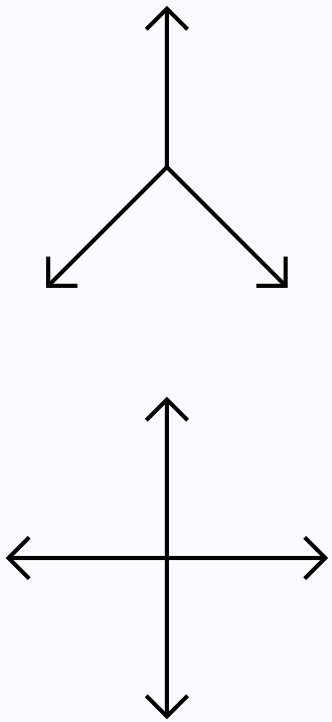
Randomly generate *independent* polling sets, possibly of less than $n+1$ vectors!

Our randomizing framework

Idea (Gratton, Vicente - 2013)

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

From PSS...

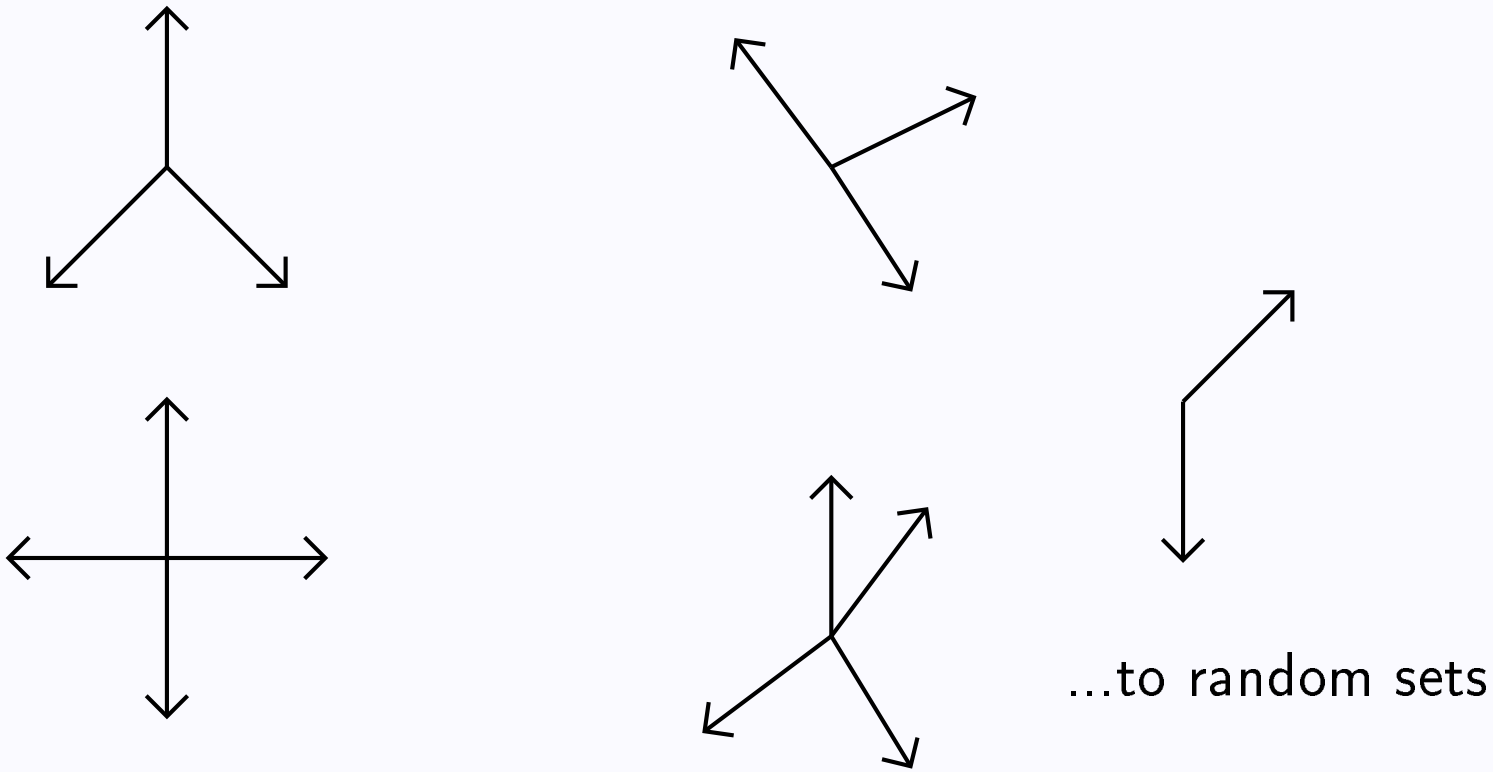


Our randomizing framework

Idea (Gratton, Vicente - 2013)

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

From PSS...



...to random sets

Numerical motivations

Problem	$[Q\ -Q]$	$[I\ -I]$	$[Qk\ -Qk]$	$n/2$	$n/4$	2	1
arglina	4.87	1.79	8.25	1.98	1.19	1	4.17 (93)
arglinb	8.74	8.96	5.93	1.81	1.15	1	3.44 (6.7)
arwhead	2.76	0.08	7.09	1.95	1.29	1	3.95 (97)
bdqrtic	7.19	4.28	9.86	1.96	1.23	1	4.55 (80)
broydn3d	4.08	2.89	5.79 (3.3)	1.38	0.96	1	- (100)
dqrtic	5.16	3.55	9.22	1.73	1.17	1	5.59 (90)
engval1	8.23	4.56	10.08	1.81	1.19	1	4.38 (60)
freuroth	3.17	4.00	2.50	0.67	0.83	1	1.17
integreq	8.38	8.88	9.11	2.16	1.36	1	3.38 (97)
nondia	1.61	0.002	1.40	0.82	0.66	1	0.12 (3.3)
nondquar	14.07	3.71	6.27	1.36	1.02	1	- (100)
penalty1	4.97	4.78	9.12	1.95	1.27	1	6.65 (70)
penalty2	9.04	3.24	9.31	2.00	1.23	1	3.71 (67)
tquartic	1.18	- (100)	2.02 (13)	0.86	0.73	1	- (100)
vardim	2.06	0.38	2.69	1.13	1.00	1	6.81 (3.3)

Table: Ratio function evaluations with respect to the case $m = 2$ directions and percentage of unsuccessful runs when relevant; tolerance is 10^{-3}, $n = 20$, mean on 30 runs.

Our probabilistic direct-search algorithm

Random variables and realizations

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

Our probabilistic direct-search algorithm

Random variables and realizations

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

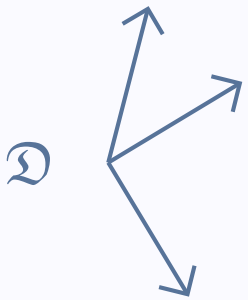
1. **Initialization:** Set $x_0, \alpha_0, \gamma > 1, \theta < 1$.
2. **For** $k = 0, 1, 2, \ldots$, while $\mathcal{A}_k$ is not too small:
   - Choose a set $\mathfrak{D}_k$ of m unitary vectors uniformly randomly distributed.
   - If it exists $\mathfrak{d}_k \in \mathfrak{D}_k$ so that

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

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

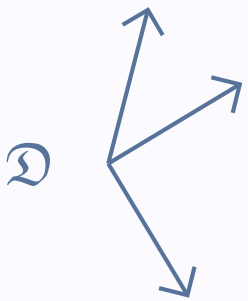
Random sets quality

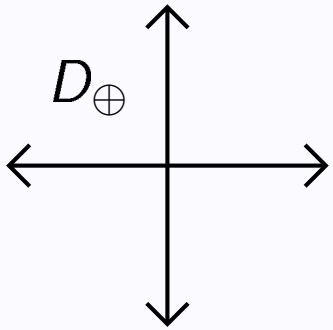
- $\mathfrak{D}$ is not a PSS...



Random sets quality

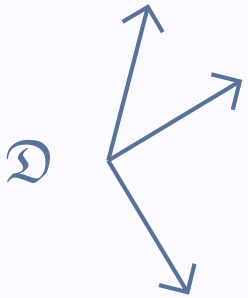
$\mathfrak{D}$ is not a PSS... ...$D_\oplus$ is...



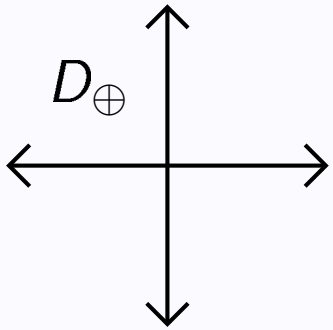


Random sets quality

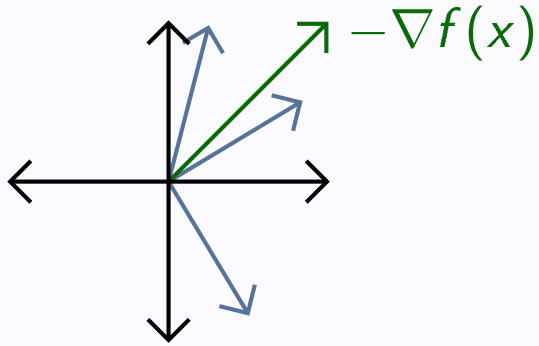
$\mathfrak{D}$ is not a PSS...



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

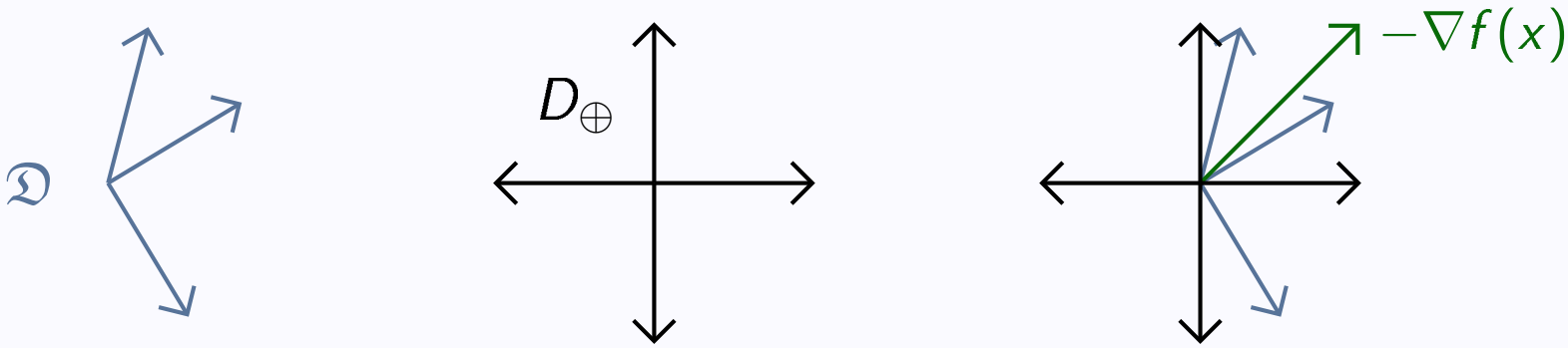


...but $\mathfrak{D}$ is closer to $-\nabla f(x)$!



Random sets quality

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



f is smooth $\Rightarrow$ there are descent directions near $-\nabla f(x)$.

What we really want is $\max_{d \in D} d^T \frac{-\nabla f(x)}{\|\nabla f(x)\|} > 0$.

Local cosine measure

Let $v \in \mathbb{R}^n$ and D a set of unitary vectors.
The local cosine measure of D at v is

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

- Having properties on the local cosine measure is easier;
- The properties will be probabilistic.

κ-descent sets

We want to look at $P\left[\text{cm}\left(\mathfrak{D}_k, -\frac{\nabla f(X_k)}{\|\nabla f(X_k)\|}\right) > \kappa\right]$. $\mathfrak{D}_k$ and X_k are independent random variables.
One solution is conditioning to the past.

- Bandeira, Scheinberg, Vicente, *Convergence on trust-region methods based on probabilistic models* - 2013.

κ-descent sets

We want to look at $P\left[\text{cm}\left(\mathfrak{D}_k, -\frac{\nabla f(X_k)}{\|\nabla f(X_k)\|}\right) > \kappa\right]$. $\mathfrak{D}_k$ and X_k are independent random variables.
One solution is conditioning to the past.

- Bandeira, Scheinberg, Vicente, *Convergence on trust-region methods based on probabilistic models* - 2013.

The κ-descent property

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

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

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 $\{\mathcal{A}_k\}$:

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

Theorem

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

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

Sketch of the proof

Two main ideas:

Lemma

If K is an unsuccessful iteration, then

$$P\left[\mathcal{A}_k \geq \mathcal{C}(\nu,\kappa)\,\|\nabla f(X_k)\|\right] \geq p.$$

Lemma

Let $F_k = \{\mathrm{cm}\,(\mathfrak{D}_k, -\nabla f(X_k)/\|\nabla f(X_k)\|) > \kappa\}$, and define:

$$Z_k = \sum_{i=0}^{k-1} \left(1 - \frac{\ln \gamma}{\ln \theta}\right) \cdot 1_{F_i} - 1.$$

Then $\{Z_k\}$ is a submartingale and $P\left[\limsup Z_k = \infty\right] = 1$.

Convergence of probabilistic descent (2)

Probabilistic lemma (Bandeira, Scheinberg, Vicente - 2013)

Let $\epsilon \in (0,1)$, $\{K_i\}_i = \{k \mid \|\nabla f(X_k)\| \geq \epsilon\}$ and $p \geq \ln(\theta)\ \ln(\theta/\gamma)^{-1}$.

$$P\left[\sum_i \mathcal{A}_{K_i}^2 < \infty\right] = 1.$$

Theorem

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

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

Global rates for probabilistic descent

Probabilistic formula

Let $\{\mathfrak{D}_k\}$ be (p, κ)-descent with $p \geq \ln(\theta)\, \ln(\theta/\gamma)^{-1}$, $\epsilon \in (0, 1)$ and $F_{ev}(\epsilon)$ the number of function evaluations needed to decrease $\|\nabla f(X_k)\|$ below ϵ. Then

$$P\left[F_{ev}(\epsilon) \leq \mathcal{O}(m\,\nu\,(\kappa\epsilon)^{-2}) \;\middle|\; \mathfrak{S}(I_\epsilon)\right] \geq 2\,p - 1 \geq \frac{\ln(\theta\,\gamma)}{\ln(\theta/\gamma)}.$$

where I_ϵ is the index of smallest progress.

Global rates for probabilistic descent

Probabilistic formula

Let $\{\mathfrak{D}_k\}$ be (p, κ)-descent with $p \geq \ln(\theta)\, \ln(\theta/\gamma)^{-1}$, $\epsilon \in (0, 1)$ and $F_{ev}(\epsilon)$ the number of function evaluations needed to decrease $\|\nabla f(X_k)\|$ below ϵ. Then

$$P\left[F_{ev}(\epsilon) \leq \mathcal{O}(m\,\nu\,(\kappa\epsilon)^{-2}) \;\middle|\; \mathfrak{S}(I_\epsilon)\right] \geq 2\,p - 1 \geq \frac{\ln(\theta\,\gamma)}{\ln(\theta/\gamma)}.$$

where I_ϵ is the index of smallest progress.

- Interpretation
  The global rate is a performance indicator.
  The probability bound expresses the robustness.
- Link with deterministic case

Global rates for probabilistic descent

Probabilistic formula

Let $\{\mathfrak{D}_k\}$ be (p, κ)-descent with $p \geq \ln(\theta)\, \ln(\theta/\gamma)^{-1}$, $\epsilon \in (0, 1)$ and $F_{ev}(\epsilon)$ the number of function evaluations needed to decrease $\|\nabla f(X_k)\|$ below ϵ. Then

$$P\left[F_{ev}(\epsilon) \leq \mathcal{O}\left(m\,\nu\,(\kappa\epsilon)^{-2}\right) \,\middle|\, \mathfrak{S}(I_\epsilon)\right] \geq 2\,p - 1 \geq \frac{\ln(\theta\,\gamma)}{\ln(\theta/\gamma)}.$$

where I_ϵ is the index of smallest progress.

- Interpretation
- Link with deterministic case
  By taking $\mathfrak{D}_k = D_\oplus$, one has $\kappa = 1/\sqrt{n}$, $m = 2\,n$ and $p = 1$, we recover:

$$\mathcal{O}(\nu\, n^2\, \epsilon^{-2})$$

Summary

What do we have?

- A proof that convergence is possible *with less than* $n+1$ *vectors* ;
- The theoretical basis for a randomized direct search method ;
- Clear improvement of the performance in practice.

Summary

What do we have?

- A proof that convergence is possible *with less than* $n+1$ *vectors* ;
- The theoretical basis for a randomized direct search method ;
- Clear improvement of the performance in practice.

- Robustness issues ;
- Conditioned (and weakly ensured) global rates.

To be continued

- Get rid of the past $\rightarrow$ closer to the numerical behaviour:

$$P\left[\mathrm{cm}\left(\mathfrak{D}_k, \frac{-\nabla f(X_k)}{\|\nabla f(X_k)\|}\right) > \kappa\right] \geq p;$$

- Improve the algorithm (ex: keep good directions);
- Extend the study to nonsmooth or noisy functions;
- Adapt it to constrained problems.

Thank you for your attention!