Probabilistic analysis of derivative-free methods

Clément Royer
IRIT, University of Toulouse, France

Based on joint works with S. Gratton, L. N. Vicente, Z. Zhang

Argonne National Laboratory - LANS Seminar
April 13, 2016

Introduction: Randomized optimization methods

- Randomness has triggered significant recent developments in numerical optimization.
- Multiple reasons:
  - Large-scale setting: Applying classical methods is too expensive;
  - Distributed computing: The data is not stored on a single computer/processor;
  - Applications: already popular in machine learning community.

Randomized optimization: methods

What methods does it refer to ?

- Mostly first-order algorithms:
  - Randomized Coordinate Descent;
  - Randomized Conditional Gradient/Frank-Wolfe;
  - Stochastic Gradient.
- Some exploit second-order aspects:
  - Randomized Newton;
  - Stochastic Quasi-Newton/BFGS/Variable metric.

Randomized optimization: methods

What methods does it refer to ?

- Mostly first-order algorithms:
  - Randomized Coordinate Descent;
  - Randomized Conditional Gradient/Frank-Wolfe;
  - Stochastic Gradient.
- Some exploit second-order aspects:
  - Randomized Newton;
  - Stochastic Quasi-Newton/BFGS/Variable metric.

- Distinction stochastic/randomized ?
- What about zero-th order/derivative-free methods ?

Randomized optimization: complexity analysis

Complexity of optimization methods

- Studies the convergence speed of some criterion: gradient norm, iterates variation, function value,...
- **Stochastic/Randomized**: bounds the rate of decrease in expectation, typically in $\mathcal{O}(\frac{1}{k^r})$;
- **Deterministic**: global rates (convex) /complexity bounds (nonconvex) Results in $\mathcal{O}(\epsilon^{-r})$;
- Recent advancies in zero-th order methods.

General questions:

- Comparison deterministic/randomized ?
- Practical relevance ?

Objectives of the talk

- Illustrate how to randomize a classical derivative-free method;
- Show that it can be beneficial for both theory and practice;
- Generalize to other methods and higher-order aspects.

Outline

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 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 *first-order/gradient* methods.

The derivative-free/zero-th order context

Although the gradient exists, we assume it is unavailable.

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

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

Derivative-Free Optimization (DFO) algorithms

Deterministic DFO methods

- Inspired by the classical methods in nonlinear optimization;
- Model-based methods $\sim$ Trust-Region, Levenberg-Marquardt,...
- Directional methods $\sim$ Line Search, Gradient Descent,...

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

Key concerns

- Well-established: convergence theory.
- Recent advancies: 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 talk will NOT address those methods;
- Distinction: stochastic VS randomized algorithm.

Randomized DFO methods

A category between stochastic and deterministic

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

Randomized DFO methods

A category between stochastic and deterministic

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

Motivation:

- In DFO, we lack the knowledge of the derivative;

Randomized DFO methods

A category between stochastic and deterministic

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

Motivation:

- In DFO, we lack the knowledge of the derivative;
- Probability is the theory of lack of information;

Randomized DFO methods

A category between stochastic and deterministic

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

Motivation:

- In DFO, we lack the knowledge of the derivative;
- Probability is the theory of lack of information;
- Using randomness in DFO methods seems natural.

Main literature

From the DFO community:

- *Seminal work:* **Convergence of trust-region methods based on probabilistic models**
  Bandeira, Scheinberg and Vicente (2013).
  Convergence addressed, not complexity.

Main literature

From the DFO community:

- *Seminal work:* **Convergence of trust-region methods based on probabilistic models**
  Bandeira, Scheinberg and Vicente (2013).
  Convergence addressed, not complexity.
- *Follow-ups:* Larson, Billups (2013), Chen, Menickelly and Scheinberg (2015)
  Convergence for noisy/stochastic functions, but no complexity.

Main literature

From the DFO community:

- *Seminal work:* **Convergence of trust-region methods based on probabilistic models**
  Bandeira, Scheinberg and Vicente (2013).
  Convergence addressed, not complexity.
- *Follow-ups:* Larson, Billups (2013), Chen, Menickelly and Scheinberg (2015)
  Convergence for noisy/stochastic functions, but no complexity.
- **Direct search based on probabilistic descent**
  Gratton, Royer, Vicente and Zhang (2014).
  Convergence and complexity results.

Main literature

From the DFO community:

- *Seminal work:* **Convergence of trust-region methods based on probabilistic models**
  Bandeira, Scheinberg and Vicente (2013).
  Convergence addressed, not complexity.
- *Follow-ups:* Larson, Billups (2013), Chen, Menickelly and Scheinberg (2015)
  Convergence for noisy/stochastic functions, but no complexity.
- **Direct search based on probabilistic descent**
  Gratton, Royer, Vicente and Zhang (2014).
  Convergence and complexity results.
- *Model-based methods with complexity:* Cartis and Scheinberg (2015).

Main literature (2)

From other communities (convex or stochastic optimization, machine learning):

- *Random gradient-free optimization of convex functions*
  Nesterov (2011)
- *Query complexity of derivative-free optimization*
  Jamieson, Nowak and Recht (2012)
- *Optimal rates for zero-order convex optimization: the power of two function evaluations*
  Duchi et al. (2014)
- *Stochastic first- and zeroth-order methods for nonconvex stochastic programming*
  Ghadimi and Lan (2014)

The literature is still growing from both ends !

Outline

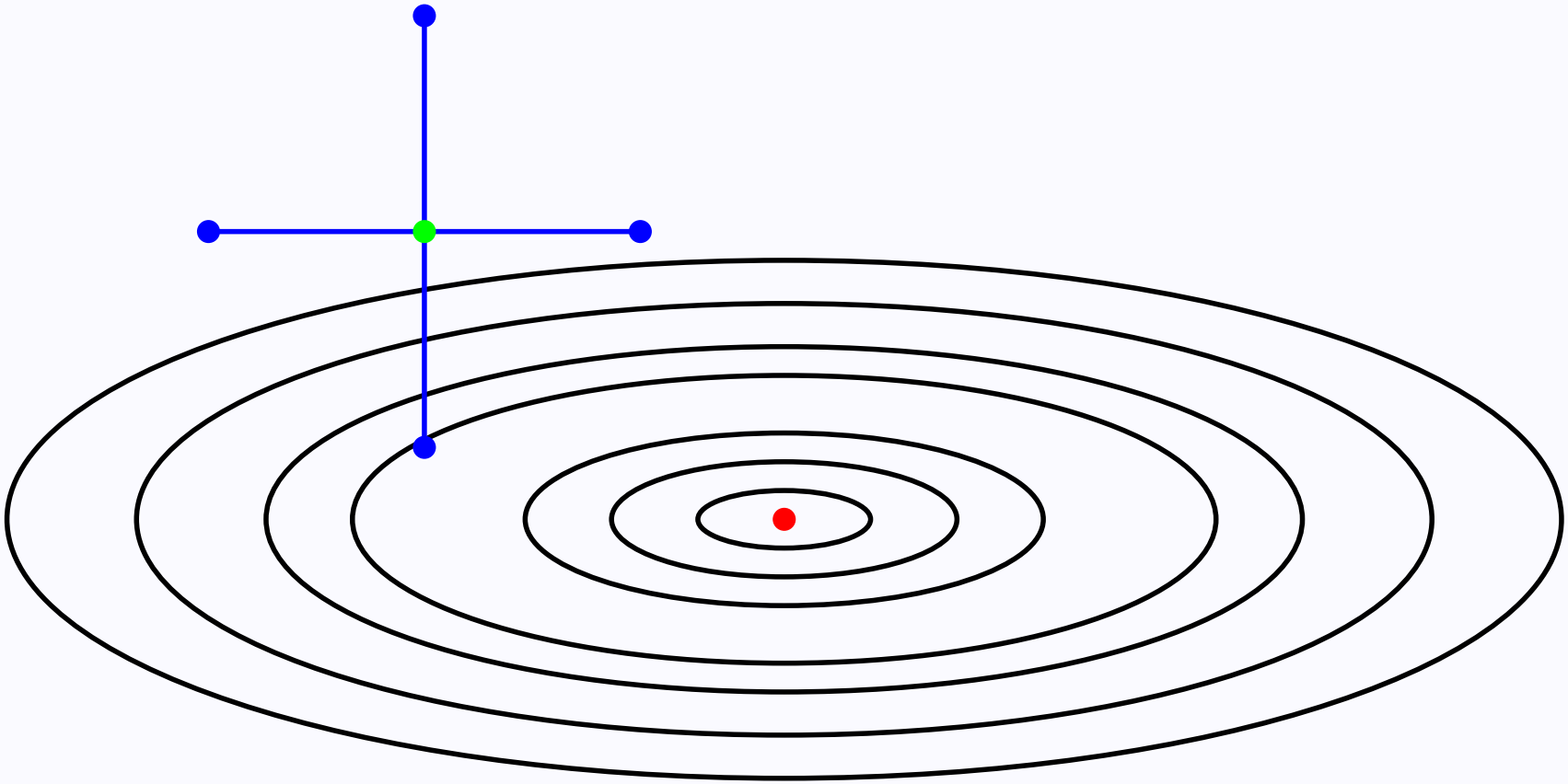
Outline

Direct search (DS)

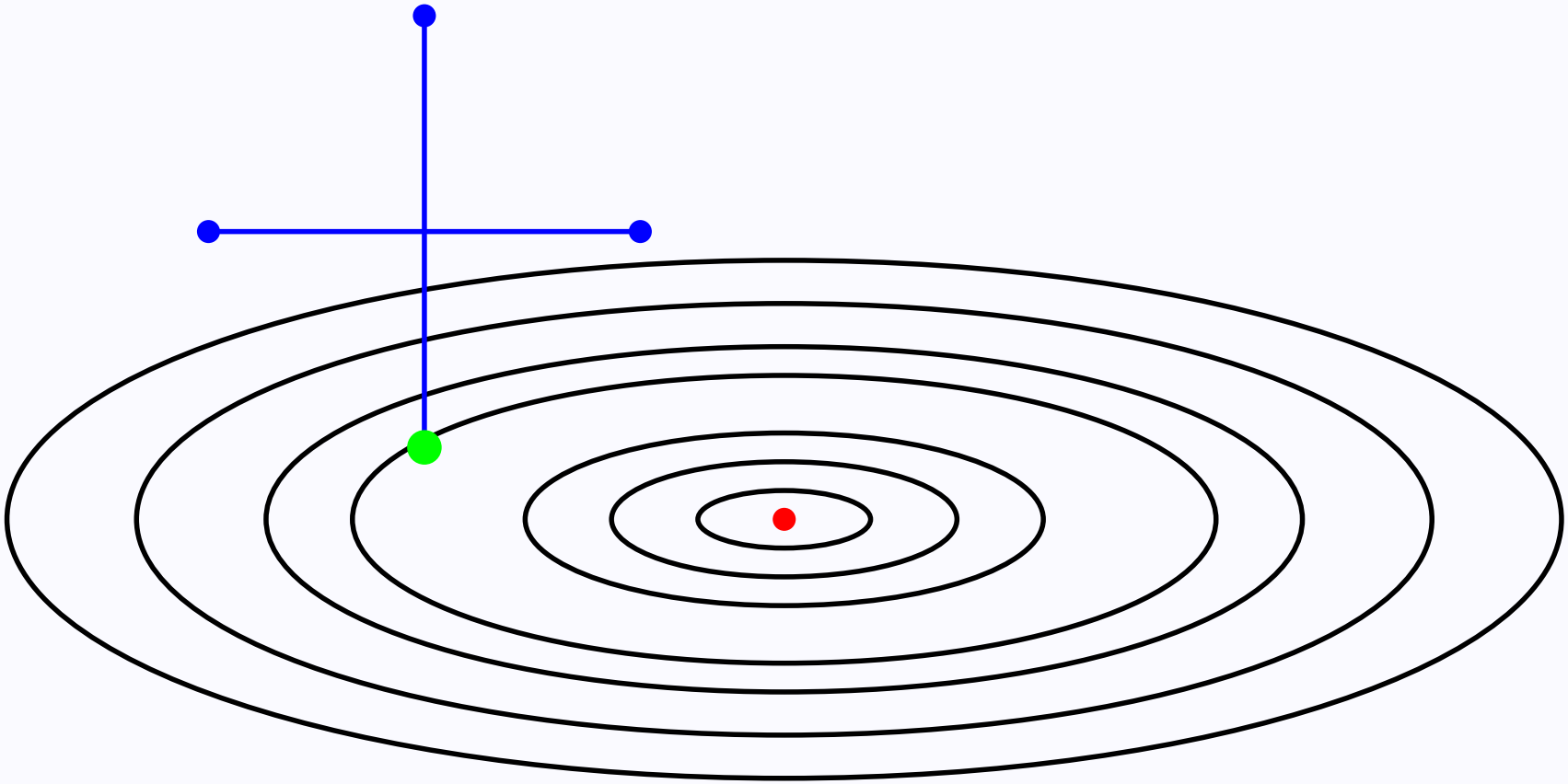
- Directional methods $\sim$ Steepest/Gradient Descent;
- Early appearance in the 1960s, convergence theory in the 1990s;
- Attractive: simplicity, parallel potential;
- Software: `NOMAD`, `APPSPACK`.

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

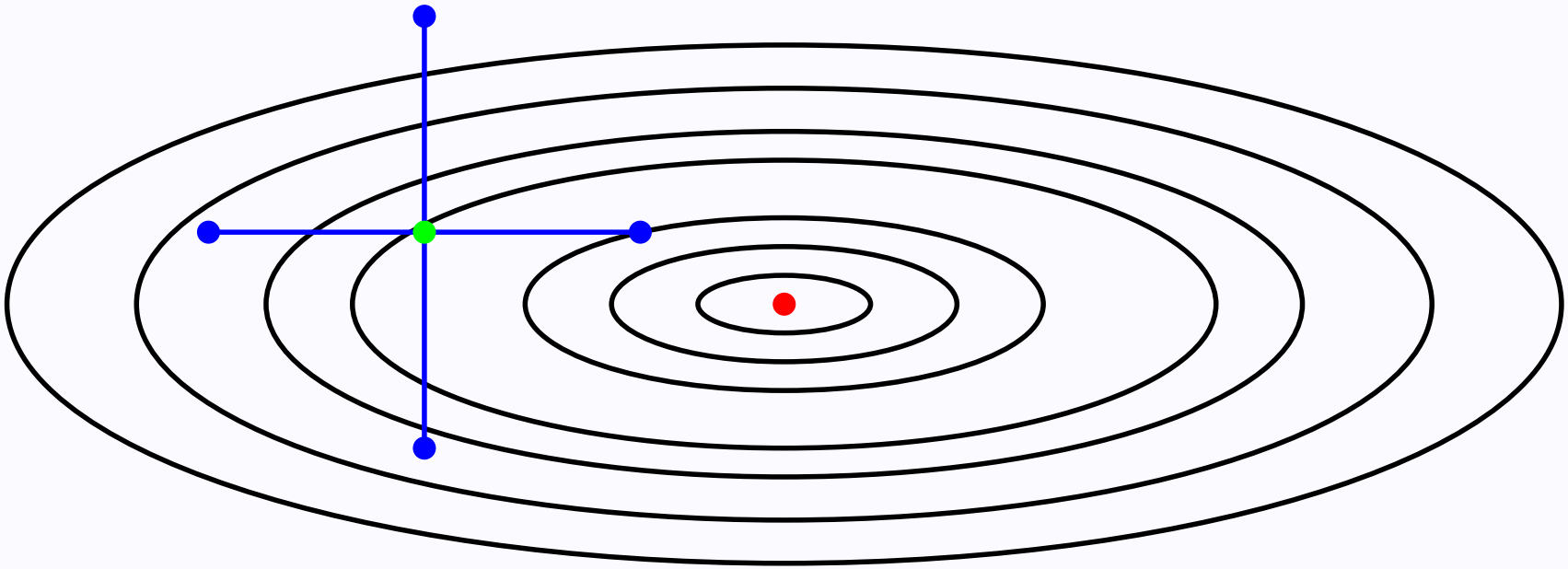
An example of DS : Coordinate Search



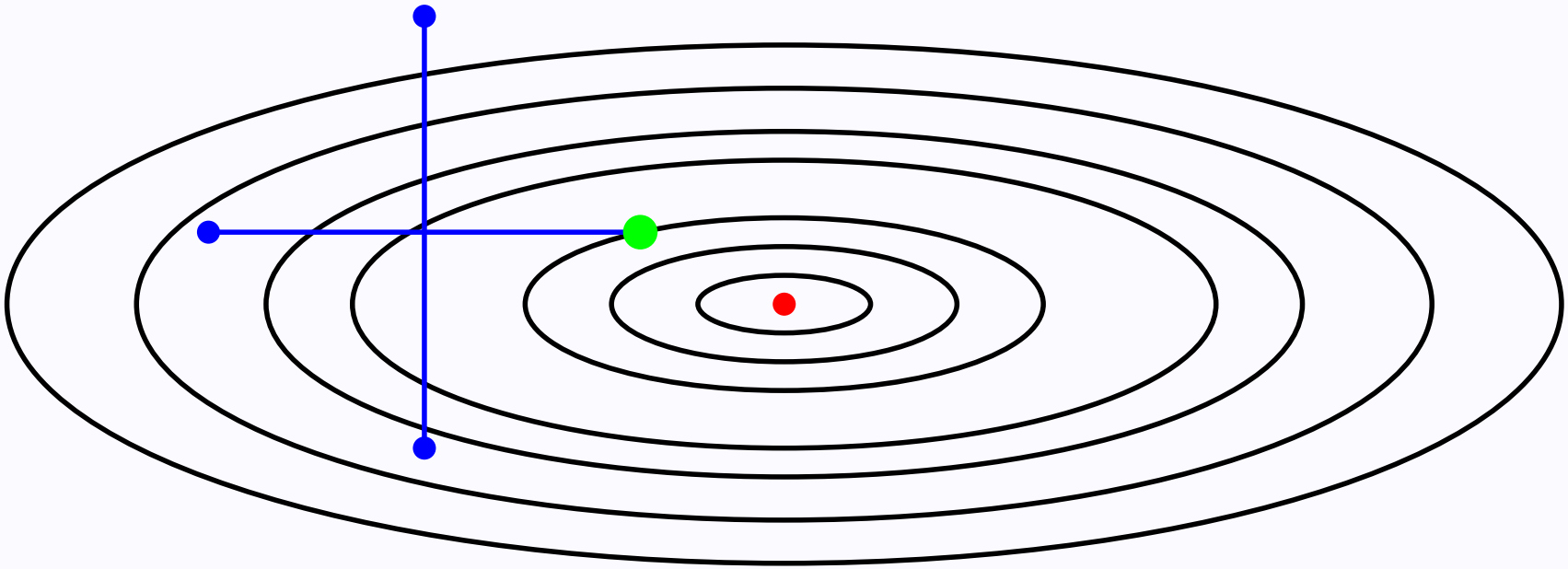
An example of DS : Coordinate Search



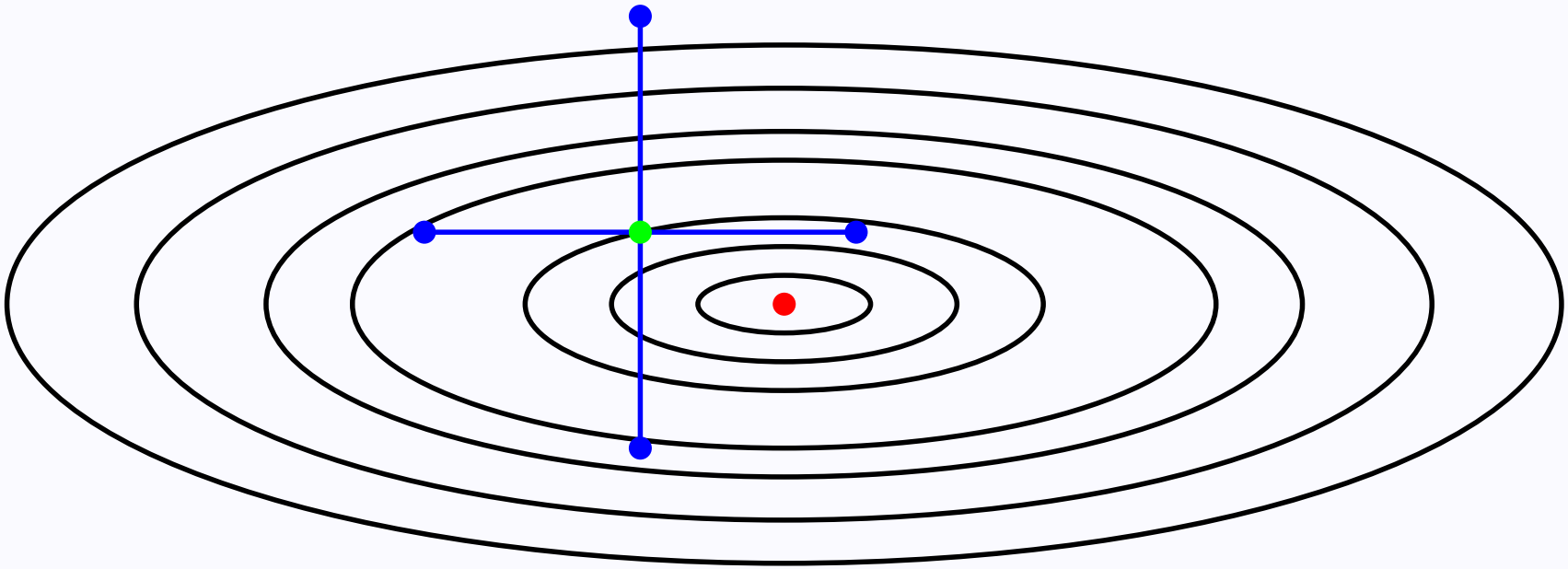
An example of DS : Coordinate Search



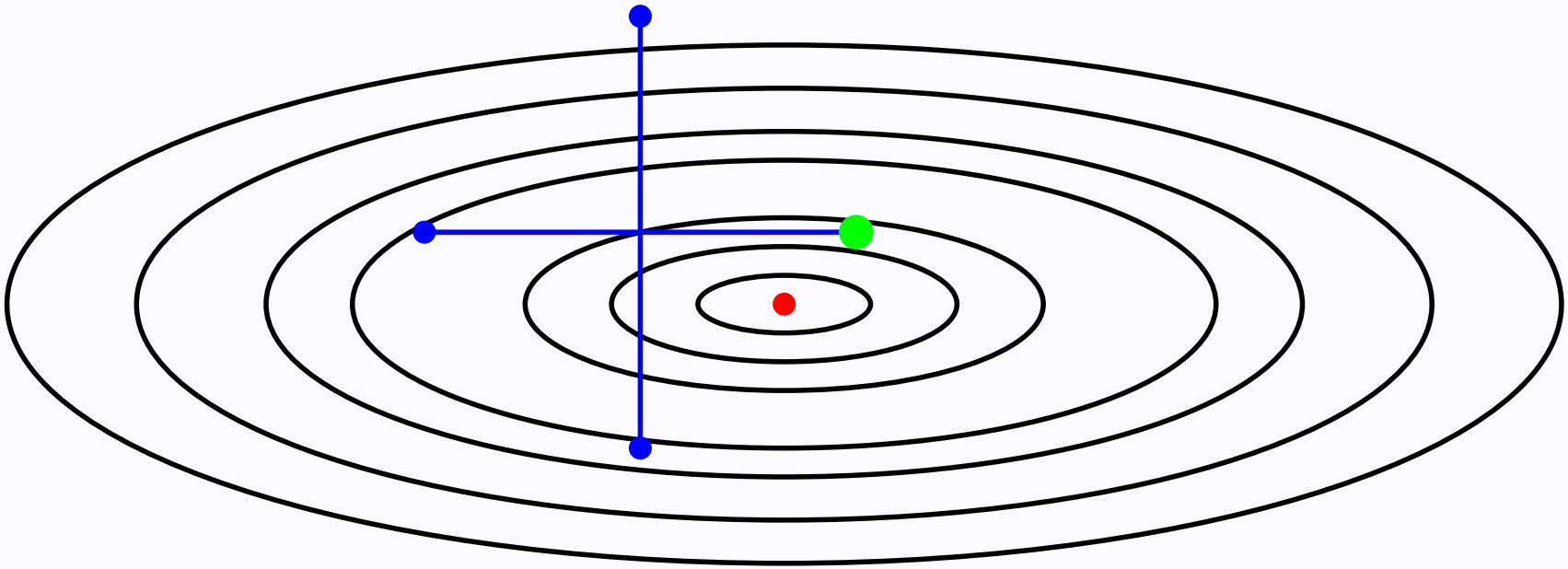
An example of DS : Coordinate Search



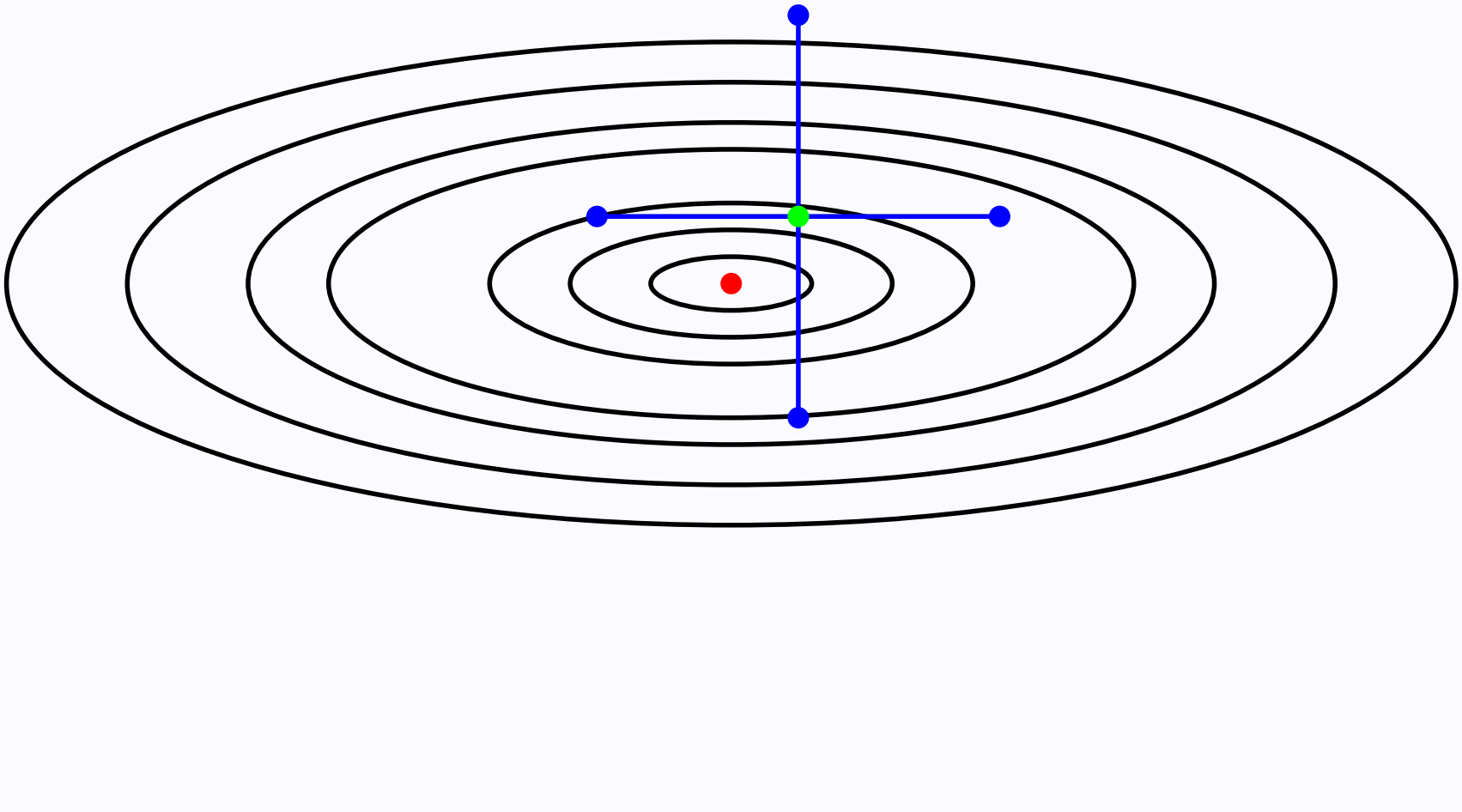
An example of DS : Coordinate Search



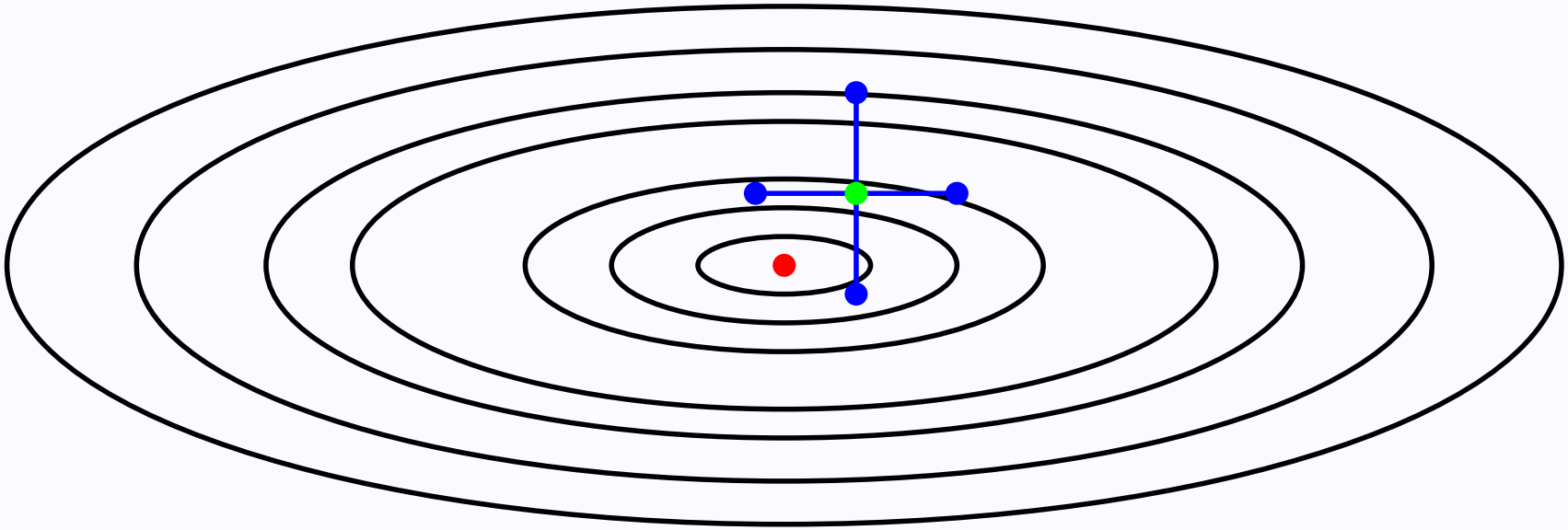
An example of DS : Coordinate Search



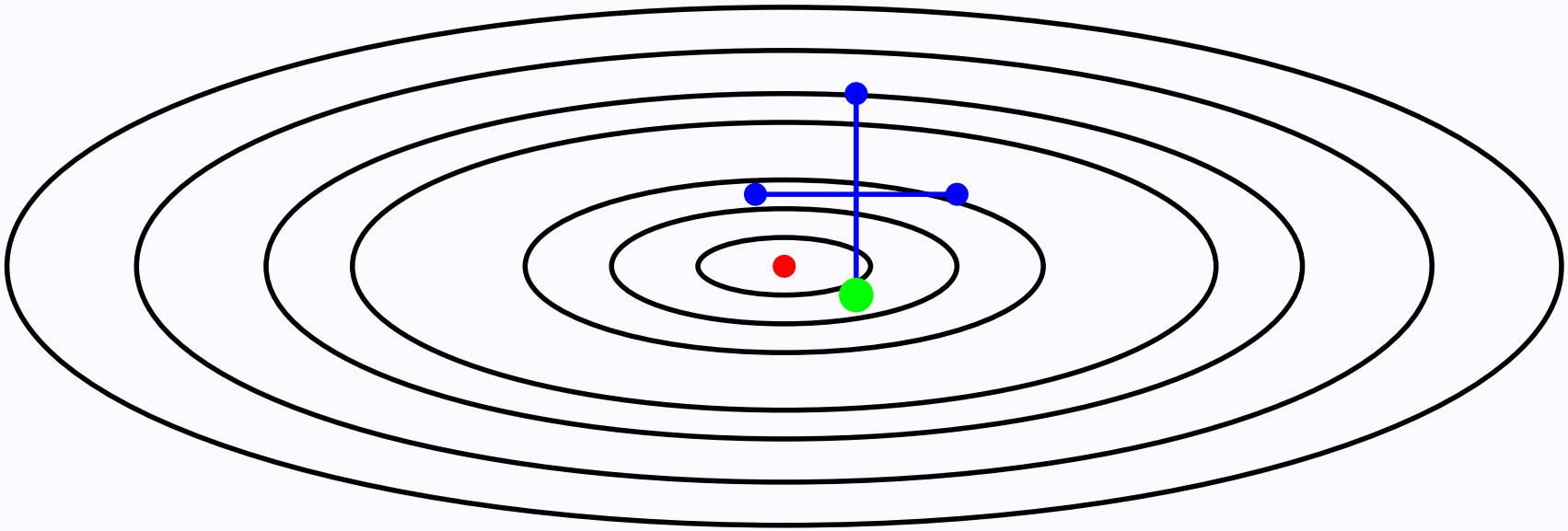
An example of DS : Coordinate Search



An example of DS : Coordinate Search



An example of DS : Coordinate Search



A basic framework for direct-search algorithms

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

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

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

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

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

Set quality

We would like to have $\mathrm{cm}(D) > 0$.
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** $\mathrm{cm}(D) > 0$;
- a PSS contains at least $n + 1$ vectors.

Example

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

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

Convergence for deterministic direct search

Lemma

Independently of $\{D_k\}$,

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

Proposition

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

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

Convergence for deterministic direct search

Lemma

Independently of $\{D_k\}$,

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

Proposition

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

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

Convergence Theorem

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

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

Outline

Introducing randomness

Idea from Gratton and Vicente (2013)

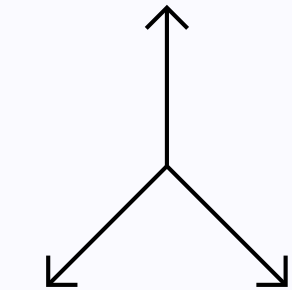
Randomly independently generate polling sets, possibly of less than $n + 1$ vectors!

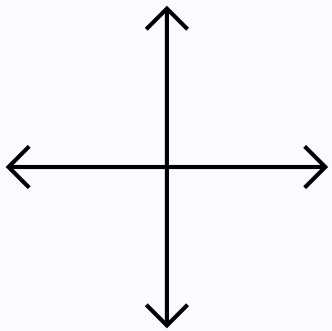
Introducing randomness

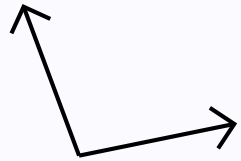
Idea from Gratton and Vicente (2013)

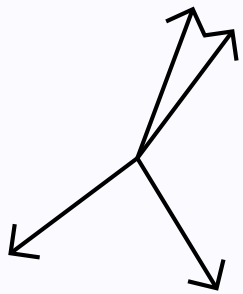
Randomly independently generate polling sets, possibly of less than $n+1$ vectors!

From PSS...









...to random sets

Numerical motivations

- Performances in dimension $n = 40$:

Problem	$[I\ -I]$	$[Q\ -Q]$	$2\,n$	$n+1$	$n/2$	2	1
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)

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: $\alpha_k = \mathcal{A}_k(\omega)$.

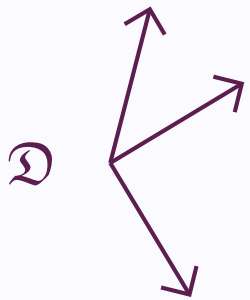
1. **Initialization:** Set $x_0 \in \mathbb{R}^n, \alpha_0 > 0, 0 < \theta < 1 \leq \gamma$.
2. **For** $k = 0, 1, 2, \ldots,$
   - Choose a set $\mathfrak{D}_k$ of m independent random vectors.
   - If it exists $\mathfrak{d}_k \in \mathfrak{D}_k$ so that

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

   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 $\mathcal{A}_{k+1} := \theta \mathcal{A}_k$.

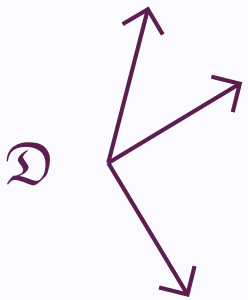
A new measure of set quality

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

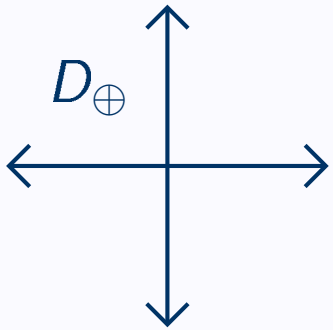


A new measure of set quality

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

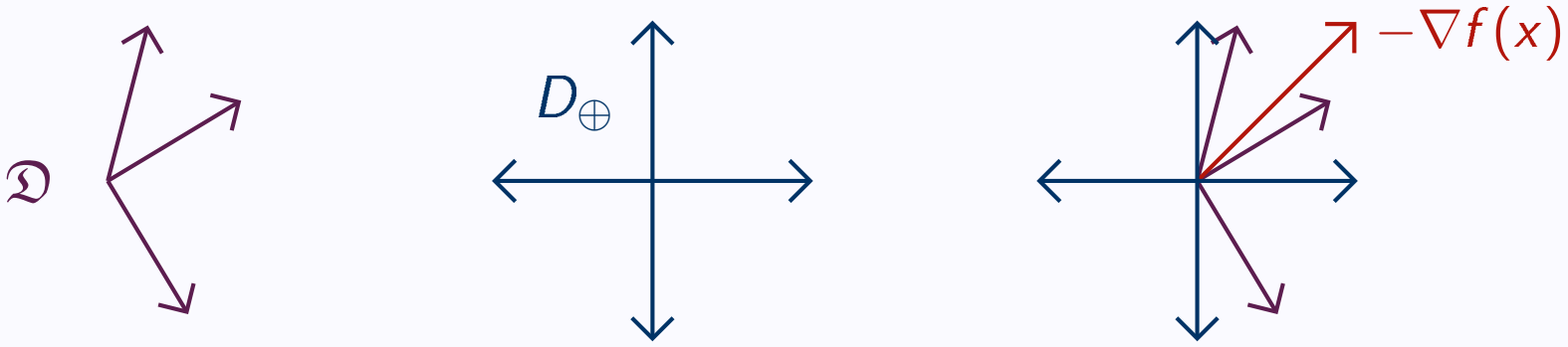


...$D_{\oplus}$ is...



A new measure of set quality

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



Is being close to the negative gradient enough ?

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 ?

Analysis of a randomized method

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

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

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

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

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

We need to show that $\{\mathrm{cm}\left(\mathfrak{D}_k, -\nabla f(X_k)\right) \geq \kappa\}$ 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

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

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

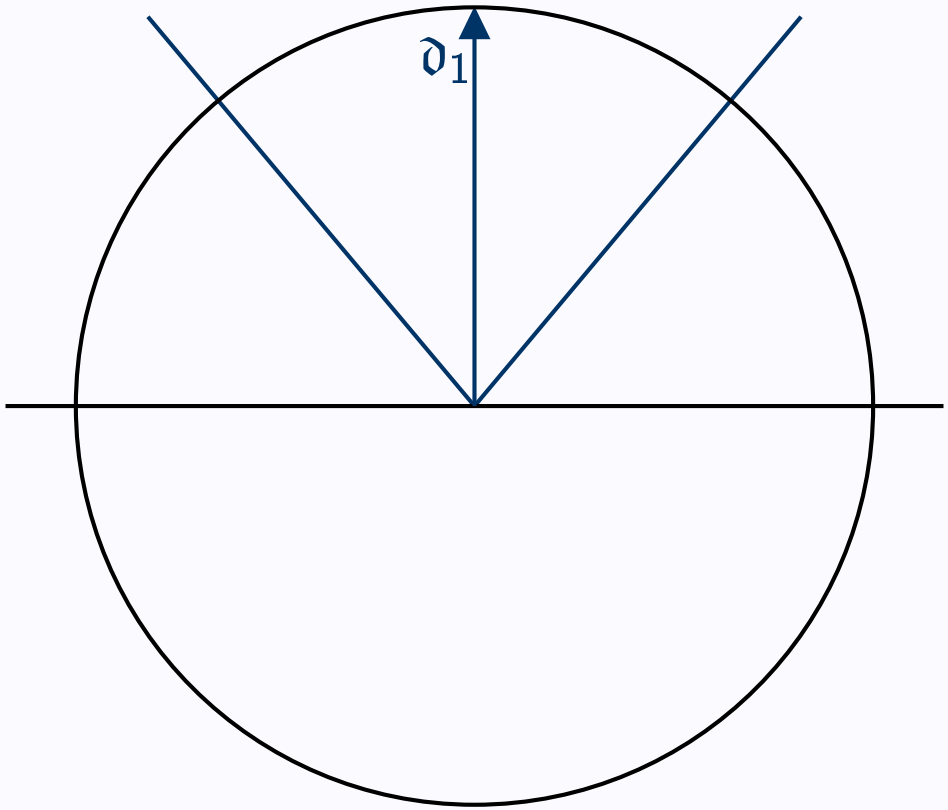
If

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

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

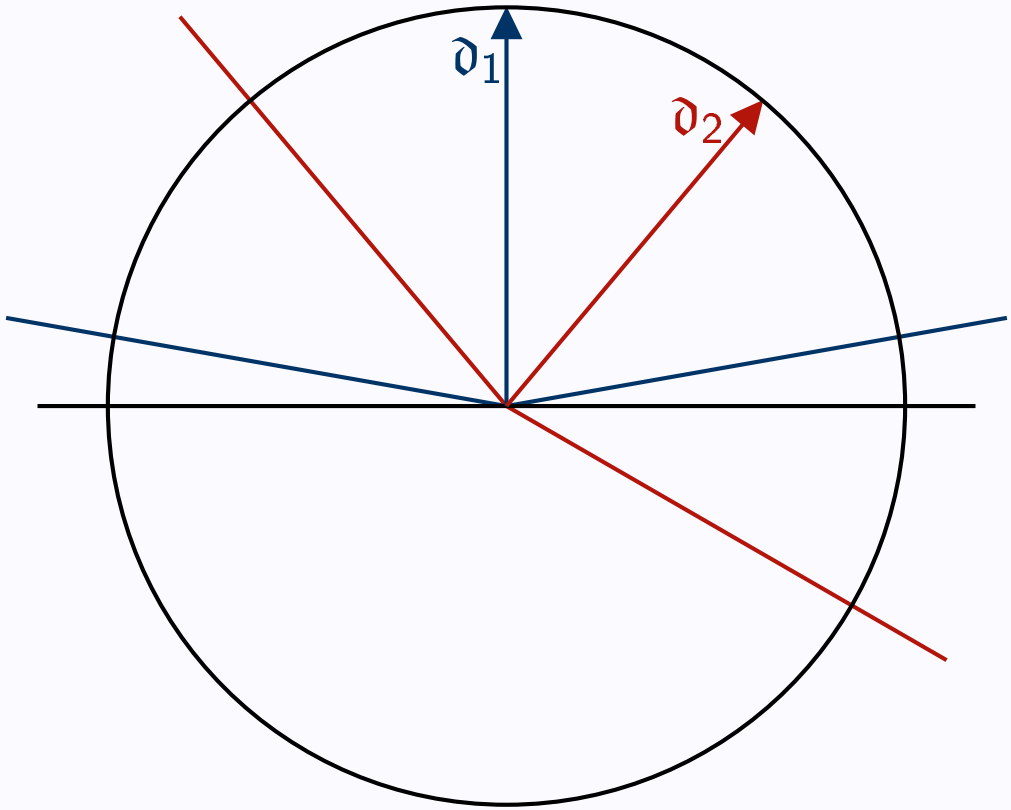
If $\gamma = \theta^{-1} = 2$, it suffices to choose $m \geq 2$ to have $p > \frac{1}{2}$.

Two is enough, one is not



$$\forall \kappa \in (0,1),\ \forall g, \quad \mathbb{P}\left(\mathfrak{d}_1^\top g \geq \kappa\right) < 1/2$$

Two is enough, one is not

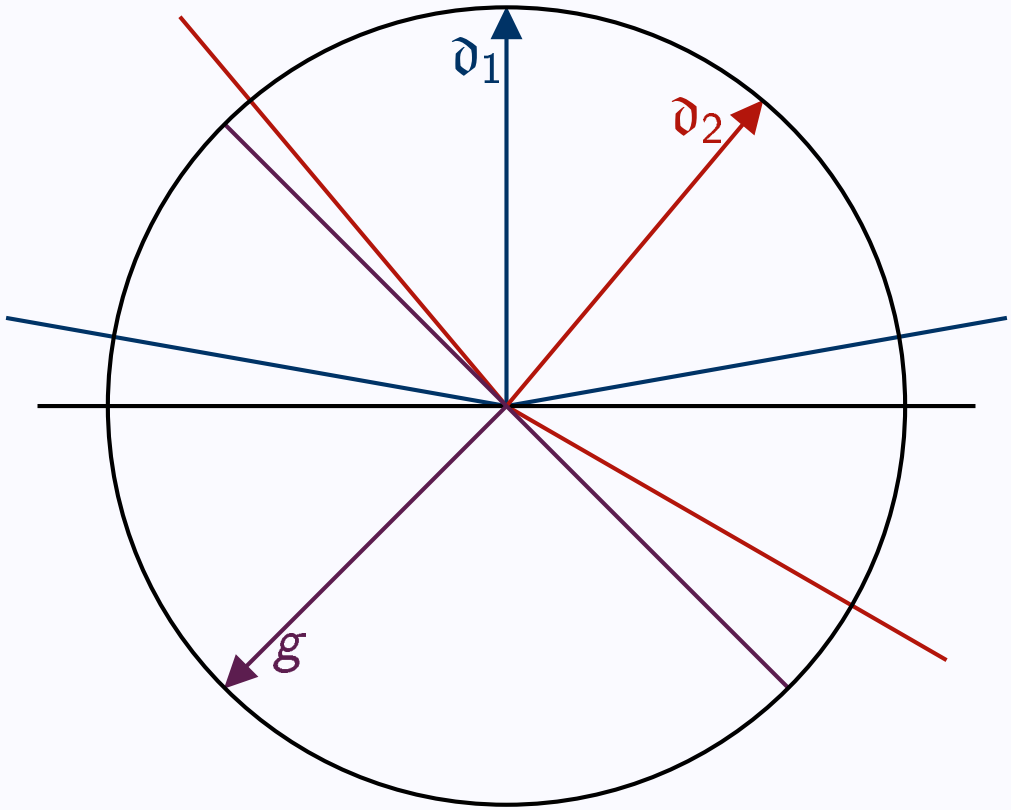


$$\forall \kappa \in (0,1),\ \forall g, \quad \mathbb{P}\left(\mathfrak{d}_1^\top g \geq \kappa\right) = \mathbb{P}\left(\mathfrak{d}_1^\top g \geq \kappa\right) < 1/2$$

but

$$\exists \kappa^* \in (0,1), \forall g, \quad \mathbb{P}\left(\max\left\{\mathfrak{d}_1^\top g, \mathfrak{d}_2^\top g\right\} \geq \kappa^*\right) > 1/2$$

Two is enough, one is not



$$\forall \kappa \in (0,1),\ \forall g, \quad \mathbb{P}\left(\mathfrak{d}_1^\top g \geq \kappa\right) = \mathbb{P}\left(\mathfrak{d}_1^\top g \geq \kappa\right) < 1/2$$

but

$$\exists \kappa^* \in (0,1), \forall g, \quad \mathbb{P}\left(\max\left\{\mathfrak{d}_1^\top g, \mathfrak{d}_2^\top g\right\} \geq \kappa^*\right) > 1/2$$

What about the practical choice ?

Finding the best number of random directions

- m^* depends on γ and θ;
- Technique: Tune γ to ensure $(p, \tau/\sqrt{n})$-descent with $p > 1/2$.

Numerical testing

- $\theta = 1/2$, γ varies with the solver;
- Convergence test: $f(x_k) - f_{best} < 10^{-3}\,(f(x_0) - f_{best})$;
- Performance criterion: # of evaluations.

More numerical experiments

Problem	$[I\ -I]\,(\gamma=1)$	$[Q\ -Q]\,(\gamma=1)$	$4\,(\gamma=1.1)$	$2\,(\gamma=2)$
arglina	1.00	3.17	6.73	5.86
arglinb	34.12	5.34	2.02	1.00
broydn3d	1.00	1.91	3.47	2.04
dqrtic	1.18	1.36	1.48	1.00
engval1	1.05	1.00	2.89	2.29
freuroth	17.74	7.39	1.00	1.35
integreq	1.54	1.49	1.34	1.00
nondquar	1.00	2.82	1.73	1.37
sinquad	-	1.26	-	1.00
vardim	20.31	11.02	1.84	1.00

Table : Relative number of evaluations for different polling sets ($n = 40$)

More numerical experiments

Problem	$[I\ -I]\,(\gamma=1)$	$[Q\ -Q]\,(\gamma=1)$	$4\,(\gamma=1.1)$	$2\,(\gamma=2)$
arglina	1.00	3.86	7.58	5.86
arglinb	138.28	107.32	1.99	1.00
broydn3d	1.00	2.57	3.21	1.92
dqrtic	3.01	3.25	1.46	1.00
engval1	1.04	1.00	2.84	2.06
freuroth	31.94	17.72	1.36	1.00
integreq	1.83	1.66	1.22	1.00
nondquar	1.18	2.83	1.17	1.00
sinquad	-	-	-	-
vardim	112.22	19.72	2.36	1.00

Table : Relative number of evaluations for different polling sets ($n = 100$)

Outline

Theorem (Vicente 2013)

Let N_ϵ be the number of function evaluations needed to reach an point such that $\inf_{0 \leq l \leq k} \|\nabla f(x_l)\| < \epsilon$; then

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

The corresponding global rate for $\|\nabla f(x_k)\|$ is $\mathcal{O}\left(\frac{1}{\sqrt{k}}\right)$.

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

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

Worst-case complexity in deterministic direct search

Theorem (Vicente 2013)

Let N_ϵ be the number of function evaluations needed to reach an point such that $\inf_{0 \leq l \leq k} \|\nabla f(x_l)\| < \epsilon$; then

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

The corresponding global rate for $\|\nabla f(x_k)\|$ is $\mathcal{O}\left(\frac{1}{\sqrt{k}}\right)$.

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

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

What about the randomized case ?

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 $Z_k = \mathbf{1}\left(\text{cm}(\mathfrak{D}_k, -G_k) \geq \kappa\right)$.

- 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

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

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

WCC for probabilistic descent (2)

An inclusion argument

$$\left\{ \|\tilde{G}_k\| \geq \epsilon \right\} \subset \left\{ \sum_{l=0}^{k-1} Z_l \leq \lambda\, k \right\}$$

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

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

The convergence rate of $\|\tilde{G}_k\|$ gets exponentially close to $\frac{1}{\sqrt{k}}$.

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

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

The convergence rate of $\|\tilde{G}_k\|$ gets exponentially close to $\frac{1}{\sqrt{k}}$.

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

Worst-case complexity: conclusions

- Theoretical improvement for # of evaluations.
- Matches the practical performance.
- **Question:** *Can this theory be applied to other DFO methods ?*

A derivative-free trust-region algorithm

- **Initialization:** Set $x_0 \in \mathbb{R}^n, \delta_0 > 0, \gamma > 1, 0 < \eta_1 < \eta_2 < 1$.
- **For** $k = 0, 1, 2, \ldots$
  1. Generate a model m_k of f in $B(x_k, \delta_k)$ using m points.
  2. Compute an approximate solution of the model subproblem

$$s_k \approx \operatorname{argmin}_{\|s\| \leq \delta_k} m_k(s).$$

  3. Evaluate $f(x_k + s_k)$ and $\rho_k = \frac{f(x_k) - f(x_k + s_k)}{m_k(0) - m_k(s_k)}$.
  4. If $\rho_k \geq \eta_1$, set $x_{k+1} = x_k + s_k$ and

$$\begin{cases} \delta_{k+1} = \gamma\, \delta_k & \text{if } \|\nabla m_k(x_k)\| \geq \eta_2\, \delta_k, \\ \delta_{k+1} = \delta_k & \text{otherwise.} \end{cases}$$

  5. Otherwise, set $x_{k+1} = x_k$ and $\delta_{k+1} = \gamma^{-1}\, \delta_k$.

Randomizing the models

First-order convergent DFO trust-region rely on fully linear models:

Fully linear model

A model m_k is a κ_f, κ_g-fully linear model of f on $B(x_k, \delta_k)$ if

$$\begin{aligned} \forall s \in B(x_k, \delta_k), \quad \|m_k(x_k + s) - f(x_k + s)\| &\leq \kappa_f \, \delta_k^2 \\ \|\nabla f(x_k + s) - \nabla m_k(x_k + s)\| &\leq \kappa_g \, \delta_k . \end{aligned}$$

If the models $\{M_k\}$ are random:

Probabilistic fully linear sequence

$\{M_k\}$ is (p, κ_f, κ_g)-fully linear if

$$\begin{aligned} \mathbb{P}(\ M_0 \ \kappa_f, \kappa_g - \textit{fully linear}\ ) &\geq p \\ \forall k \geq 1, \quad \mathbb{P}\left(\ M_0 \ \kappa_f, \kappa_g - \textit{fully linear} \,\middle|\, \mathfrak{S}_{k-1}^M\right) &\geq p, \end{aligned}$$

where $\mathfrak{S}_{k-1}^M = \sigma(M_0, \ldots, M_{k-1})$.

Proof analogies

Direct search	Trust-region
$\{\mathfrak{D}_k\}$ (p, κ)-descent	$\{M_k\}$ (p, κ_f, κ_g)-fully linear
$\alpha_k \to 0$	$\delta_k \to 0$
(D_k descent + $\alpha_{k+1} < \alpha_k$) $\Rightarrow \|\nabla f(x_k)\| \leq \mathcal{O}(\alpha_k)$	(m_k fully linear + $\delta_{k+1} < \delta_k$) $\Rightarrow \|\nabla f(x_k)\| \leq \mathcal{O}(\delta_k)$
$\alpha_{k+1} \geq \alpha_k$ $\Rightarrow$ $f(x_k) - f(x_{k+1}) \geq \mathcal{O}(\alpha_k^2)$	$\delta_{k+1} > \delta_k$ $\Rightarrow$ $f(x_k) - f(x_{k+1}) \geq \mathcal{O}(\delta_k^2)$.

- Both methods: $N_\epsilon \leq \mathcal{O}\left(m\, n\, \epsilon^{-2}\right)$ with overwhelming probability;
- $\mathcal{O}\left(n\, \epsilon^{-2}\right)$ bound only proved in the DS case.

Comments and other methods

Model-based methods

- Also applicable to Levenberg-Marquardt with probabilistic gradients (Bergou et al, 2013);
- Models typically require up to $\mathcal{O}(n)$ evaluations.

Directional methods

- Results apply to Line search:
  - Seen as direct-search methods;
  - Seen as model-based methods (Cartis and Scheinberg, 2015).
- In the first case, we can ensure a $\mathcal{O}(n\,\epsilon^{-2})$ bound on the number of evaluations !

Outline

Second-order optimality

Assumption

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

Goals

- Exploit (negative) curvature information;
- Converge to a second-order stationary point

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

- Randomized strategy.

Second-order DFO (Model-based)

For trust-region methods

- Use negative curvature directions of fully quadratic models;
- Convergence to second-order stationary points;
- Requires $\mathcal{O}(n^2)$ evaluations.

Randomized case

- Theoretical results based on probabilistically fully quadratic models;
- No practical propositions.

Second-order DFO methods (directional)

A second-order globally convergent direct-search method and its worst-case complexity. Gratton, Royer and Vicente (2015).

Key features

- A PSS D_k, as before;
- A linear basis B_k used to gather curvature information;

Theorem

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

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

then

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

Two avenues for randomization

Randomizing the "first-order" directions

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

Two avenues for randomization

Randomizing the "first-order" directions

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

Randomizing the "second-order" directions

- Focus on ensuring $\mathbb{P}(\sigma_{\min}(B_k) \geq \sigma)$;
- Use results from randomized linear algebra;
- Implicit randomized Hessian;
- Still almost-sure convergence.

Towards second-order directions

- Using the B_k is expensive.
- Can we replace B_k by random directions ?

A general problem

Generate $d \in \mathbb{R}^n, \|d\| = 1$ such that the Taylor expansion

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

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

Towards second-order directions

- Using the B_k is expensive.
- Can we replace B_k by random directions ?

A general problem

Generate $d \in \mathbb{R}^n, \|d\| = 1$ such that the Taylor expansion

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

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

C. Royer, **Probabilistic tools for derivative-free optimization.** PhD Thesis, 2016.

Summary

- Derivative-free optimization is easy to combine with randomization.

Summary

- Derivative-free optimization is easy to combine with randomization.
- Convergence and complexity guarantees follow.

Summary

- Derivative-free optimization is easy to combine with randomization.
- Convergence and complexity guarantees follow.
- Practical performance is enhanced in the direct-search case.

Summary

- Derivative-free optimization is easy to combine with randomization.
- Convergence and complexity guarantees follow.
- Practical performance is enhanced in the direct-search case.

Open questions

- In DFO: random models with less than $\mathcal{O}(n)$ points;
- Second-order aspects: randomized Hessians VS randomized curvature directions ?

Thank you for your attention !

clement.royer@enseeiht.fr