Propriétés probabilistes dans les algorithmes d'optimisation sans et avec dérivées

Clément Royer - University of Wisconsin-Madison

Séminaire SPOC
Institut de Mathématiques de Bourgogne

12 avril 2017

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.
- With randomness: results in expectation/probability.

Using complexity

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

Objectives of the talk

Main track

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

Objectives of the talk

Main track

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

- In this talk: focus on direct-search methods;
- Apply to other frameworks, like trust-region.

Outline

Outline

1. Deterministic direct search
   - Derivative-free optimization
   - Direct search
2. Direct search based on probabilistic descent
3. Extension to bound and linearly constrained problems
4. Probabilistic properties in derivative-based algorithms

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, a priori not convex.
- 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, a priori not convex.
- 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, medicine,...

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, medicine,...

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.
- No deterministic variant.

- This talk does 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 r 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 r 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 $\mathrm{cm}(D) > 0$.

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.

Set quality

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

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

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

Lemma

Independently of $\{D_k\}$,

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

Convergence for deterministic direct search

Lemma

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

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

Lemma

Independently of $\{D_k\}$,

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

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

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,

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

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

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

Outline

Introducing randomness

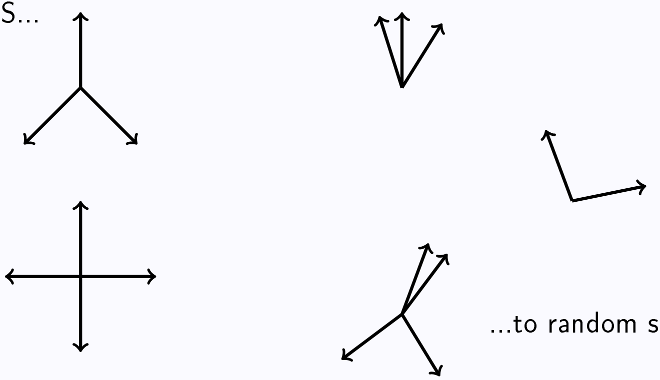
Idea (Gratton and Vicente, 2013)

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

Idea (Gratton and Vicente, 2013)

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

From PSS...



...to random sets

- Convergence test: $f(x_k) < f_{\rm low} + 10^{-3}\,(f(x_0) - f_{\rm low})$;
- 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: $\alpha_k = A_k(\omega)$.

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 r **independent random** vectors.
   - If it exists $\mathfrak{d}_k \in \mathfrak{D}_k$ so that

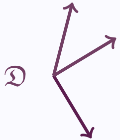
     $$f(X_k + A_k\, \mathfrak{d}_k) < f(X_k) - A_k^2,$$

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

Outline

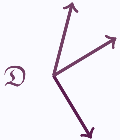
First step: What is a good random polling set ?

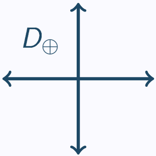
$\mathfrak{D}$ is not a PSS...



First step: What is a good random polling set ?

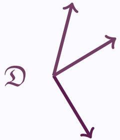
$\mathfrak{D}$ is not a PSS... ...$D_\oplus$ is...

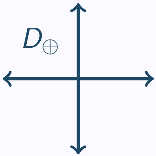


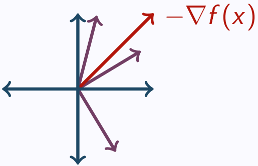


First step: What is a good random polling set ?

$\mathfrak{D}$ 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

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

(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:

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

Convergence results

Lemma

For all realizations $\{\alpha_k\}$ of $\{A_k\}$, independently of $\{\mathfrak{D}_k\}$,

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

Lemma

If k is an unsuccessful iteration; then

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

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 $Z_k = \mathbb{1}\left(\text{cm}\left(\mathfrak{D}_k, -\nabla f(X_k)\right) \geq \kappa\right)$.

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

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

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 $\kappa \, \|G_k\| < \mathcal{O}(A_k)$...

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 $\kappa \|G_k\| < \mathcal{O}(A_k)$...
- ...A_k goes to zero...

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 $\kappa \|G_k\| < \mathcal{O}(A_k)$...
- ...A_k goes to zero...
- ...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 $\kappa \, \|G_k\| < \mathcal{O}(A_k)$...
- ...A_k goes to zero...
- ...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)

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

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

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

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

- Deterministic: $\mathcal{O}(n^2\,\epsilon^{-2})$.
- Probabilistic: $\mathcal{O}(r\,n\,\epsilon^{-2})$ in probability
  $\Rightarrow \mathcal{O}(n\,\epsilon^{-2})$ when $r=2$!
- 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 $r = |\mathfrak{D}_k|$ possible.

A practical example: uniform distribution over the unit sphere

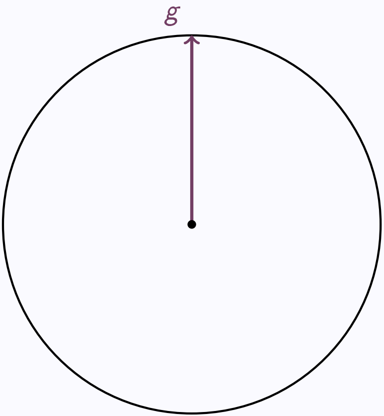
If

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

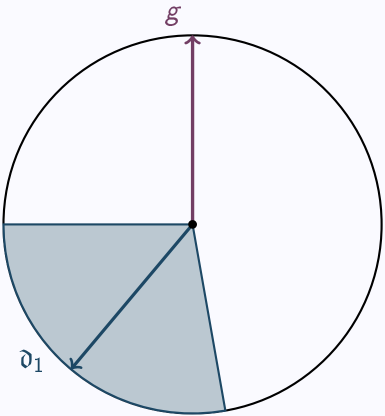
then there exist p and τ independent of n such that the sequence $\mathfrak{D}_k$ is $(p, \tau/\sqrt{n})$-descent, with $p > p_0$.

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

Two uniform directions are enough, one is not

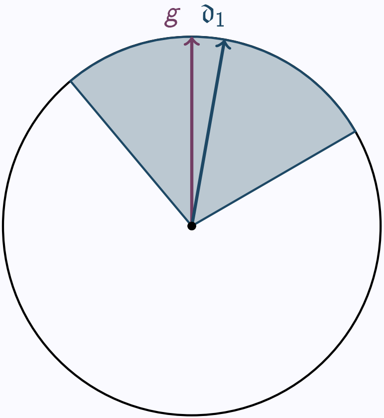


Two uniform directions are enough, one is not



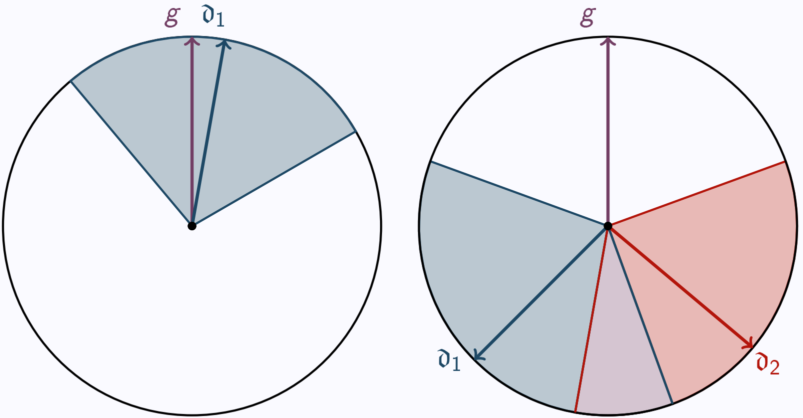
$$\mathfrak{d}_1 \sim \mathcal{U}(\mathbb{S}^1) \Rightarrow \forall \kappa \in (0,1), \quad \mathbb{P}\left(\mathrm{cm}\left(\mathfrak{d}_1, g\right) = \mathfrak{d}_1^\top g \geq \kappa\right) < 1/2.$$

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

Two linearly constrained problems

Linear equality constraints

$$\begin{cases} \min_{x\in\mathbb{R}^n} & f(x) \\ \text{s.t.} & Ax = b. \end{cases}$$

- Equivalent to the unconstrained problem $\min_{\tilde{x}\in\mathbb{R}^{n-m}} f\left(x_0 + W\tilde{x}\right)$ with $W \in \mathbb{R}^{n\times(n-m)}$ orthonormal basis for null(A) and $Ax_0 = b$.
- Deterministic and probabilistic analyses apply !

Two linearly constrained problems

Linear equality constraints

$$\begin{cases} \min_{x \in \mathbb{R}^n} & f(x) \\ \text{s.t.} & Ax = b. \end{cases}$$

- Equivalent to the unconstrained problem $\min_{\tilde{x} \in \mathbb{R}^{n-m}} f(x_0 + W\tilde{x})$ with $W \in \mathbb{R}^{n \times (n-m)}$ orthonormal basis for $\mathrm{null}(A)$ and $Ax_0 = b$.
- Deterministic and probabilistic analyses apply !

Bound constrained case

$$\begin{cases} \min_{x \in \mathbb{R}^n} & f(x) \\ \text{s.t.} & l \le x \le u. \end{cases}$$

- Deterministic practice: Uses $D_\oplus = \{e_1, \dots, e_n, -e_1, \dots, -e_n\}$ to guarantee convergence and moves parallel to the constraints.

Algorithm (deterministic version)

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 at most r vectors.
   - If it exists $d_k \in D_k$ so that $x_k + \alpha_k d_k$ **is feasible** and

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

     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 $\alpha_{k+1} := \theta \alpha_k$.

Bound constraints

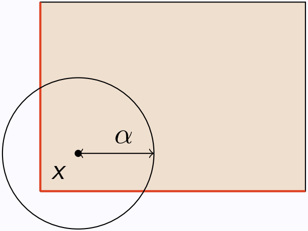
- Feasible set: $\mathcal{F} = \{l \leq x \leq u\}$.

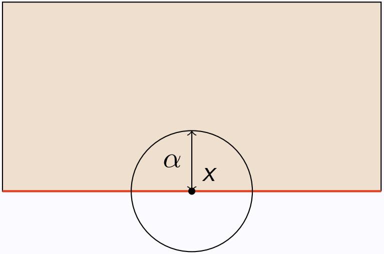
Nearby constraints

The indexes

$$
\begin{aligned}
I_u(x,\alpha) &= \{i : |u_i - [x]_i| \leq \alpha\} \\
I_l(x,\alpha) &= \{i : |l_i - [x]_i| \leq \alpha\}
\end{aligned}
$$

define the nearby constraints at $x \in \mathcal{F}$ given $\alpha > 0$.



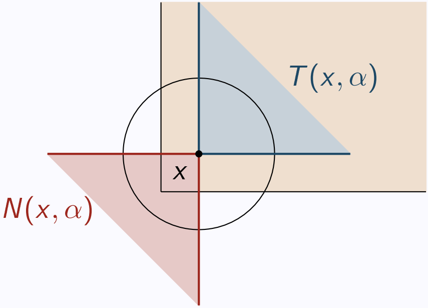


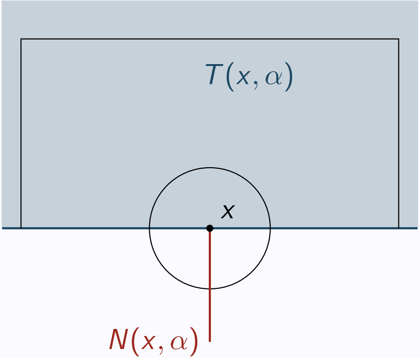
Bound constraints (2)

- **Approximate normal cone $N(x,\alpha)$:** Positive span of

$$\{e_i\}_{i \in I_u(x,\alpha)} \cup \{-e_i\}_{i \in I_l(x,\alpha)} .$$

- **Approximate tangent cone $T(x,\alpha)$:** polar of $N(x,\alpha)$.





Feasible descent property

- Recall the cosine measure that identifies descent directions

$$\text{cm}(D, -\nabla f(x)) = \max_{d \in D} \frac{d^\top[-\nabla f(x)]}{\|d\| \| -\nabla f(x)\|}.$$

Feasible descent property

D is a κ-feasible descent set for $T(x,\alpha)$ if $D \subset T(x,\alpha)$ and

$$\text{cm}_{T(x,\alpha)}(D, -\nabla f(x)) = \max_{d \in D} \frac{d^\top[-\nabla f(x)]}{\|d\| \|P_{T(x,\alpha)}[-\nabla f(x)]\|} \geq \kappa.$$

Feasible descent property

- Recall the cosine measure that identifies descent directions

$$\mathrm{cm}(D, -\nabla f(x)) \; = \; \max_{d \in D} \frac{d^\top[-\nabla f(x)]}{\|d\| \| - \nabla f(x)\|}.$$

Feasible descent property

D is a κ-feasible descent set for $T(x,\alpha)$ if $D \subset T(x,\alpha)$ and

$$\mathrm{cm}_{T(x,\alpha)}(D, -\nabla f(x)) \;=\; \max_{d \in D} \frac{d^\top[-\nabla f(x)]}{\|d\| \|P_{T(x,\alpha)}[-\nabla f(x)]\|} \;\geq\; \kappa.$$

- Using κ-feasible descent sets guarantee both convergence and complexity (similar analysis than unconstrained case).
- $D_\oplus \cap T(x,\alpha)$ is always $\frac{1}{\sqrt{n}}$-feasible descent.

Probabilistic feasible descent

Definition

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

$$\begin{aligned} \mathbb{P}\left(\mathrm{cm}_{T_0}\left(\mathfrak{D}_0, -\nabla f(x_0)\right) \geq \kappa\right) &\geq p \\ \forall k \geq 1, \quad \mathbb{P}\left(\mathrm{cm}_{T_k}\left(\mathfrak{D}_k, -\nabla f(X_k)\right) \geq \kappa \,\middle|\, \mathfrak{D}_0, \dots, \mathfrak{D}_{k-1}\right) &\geq p. \end{aligned}$$

where $T_k = T(X_k, A_k)$.

Theoretical guarantees

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

- **Almost-sure convergence towards stationary point**;
- **Complexity bound for ϵ-stationarity:**

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

Towards randomization

Main concerns

- How to define probabilistic feasible descent sets ?
- What are the orders of r and κ ?
- Can we use **less directions** than in the deterministic case ?

Towards randomization

Main concerns

- How to define probabilistic feasible descent sets ?
- What are the orders of r and κ ?
- Can we use **less directions** than in the deterministic case ?

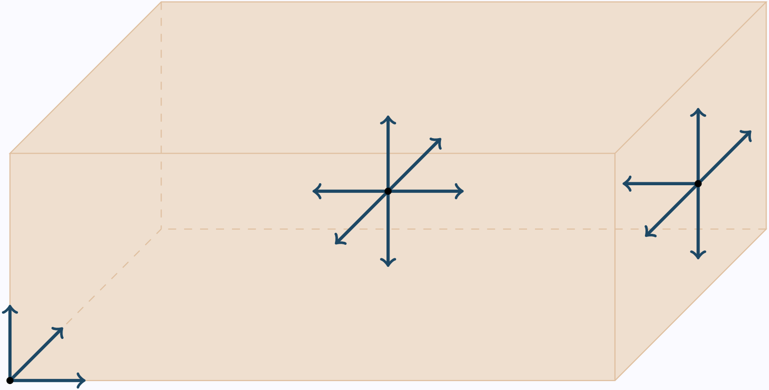
Our approach

- Based on tangent cone generators;
- Choice of a random feasible polling set;
- **At worst as expensive as the deterministic case**.

Possible direction generation techniques

A random generator sampling approach

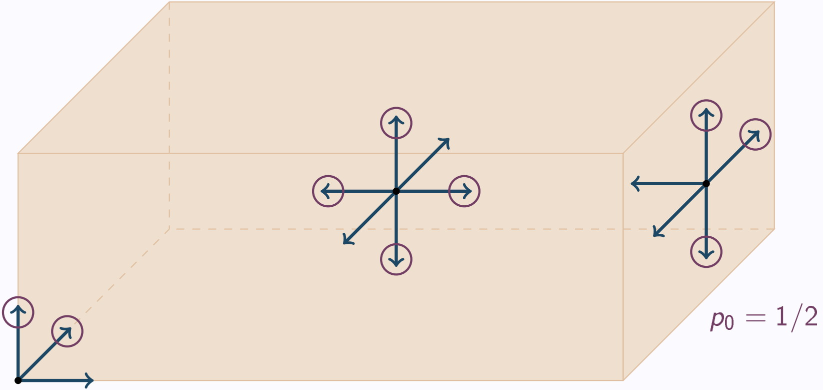
1. Compute a deterministic generating set V_k for T_k;



Possible direction generation techniques

A random generator sampling approach

1. Compute a deterministic generating set V_k for T_k;
2. Take a random sample $\mathfrak{D}_k$ of V_k of size $> |V_k| p_0$;
3. $\{\mathfrak{D}_k\}$ is (p, κ)-descent with $p > p_0$.



Using fewer directions by exploiting subspaces

Idea

- Unconstrained case: probabilistic descent can use less directions;
- Also for linear equalities: unconstrained problem in the null space of A;
- Benefit in exploiting unconstrained subspaces ?

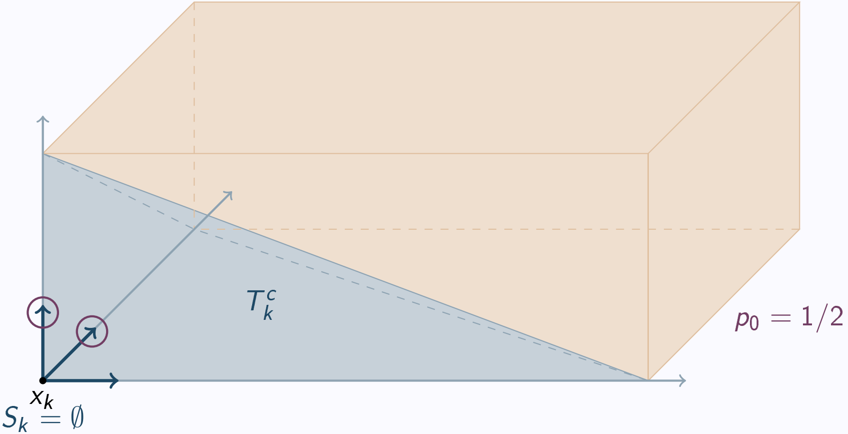
Lemma

Let S_k be a linear subspace within a cone T_k. Then $T_k = S_k + T_k^c$, where T_k^c is a cone lying in $S_k^{\perp}$.

Second technique: Illustration

Two types of directions

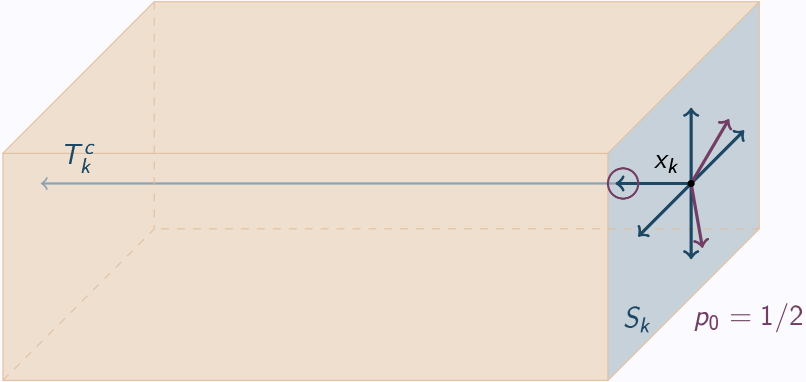
- Subspace S_k: Generate randomly directions;
- Orthogonal part T_k^c: Use a random subset of the generators.



Second technique: Illustration

Two types of directions

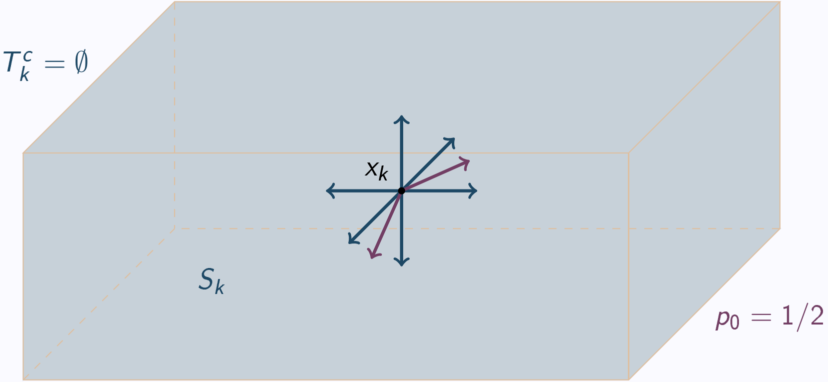
- Subspace S_k: Generate randomly directions;
- Orthogonal part T_k^c: Use a random subset of the generators.



Second technique: Illustration

Two types of directions

- Subspace S_k: Generate randomly directions;
- Orthogonal part T_k^c: Use a random subset of the generators.



Complexity aspects

- The general bound is $\mathcal{O}\left(r\kappa^{-2}\epsilon^{-2}\right)$.

Comparison of results - Linear constraints only

Method	r	κ	Bound
Determ.	$2(n-m)$	$\frac{1}{\sqrt{n-m}}$	$\mathcal{O}\left((n-m)^2\epsilon^{-2}\right)$
Proba. 1	$\mathcal{O}(2(n-m)p_0)$	$\frac{1}{\sqrt{n-m}}$	$\mathcal{O}\left((n-m)^2\epsilon^{-2}\right)$
Proba. 2 (subspace)	$\mathcal{O}(1)$	$\frac{\tau}{\sqrt{n-m}}$	$\mathcal{O}\left((n-m)\epsilon^{-2}\right)$

Comparison of results - Bounds on $n_b < n$ variables only

Method	r	κ	Bound
Determ.	$2n$	$\frac{1}{\sqrt{n}}$	$\mathcal{O}\left(n^2\epsilon^{-2}\right)$
Proba. 1	$\mathcal{O}(2np_0)$	$\frac{1}{\sqrt{n}}$	$\mathcal{O}\left(n^2\epsilon^{-2}\right)$
Proba. 2 (subspace)	$\mathcal{O}(1)+\mathcal{O}\left(n_b\, p_0\right)$	$\frac{1}{\sqrt{n}}$	$\mathcal{O}\left(n\, n_b\epsilon^{-2}\right)$

Numerical experiments-Bound constraints

- Comparison with MATLAB built-in `patternsearch` function.

Four solvers

Name	Polling in $T(x_k, \alpha_k) = T_k = S_k + T_k^c$	Guarantees
`dspfd-0`	Shuffled $D_\oplus \cap T_k$	Deterministic
`dspfd-1`	Random subset of $D_\oplus \cap T_k$	Probabilistic
`dspfd-2`	Random vectors in S_k/subset of $D_\oplus \cap T_k^c$	Probabilistic
`matlab`	$D_\oplus \cap T(x_k, t\alpha_k), t \in (0,1)$	Deterministic

Performance profiles

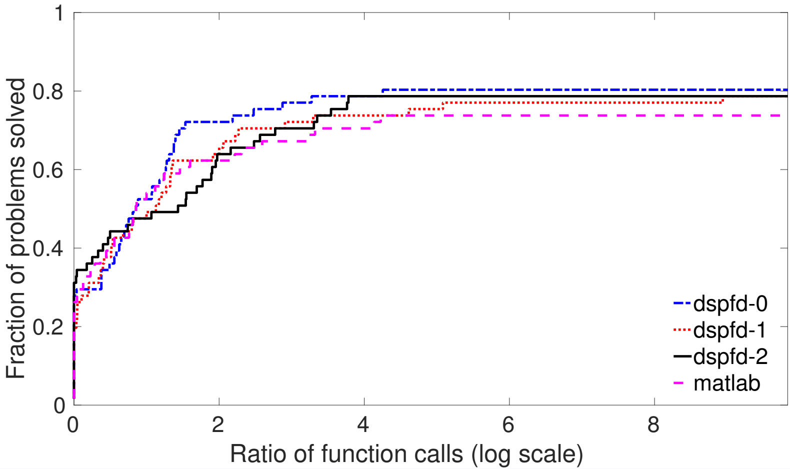
- Criterion: # of function evaluations (budget of $2000n$) to satisfy

$$f(x_k) - f_{best} < 10^{-3}(f(x_0) - f_{best}).$$

- Benchmark: Problems from the CUTEst collection.

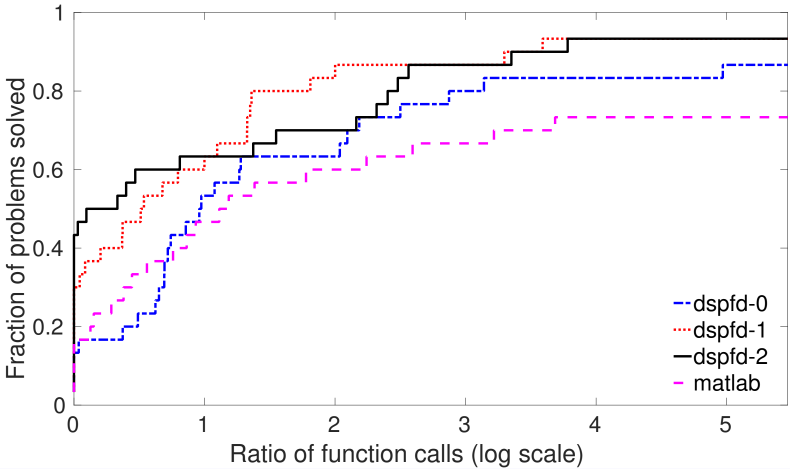
Profiles for bound-constrained problems

- Performance on 63 problems with bounds, small dimensions: $2 \leq n \leq 20$.



Profiles for bound-constrained problems (2)

- Performance on 31 problems with bounds constraints, **larger dimensions:** $20 \leq n \leq 52$.



Outline

Context

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

- f twice continuously differentiable,
- f typically nonconvex.

From first to second-order algorithms

- Expensive to access the Hessian...
- ...or to perform associated procedures:
  - Eigenvectors;
  - Linear systems.

Can probabilistic properties come to the rescue ?

A hot topic

Probabilistic analysis of linear algebra methods

With a random start...

- Power method finds an ϵ-approximate eigenvector in $\mathcal{O}\left(\epsilon^{-1}\right)$ steps;
- Lanczos method finds an ϵ-approximate eigenvector in $\mathcal{O}\left(\epsilon^{-1/2}\right)$ steps.

...**with high probability**.

Uses

- Combined with first-order methods;
- For convex and nonconvex problems.

Second-order convergence

Key points

- Escape saddle points;
- Detect negative Hessian eigenvalues;
- Use **negative curvature directions**;

The best methods in terms of complexity guarantee second-order convergence!

Our approach

- Revisit practically efficient methods;
- Develop a full complexity analysis;
- Complete known observations.

Tools

- Newton-Krylov (e.g. CG) methods:
  Matrix-free, efficient.
- Line search techniques and properties;
- Probabilistic analysis.

Main conclusions and contributions

For unconstrained problems

- Convergence results hold for probabilistic properties.
- Requires less evaluations in theory and practice.

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

Bounds and linear constraints

- Using (probabilistic) feasible descent is the key.
- Random generation in "unconstrained" subspaces.
- Can improve complexity bounds, practically efficient.

  Direct search based on probabilistic feasible descent for bound and linearly constrained problems. Gratton, Royer, Vicente and Zhang, *Submitted*, 2017.

General perspectives

- In derivative-free
  - Nonlinear constraints;
  - Parallel setting.
- More generally
  - Complexity analysis;
  - **Negative curvature**.

General perspectives

- In derivative-free
  - Nonlinear constraints;
  - Parallel setting.
- More generally
  - Complexity analysis;
  - **Negative curvature**.

Thank you for your attention !

croyer2@wisc.edu