Direct Search Based on Probabilistic Descent

Zaikun Zhang

University of Coimbra, moving to CERFACS-IRIT joint lab

Joint work with S. Gratton, C. W. Royer, and L. N. Vicente

SIOPT — May 22, 2014, San Diego

Privilege

It is a privilege to conclude the whole conference!

Problem setting

Unconstrained derivative-free optimization (DFO)

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

$$f : \mathbb{R}^n \to \mathbb{R}$$

f is bounded from below and differentiable
∇f is Lipschitz continuous but unavailable

Problem setting

Unconstrained derivative-free optimization (DFO)

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

$$f : \mathbb{R}^n \to \mathbb{R}$$

f is bounded from below and differentiable
∇f is Lipschitz continuous but unavailable

- Many real-world problems: derivatives are expensive or unreliable.

Problem setting

Unconstrained derivative-free optimization (DFO)

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

$$f : \mathbb{R}^n \to \mathbb{R}$$

f is bounded from below and differentiable
∇f is Lipschitz continuous but unavailable

- Many real-world problems: derivatives are expensive or unreliable.
- S. Gratton, P. Laloyaux, and A. Sartenaer, "Derivative-free Optimization for Large-scale Nonlinear Data Assimilation Problems", 2013.

Derivative-free optimization: What is desirable?

Algorithms that

- do not use derivatives, and

Derivative-free optimization: What is desirable?

Algorithms that

- do not use derivatives, and
- use function evaluations as few as possible.

Derivative-free optimization: What is desirable?

Algorithms that

- do not use derivatives, and
- use function evaluations as few as possible.

And convergence theory of the algorithms.

Existing methods

- Two main classes of rigorous methods in DFO
  - Directional methods, like direct search (`GPS`, `GSS`, `MADS` ...)
  - Model-based methods, like trust region methods (`DFO`, `NEWUOA`, `CONDER`, `BOOSTER`, `ORBIT` ...)

Choose: x_0, α_0, $\gamma \in [1, \infty)$, $\theta \in (0, 1)$, and a forcing function ρ.

Direct search (DS)

Choose: x_0, α_0, $\gamma \in [1, \infty)$, $\theta \in (0, 1)$, and a forcing function ρ.

For $k = 0, 1, 2, \ldots$

- **Polling:** Select a polling set D_k of directions, and seek $d_k \in D_k$:

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

Direct search (DS)

Choose: x_0, α_0, $\gamma \in [1,\infty)$, $\theta \in (0,1)$, and a forcing function ρ.

For $k = 0, 1, 2, \ldots$

- **Polling:** Select a polling set D_k of directions, and seek $d_k \in D_k$:

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

  If d_k is found, the iteration is successful. Otherwise, it is unsuccessful.

Direct search (DS)

Choose: x_0, α_0, $\gamma \in [1, \infty)$, $\theta \in (0, 1)$, and a forcing function ρ.

For $k = 0, 1, 2, \ldots$

- **Polling:** Select a polling set D_k of directions, and seek $d_k \in D_k$:

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

  If d_k is found, the iteration is successful. Otherwise, it is unsuccessful.

- **Update:**

$$x_{k+1} = \begin{cases} x_k + \alpha_k d_k & \text{if successful} \\ x_k & \text{if unsuccessful,} \end{cases}$$

Direct search (DS)

Choose: x_0, α_0, $\gamma \in [1, \infty)$, $\theta \in (0, 1)$, and a forcing function ρ.

For $k = 0, 1, 2, \ldots$

- **Polling:** Select a polling set D_k of directions, and seek $d_k \in D_k$:

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

If d_k is found, the iteration is successful. Otherwise, it is unsuccessful.

- **Update:**

$$x_{k+1} = \begin{cases} x_k + \alpha_k d_k & \text{if successful} \\ x_k & \text{if unsuccessful,} \end{cases}$$

$$\alpha_{k+1} = \begin{cases} \gamma\alpha_k & \text{if successful} \\ \theta\alpha_k & \text{if unsuccessful.} \end{cases}$$

Direct search (DS)

Choose: x_0, α_0, $\gamma \in [1, \infty)$, $\theta \in (0, 1)$, and a forcing function ρ.

For $k = 0, 1, 2, \ldots$

- **Polling:** Select a polling set D_k of directions, and seek $d_k \in D_k$:

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

  If d_k is found, the iteration is successful. Otherwise, it is unsuccessful.

- **Update:**

$$x_{k+1} = \begin{cases} x_k + \alpha_k d_k & \text{if successful} \\ x_k & \text{if unsuccessful,} \end{cases}$$

$$\alpha_{k+1} = \begin{cases} \gamma \alpha_k & \text{if successful} \\ \theta \alpha_k & \text{if unsuccessful.} \end{cases}$$

More ...

A forcing function ρ is a positive and monotonically nondecreasing function such that

$$\lim_{\alpha \downarrow 0} \frac{\rho(\alpha)}{\alpha} = 0.$$

More ...

A forcing function ρ is a positive and monotonically nondecreasing function such that

$$\lim_{\alpha \downarrow 0} \frac{\rho(\alpha)}{\alpha} = 0.$$

In this talk:

$$\rho(\alpha) = \frac{\alpha^2}{2}$$

$$\alpha_0 = 1 \qquad \text{(initial stepsize)}$$

$$\gamma = 2 \qquad \text{(increasing factor)}$$

$$\theta = \frac{1}{2} \qquad \text{(decreasing factor)}$$

Traditional polling set: PSS

- Positive spanning set (PSS):

  $D = \{d_1, \ldots, d_m\}$ is a PSS if it spans $\mathbb{R}^n$ positively:

$$\mathbb{R}^n = \left\{ \sum_{i=1}^{m} \mu_i d_i : \mu_i \geq 0 \ (1 \leq i \leq m) \right\}.$$

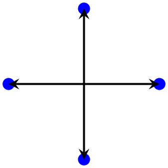
Traditional polling set: PSS

- Positive spanning set (PSS):

  $D = \{d_1, \ldots, d_m\}$ is a PSS if it spans $\mathbb{R}^n$ positively:

$$\mathbb{R}^n = \left\{ \sum_{i=1}^{m} \mu_i d_i : \mu_i \geq 0 \ (1 \leq i \leq m) \right\}.$$

Example:



$$D_\oplus = \{e_1, \ldots, e_n, -e_1, \ldots, -e_n\}$$

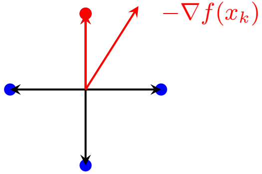
Traditional polling set: PSS

- Positive spanning set (PSS):

  $D = \{d_1, \ldots, d_m\}$ is a PSS if it spans $\mathbb{R}^n$ positively:

$$\mathbb{R}^n = \left\{ \sum_{i=1}^{m} \mu_i d_i : \mu_i \geq 0 \ (1 \leq i \leq m) \right\}.$$

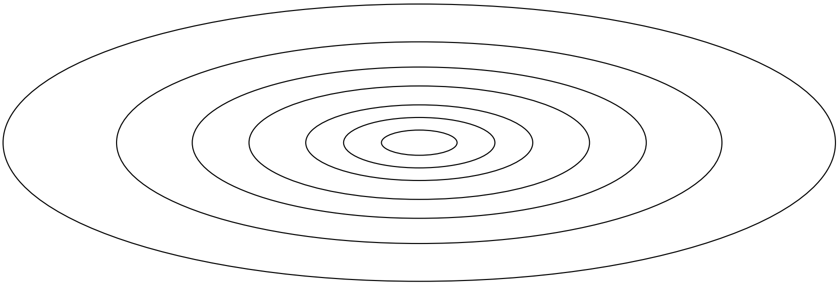
  Example:



$$D_{\oplus} = \{e_1, \ldots, e_n, -e_1, \ldots, -e_n\}$$

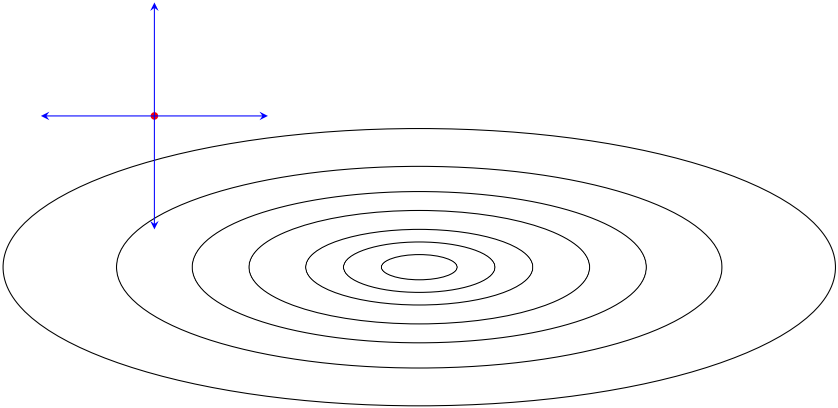
- $\exists\, d \in D$ that 'approximates' $-\nabla f(x_k)$, meaning $d^\top[-\nabla f(x_k)] > 0$.

Example: Coordinate search



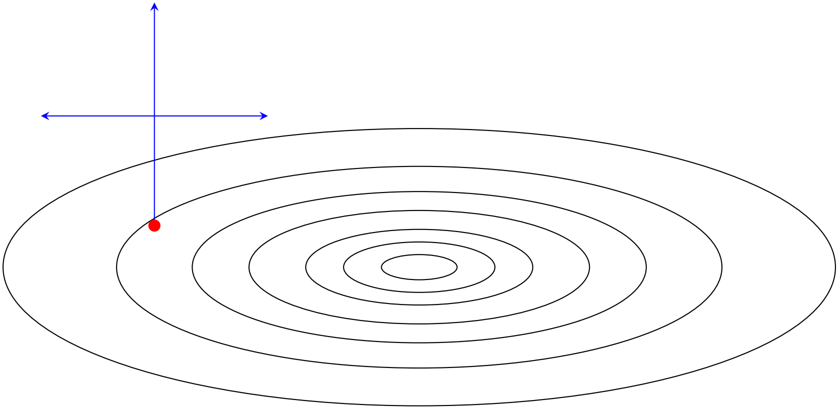
$$n = 2, \quad D = D_{\oplus} = \{e_1, e_2, -e_1, -e_2\}$$

Example: Coordinate search



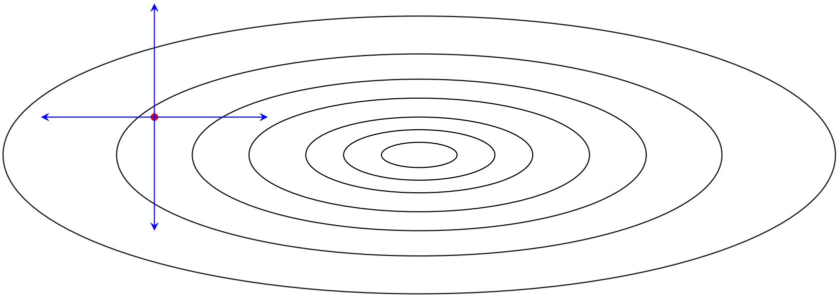
$$n = 2, \quad D = D_{\oplus} = \{e_1, e_2, -e_1, -e_2\}$$

Example: Coordinate search

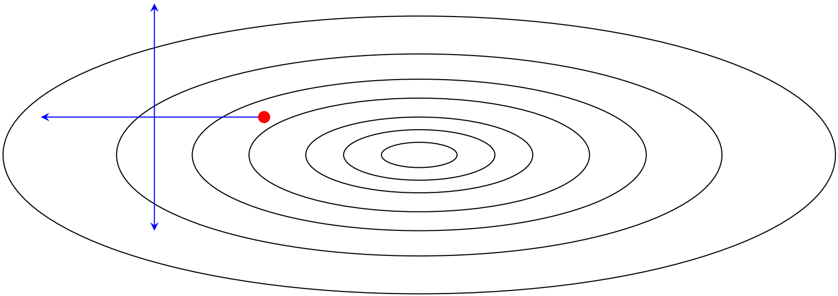


$$n = 2, \quad D = D_{\oplus} = \{e_1, e_2, -e_1, -e_2\}$$

Example: Coordinate search

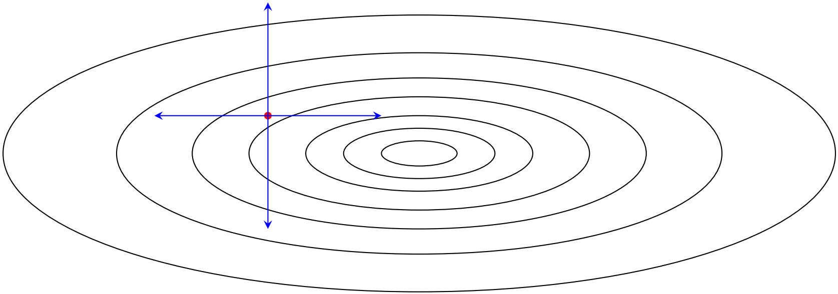


$$n = 2, \quad D = D_{\oplus} = \{e_1, e_2, -e_1, -e_2\}$$



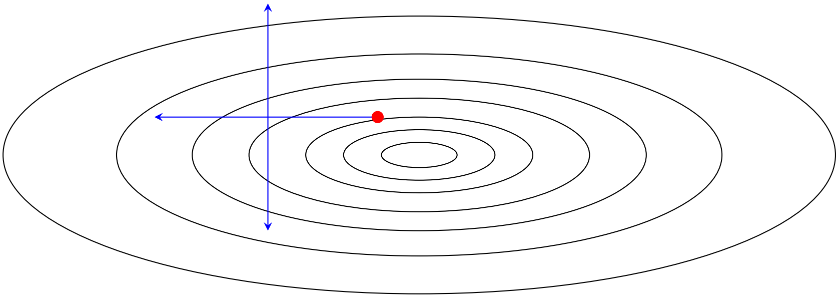
$$n = 2, \quad D = D_{\oplus} = \{e_1, e_2, -e_1, -e_2\}$$

Example: Coordinate search



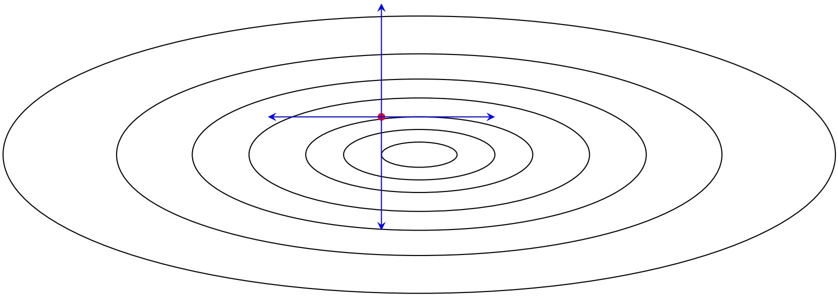
$$n = 2, \quad D = D_{\oplus} = \{e_1, e_2, -e_1, -e_2\}$$

Example: Coordinate search



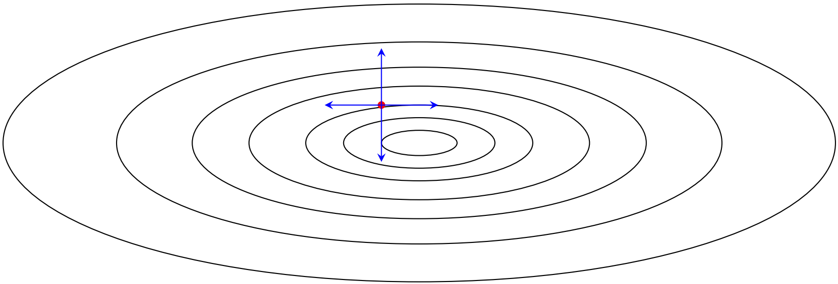
$$n = 2, \quad D = D_{\oplus} = \{e_1, e_2, -e_1, -e_2\}$$

Example: Coordinate search



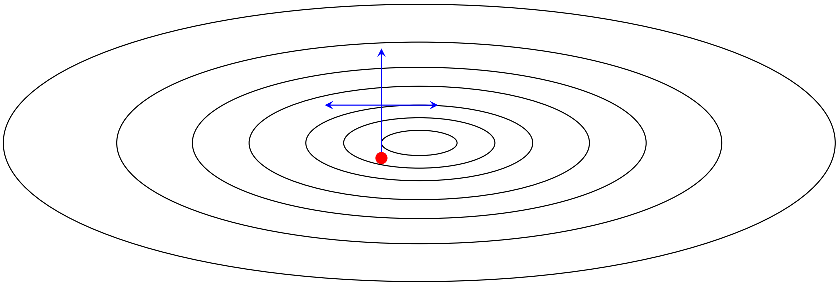
$$n = 2, \quad D = D_\oplus = \{e_1, e_2, -e_1, -e_2\}$$

Example: Coordinate search



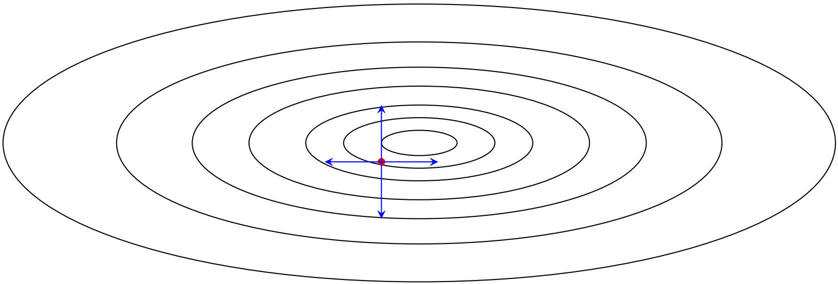
$$n = 2, \quad D = D_{\oplus} = \{e_1, e_2, -e_1, -e_2\}$$

Example: Coordinate search



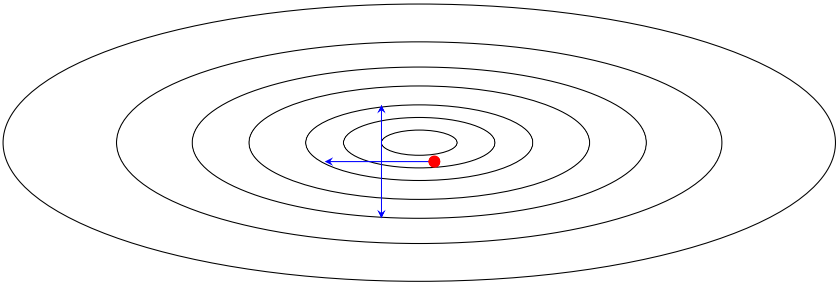
$$n = 2, \quad D = D_{\oplus} = \{e_1, e_2, -e_1, -e_2\}$$

Example: Coordinate search



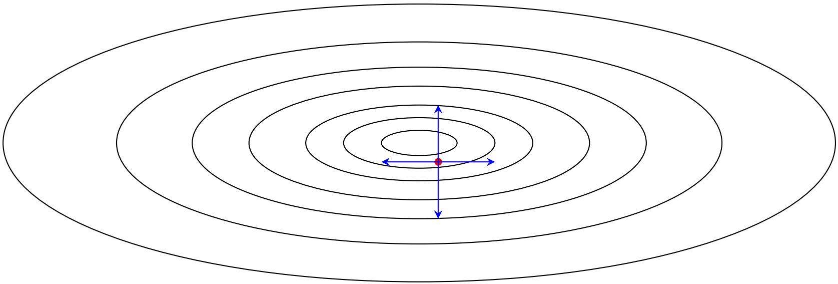
$$n = 2, \quad D = D_{\oplus} = \{e_1, e_2, -e_1, -e_2\}$$

Example: Coordinate search



$$n = 2, \quad D = D_{\oplus} = \{e_1, e_2, -e_1, -e_2\}$$

Example: Coordinate search



$$n = 2, \quad D = D_\oplus = \{e_1, e_2, -e_1, -e_2\}$$

The quality of a PSS: Cosine measure

- Cosine measure: the ability of D to 'approximate' directions in $\mathbb{R}^n$.

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

The quality of a PSS: Cosine measure

- Cosine measure: the ability of D to 'approximate' directions in $\mathbb{R}^n$.

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

- Cosine of the largest angle 'between D and the vectors in $\mathbb{R}^n$'.

The quality of a PSS: Cosine measure

- Cosine measure: the ability of D to 'approximate' directions in $\mathbb{R}^n$.

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

- Cosine of the largest angle 'between D and the vectors in $\mathbb{R}^n$'.

- For each $v \in \mathbb{R}^n$, there exists $d \in D$ such that

$$d^\top v \;\geq\; \mathrm{cm}(D) \|d\| \|v\|,$$

The quality of a PSS: Cosine measure

- Cosine measure: the ability of D to 'approximate' directions in $\mathbb{R}^n$.

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

- Cosine of the largest angle 'between D and the vectors in $\mathbb{R}^n$'.

- For each $v \in \mathbb{R}^n$, there exists $d \in D$ such that

$$d^\top v \;\geq\; \mathrm{cm}(D)\|d\|\|v\|,$$

especially when $v = -\nabla f(x_k)$.

- Cosine measure: the ability of D to 'approximate' directions in $\mathbb{R}^n$.

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

- Cosine of the largest angle 'between D and the vectors in $\mathbb{R}^n$'.

- For each $v \in \mathbb{R}^n$, there exists $d \in D$ such that

$$d^\top v \;\geq\; \mathrm{cm}(D)\|d\|\|v\|,$$

especially when $v = -\nabla f(x_k)$.

- Example:

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

DS with PSS: Theory

Let $\{D_k\}$ be a sequence of PSSs such that $\operatorname{cm}(D_k) \geq \kappa > 0$ for each k.

DS with PSS: Theory

Let $\{D_k\}$ be a sequence of PSSs such that $\mathrm{cm}(D_k) \geq \kappa > 0$ for each k.

Global converence:

Theorem (Torczon 1997, Kolda, Lewis, and Torczon 2003)

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

DS with PSS: Theory

Let $\{D_k\}$ be a sequence of PSSs such that $\mathrm{cm}(D_k) \geq \kappa > 0$ for each k.

Global converence:

Theorem (Torczon 1997, Kolda, Lewis, and Torczon 2003)

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

Global rate and worst case complexity (WCC):

Global rate and worst case complexity (Vicente 2013)

- $\min_{0\leq \ell \leq k} \|\nabla f(x_\ell)\| = \mathcal{O}(1/\sqrt{k})$.
- $\|\nabla f(x_k)\|$ *is driven under* ϵ *within* $\mathcal{O}(\epsilon^{-2})$ *iterations.*

DS with PSS: Theory

Let $\{D_k\}$ be a sequence of PSSs such that $\mathrm{cm}(D_k) \geq \kappa > 0$ for each k.

Global converence:

Theorem (Torczon 1997, Kolda, Lewis, and Torczon 2003)

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

Global rate and worst case complexity (WCC):

Global rate and worst case complexity (Vicente 2013)

- $\min_{0\leq\ell\leq k} \|\nabla f(x_\ell)\| = \mathcal{O}(1/\sqrt{k}).$
- $\|\nabla f(x_k)\|$ *is driven under* ϵ *within* $\mathcal{O}(\epsilon^{-2})$ *iterations.*

Question: Does the theory cover the most efficient implementation of DS?

A competitor against PSS: Random polling set

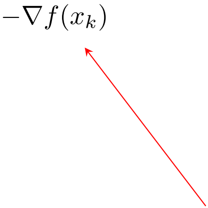
- Success of random coordinate descent, stochastic gradient . . .
  - Y. Nesterov, "Efficiency of coordinate descent methods on huge-scale optimization problems", SIAM Journal on Optimization, 22(2), 341-362, 2012
  - P. Richtárik, M. Takáč, "Iteration complexity of randomized block-coordinate descent methods for minimizing a composite function." Mathematical Programming 144(1-2): 1-38, 2014
  - Z. Lu, and L. Xiao. "On the complexity analysis of randomized block-coordinate descent methods", no. MSR-TR-2013-53, May 2013
  - . . .

A competitor against PSS: Random polling set

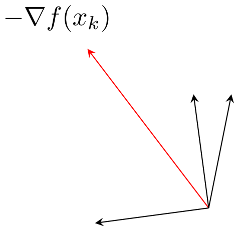
- Success of random coordinate descent, stochastic gradient . . .
  - Y. Nesterov, "Efficiency of coordinate descent methods on huge-scale optimization problems", SIAM Journal on Optimization, 22(2), 341-362, 2012
  - P. Richtárik, M. Takáč, "Iteration complexity of randomized block-coordinate descent methods for minimizing a composite function." Mathematical Programming 144(1-2): 1-38, 2014
  - Z. Lu, and L. Xiao. "On the complexity analysis of randomized block-coordinate descent methods", no. MSR-TR-2013-53, May 2013
  - . . .

- Sucess of randomization in derivative-free methods, with insightful theories:
  - A. S. Bandeira, K. Scheinberg, and L. N. Vicente, "Convergence of trust-region methods based on probabilistic models", submitted
  - K. Scheinberg, "Convergence rates of line-search and trust region methods based on probabilistic models", MS16, SIOPT14

Using random polling sets: What will happen?



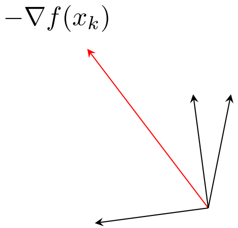
Using random polling sets: What will happen?



$n+1$ random polling directions

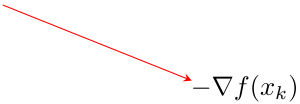
in this case not a PSS

Using random polling sets: What will happen?

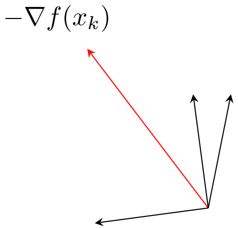


$n+1$ random polling directions

in this case not a PSS

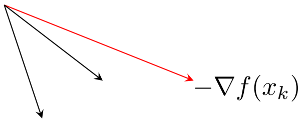


Using random polling sets: What will happen?



$n+1$ random polling directions

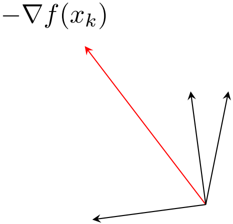
in this case not a PSS



$\leq n$ random polling directions

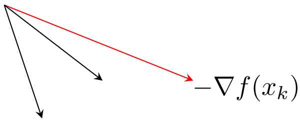
certainly not a PSS ...

Using random polling sets: What will happen?



$n+1$ random polling directions

in this case not a PSS



$\leq n$ random polling directions

certainly not a PSS ...

D_k is 'good' in some probabilistic sense ...

What do we mean by 'good'?

If derivatives were available, it would have been sufficient to require

$$\max_{d \in D} \frac{-d^\top \nabla f(x_k)}{\|d\| \|\nabla f(x_k)\|} \;\geq\; \kappa.$$

What do we mean by 'good'?

If derivatives were available, it would have been sufficient to require

$$\max_{d \in D} \frac{-d^\top \nabla f(x_k)}{\|d\| \|\nabla f(x_k)\|} \;\geq\; \kappa.$$

Define the cosine measure of D given v by

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

Then $\mathrm{cm}(D, -\nabla f(x_k)) \geq \kappa$ would have been enough.

What do we mean by 'good'?

If derivatives were available, it would have been sufficient to require

$$\max_{d\in D} \frac{-d^\top \nabla f(x_k)}{\|d\| \|\nabla f(x_k)\|} \;\geq\; \kappa.$$

Define the cosine measure of D given v by

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

Then $\mathrm{cm}(D, -\nabla f(x_k)) \geq \kappa$ would have been enough.

But derivatives are not available!

Random variables v.s. realizations

From now on, we suppose that the polling directions are not defined deterministically but taken at random from the unit sphere $\mathcal{S}^{n-1}$.

Random variables v.s. realizations

From now on, we suppose that the polling directions are not defined deterministically but taken at random from the unit sphere $\mathcal{S}^{n-1}$.

Distinguish random variables from realizations

	Iterate	Polling set
Random variables	X_k	$\mathfrak{D}_k$
Realizations	x_k	D_k

What is desirable?

- Global convergence:

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

What is desirable?

- Global convergence:

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

- Global rate:

$$\mathbb{P}\left(\min_{0\leq \ell \leq k} \|\nabla f(X_k)\| \leq \frac{C}{\sqrt{k}}\right) \text{ is 'high'}$$

for some properly selected constant C?

How to achieve the goals?

- Global convergence:

$$\left\{ \liminf_{k\to\infty} \|\nabla f(X_k)\| > 0 \right\} \subset E$$

with $\mathbb{P}(E) = 0$.

How to achieve the goals?

- Global convergence:

$$\left\{ \liminf_{k\to\infty} \|\nabla f(X_k)\| > 0 \right\} \subset E$$

with $\mathbb{P}(E) = 0$.

- Global rate:

$$\left\{ \min_{0\le\ell\le k} \|\nabla f(X_k)\| > \epsilon \right\} \subset E_{k,\epsilon},$$

with $\mathbb{P}(E_{k,\epsilon})$ being "low" when $\epsilon = C/\sqrt{k}$.

How to achieve the goals?

- Global convergence:

$$\left\{ \liminf_{k\to\infty} \|\nabla f(X_k)\| > 0 \right\} \subset E$$

with $\mathbb{P}(E) = 0$.

- Global rate:

$$\left\{ \min_{0\leq \ell \leq k} \|\nabla f(X_k)\| > \epsilon \right\} \subset E_{k,\epsilon},$$

with $\mathbb{P}(E_{k,\epsilon})$ being "low" when $\epsilon = C/\sqrt{k}$.

Let us find E and $E_{k,\epsilon}$ ...

Global convergence: An intuitive lemma

Let Z_k be the indicator function of $\left\{\mathrm{cm}\big(\mathfrak{D}_k, -\nabla f(X_k)\big) \ge \kappa\right\}$, and

$$p_0 \;=\; \frac{\ln\theta}{\ln(\gamma^{-1}\theta)} \;=\; \frac{1}{2}.$$

Global convergence: An intuitive lemma

Let Z_k be the indicator function of $\left\{\mathrm{cm}\big(\mathfrak{D}_k, -\nabla f(X_k)\big) \geq \kappa\right\}$, and

$$p_0 = \frac{\ln\theta}{\ln(\gamma^{-1}\theta)} = \frac{1}{2}.$$

Without imposing any assumption on the probabilistic behavior of $\{\mathfrak{D}_k\}$:

Lemma

$$\left\{\liminf_{k\to\infty} \|\nabla f(X_k)\| > 0\right\} \subset \left\{\sum_{k=0}^{\infty}(Z_k - p_0) = -\infty\right\}.$$

Global convergence: An intuitive lemma

Let Z_k be the indicator function of $\{\mathrm{cm}(\mathfrak{D}_k, -\nabla f(X_k)) \geq \kappa\}$, and

$$p_0 = \frac{\ln\theta}{\ln(\gamma^{-1}\theta)} = \frac{1}{2}.$$

Without imposing any assumption on the probabilistic behavior of $\{\mathfrak{D}_k\}$:

Lemma

$$\left\{\liminf_{k\to\infty} \|\nabla f(X_k)\| > 0\right\} \subset \left\{\sum_{k=0}^{\infty}(Z_k - p_0) = -\infty\right\}.$$

Meaning:

If convergence does not hold, the 'frequency' of $\{Z_k\}_{k\geq 0}$ is 'less than p_0'.

Global rate: Another intuitive lemma

Without imposing any assumption on the probabilistic behavior of $\{\mathfrak{D}_k\}$:

Lemma

$$\left\{ \max_{0 \le \ell \le k} \|\nabla f(X_k)\| > \epsilon \right\} \subset \left\{ \sum_{\ell=0}^{k-1} Z_\ell \le \left[\frac{(L+1)^2 \beta}{2\kappa^2 \epsilon^2 k} + p_0 \right] k \right\}.$$

$\beta < \infty$ is an upper bound for $\sum_{k=0}^{\infty} \rho(\alpha_k)$ (existence guaranteed).

$L < \infty$ is a Lipshitz constant of ∇f in $\mathbb{R}^n$.

Global rate: Another intuitive lemma

Without imposing any assumption on the probabilistic behavior of $\{\mathfrak{D}_k\}$:

Lemma

$$\left\{\max_{0\le\ell\le k}\|\nabla f(X_k)\|>\epsilon\right\}\subset\left\{\sum_{\ell=0}^{k-1}Z_\ell\le\left[\frac{(L+1)^2\beta}{2\kappa^2\epsilon^2k}+p_0\right]k\right\}.$$

$\beta<\infty$ is an upper bound for $\sum_{k=0}^{\infty}\rho(\alpha_k)$ (existence guaranteed).

$L<\infty$ is a Lipshitz constant of ∇f in $\mathbb{R}^n$.

Meaning:

If $\{\|\nabla f(X_0)\|\}_{0\le\ell\le k}$ are all above ϵ, the 'frequency' of $\{Z_\ell\}_{0\le\ell\le k-1}$ is 'not more than' $p_0+\mathcal{O}(\epsilon^{-2}k^{-1})$.

What assumptions to impose?

Until now, no assumption is imposed on the probabilistic behavior of $\{\mathfrak{D}_k\}$.

What assumptions to impose?

Until now, no assumption is imposed on the probabilistic behavior of $\{\mathfrak{D}_k\}$.

Definition

The sequence $\{\mathfrak{D}_k\}$ is p-probabilistically κ-descent if, for each $k \geq 0$,

$$\mathbb{P}\Big(\mathrm{cm}(\mathfrak{D}_k, -\nabla f(X_k)) \;\geq\; \kappa \mid \mathfrak{D}_0, \ldots, \mathfrak{D}_{k-1}\Big) \geq \; p.$$

Global convergence

Lemma

If $\{\mathfrak{D}_k\}$ *is* p_0*-probabilistically* κ*-descent, then* $\left\{ \sum_{\ell=0}^{k-1} (Z_\ell - p_0) \right\}$ *is a submartingale, and*

$$\mathbb{P}\left(\sum_{k=0}^{\infty} (Z_k - p_0) = -\infty \right) = 0.$$

Global convergence

Lemma

If $\{\mathfrak{D}_k\}$ is p_0-probabilistically κ-descent, then $\left\{ \sum_{\ell=0}^{k-1} (Z_\ell - p_0) \right\}$ is a submartingale, and

$$\mathbb{P}\left(\sum_{k=0}^{\infty} (Z_k - p_0) = -\infty \right) = 0.$$

Theorem

If $\{\mathfrak{D}_k\}$ is p_0-probabilistically κ-descent, then

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

The analysis is inspired by that for probabilistic trust region.

Global rate

Lemma (Chernoff bound)

Suppose that $\{\mathfrak{D}_k\}$ is p-probabilistically κ-descent and $\lambda \in (0, p)$. Then

$$\mathbb{P}\left(\sum_{\ell=0}^{k-1} Z_\ell \le \lambda k\right) \;\le\; \exp\left[-\frac{(p-\lambda)^2}{2p}k\right].$$

Global rate

Lemma (Chernoff bound)

Suppose that $\{\mathfrak{D}_k\}$ *is* *p-probabilistically κ-descent* *and* $\lambda \in (0, p)$. *Then*

$$\mathbb{P}\left(\sum_{\ell=0}^{k-1} Z_\ell \le \lambda k\right) \;\le\; \exp\left[-\frac{(p-\lambda)^2}{2p} k\right].$$

Theorem

Suppose that $\{\mathfrak{D}_k\}$ *is* *p-probabilistically κ-descent* *with* $p > p_0$. *Then*

$$\mathbb{P}\left(\min_{0\le\ell\le k} \|\nabla f(X_\ell)\| \le \left[\frac{(L+1)\beta^{\frac{1}{2}}}{(p-p_0)^{\frac{1}{2}}\kappa}\right]\frac{1}{\sqrt{k}}\right) \;\ge\; 1 - \exp\left[-\frac{(p-p_0)^2}{8p} k\right].$$

Global rate

Lemma (Chernoff bound)

Suppose that $\{\mathfrak{D}_k\}$ *is* p*-probabilistically* κ*-descent and* $\lambda \in (0, p)$*. Then*

$$\mathbb{P}\left(\sum_{\ell=0}^{k-1} Z_\ell \le \lambda k\right) \;\le\; \exp\left[-\frac{(p-\lambda)^2}{2p}k\right].$$

Theorem

Suppose that $\{\mathfrak{D}_k\}$ *is* p*-probabilistically* κ*-descent with* $p > p_0$*. Then*

$$\mathbb{P}\left(\min_{0\le\ell\le k} \|\nabla f(X_\ell)\| \le \left[\frac{(L+1)\beta^{\frac{1}{2}}}{(p-p_0)^{\frac{1}{2}}\kappa}\right]\frac{1}{\sqrt{k}}\right) \;\ge\; 1-\exp\left[-\frac{(p-p_0)^2}{8p}k\right].$$

$\Longrightarrow$ $\mathcal{O}(1/\sqrt{k})$ decaying rate for gradient holds with overwhelmingly high probability, matching the deterministic case (Vicente 2013).

For each $k \geq 0$,

- $\mathfrak{D}_k$ is independent of the previous iterations, For each $k \geq 0$,

- $\mathfrak{D}_k$ is independent of the previous iterations,
- $\mathfrak{D}_k$ is a set $\{\mathfrak{d}_1, \ldots, \mathfrak{d}_m\}$ of independent random vectors uniformly distributed on the unit sphere.

Practical probabilistic descent sets

$\{\mathfrak{D}_k\}$ generated in this way is probabilistically descent.

Proposition

Given $\tau \in [0, \sqrt{n}]$, $\{\mathfrak{D}_k\}$ *is* p*-probabilistically* $(\tau/\sqrt{n})$*-descent with*

$$p \;=\; 1 - \left(\frac{1}{2} + \frac{\tau}{\sqrt{2\pi}}\right)^{m}.$$

Practical probabilistic descent sets

$\{\mathfrak{D}_k\}$ generated in this way is probabilistically descent.

Proposition

Given $\tau \in [0, \sqrt{n}]$, $\{\mathfrak{D}_k\}$ *is* p*-probabilistically* $(\tau/\sqrt{n})$*-descent with*

$$p \;=\; 1 - \left(\frac{1}{2} + \frac{\tau}{\sqrt{2\pi}}\right)^m.$$

For instance,

$$\left.\begin{array}{rcl} m & = & 2 \\ \tau & = & \dfrac{1}{2} \end{array}\right\} \quad\Longrightarrow\quad p \;>\; \frac{1}{2} \;=\; p_0.$$

Practical probabilistic descent sets: WCC bounds

Plugging $\kappa = 1/(2\sqrt{n})$ into the global rate, one obtains

WCC (number of iterations)

$$\mathbb{P}\left(K_\epsilon \le \left\lceil \frac{4(L+1)^2\beta}{p-p_0}(n\epsilon^{-2}) \right\rceil\right) \ge 1 - \exp\left[-\frac{\beta(p-p_0)(L+1)^2}{2p}(n\epsilon^{-2})\right].$$

Practical probabilistic descent sets: WCC bounds

Plugging $\kappa = 1/(2\sqrt{n})$ into the global rate, one obtains

WCC (number of iterations)

$$\mathbb{P}\left(K_\epsilon \leq \left\lceil \frac{4(L+1)^2\beta}{p-p_0}(n\epsilon^{-2}) \right\rceil\right) \geq 1 - \exp\left[-\frac{\beta(p-p_0)(L+1)^2}{2p}(n\epsilon^{-2})\right].$$

What about the number of function evaluations?

Practical probabilistic descent sets: WCC bounds

Plugging $\kappa = 1/(2\sqrt{n})$ into the global rate, one obtains

WCC (number of iterations)

$$\mathbb{P}\left(K_\epsilon \leq \left\lceil \frac{4(L+1)^2\beta}{p-p_0}(n\epsilon^{-2}) \right\rceil\right) \geq 1 - \exp\left[-\frac{\beta(p-p_0)(L+1)^2}{2p}(n\epsilon^{-2})\right].$$

What about the number of function evaluations?

WCC (number of function evaluations)

$$\mathbb{P}\left(K_\epsilon^f \leq 2\left\lceil \frac{4(L+1)^2\beta}{p-p_0}(n\epsilon^{-2}) \right\rceil\right) \geq 1 - \text{ the tiny tail}.$$

$\Longrightarrow$ $\mathcal{O}(n\epsilon^{-2})$ with overwhelmingly high probability, better than the deterministic case $\mathcal{O}(n^2\epsilon^{-2})$ (Vicente 2013).

The competition

Relative performance: PSS v.s. Random polling sets ($n = 40$)

	$D_\oplus$	$2n$	$n+1$	$n/4$	2	1
arglina	3.42	10.30	6.01	1.88	1.00	–
arglinb	20.50	7.38	2.81	1.85	1.00	2.04
broydn3d	4.33	6.54	3.59	1.28	1.00	–
dqrtic	7.16	9.10	4.56	1.70	1.00	–
engval1	10.53	11.90	6.48	2.08	1.00	2.08
freuroth	56.00	1.00	1.67	1.67	1.00	4.00
integreq	16.04	12.44	6.76	2.04	1.00	–
nondquar	6.90	7.56	4.23	1.87	1.00	–
sinquad	–	1.65	2.01	1.00	1.55	–
vardim	1.00	1.80	2.40	1.80	1.80	4.30

Solution accuracy was 10^{-3}. Averages were taken over 10 independent runs.

Concluding remarks

- Probabilitic DS enjoys, with overwhelmingly high probability:
  - the same WCC for number of iterations,

Concluding remarks

- Probabilitic DS enjoys, with overwhelmingly high probability:
  - the same WCC for number of iterations,
  - possibly better WCC for number of function evaluations.

Concluding remarks

- Probabilitic DS enjoys, with overwhelmingly high probability:
  - the same WCC for number of iterations,
  - possibly better WCC for number of function evaluations.
- The analysis technique can be applied to probabilistic trust region method $\Longrightarrow$ $\mathcal{O}(1/\sqrt{k})$ rate for gradient.

Concluding remarks

- Probabilitic DS enjoys, with overwhelmingly high probability:
  - the same WCC for number of iterations,
  - possibly better WCC for number of function evaluations.
- The analysis technique can be applied to probabilistic trust region method $\Longrightarrow$ $\mathcal{O}(1/\sqrt{k})$ rate for gradient.
- An interesting future work: randomized subspace method.

Concluding remarks

- Probabilitic DS enjoys, with overwhelmingly high probability:
  - the same WCC for number of iterations,
  - possibly better WCC for number of function evaluations.
- The analysis technique can be applied to probabilistic trust region method $\Longrightarrow$ $\mathcal{O}(1/\sqrt{k})$ rate for gradient.
- An interesting future work: randomized subspace method. Let $\mathrm{Gr}(l, \mathbb{R}^n)$ be the set of all the l-dim linear subspaces of $\mathbb{R}^n$.

Lemma (Randomized subspace)

Suppose that S is uniformly distributed on $\mathrm{Gr}(l, \mathbb{R}^n)$. Then for any nonzero vector $v \in \mathbb{R}^n$ and constant $\delta \in (0,1)$,

$$\mathbb{P}\left(\|Pv\| \geq \sqrt{\frac{l\delta}{n}}\|v\|\right) \;\geq\; 1 - \exp\left[-\frac{l}{2}(\delta - 1 - \ln\delta)\right].$$