Non-convergence Analysis of Probabilistic Direct Search

2nd Derivative-Free Optimization Symposium

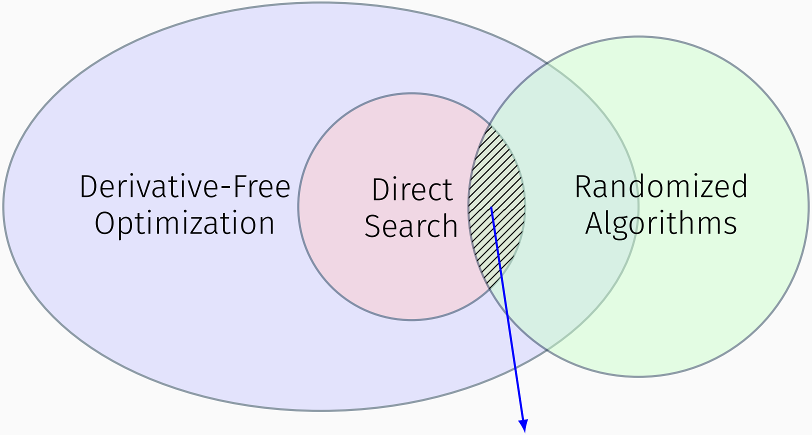
Cunxin Huang

Supervised by Prof. Xiaojun Chen and Dr. Zaikun Zhang

Padova, Italy June 28, 2024

The Hong Kong Polytechnic University

Brief introduction to Probabilistic Direct Search



The algorithm we consider in this talk:
Probabilistic Direct Search (PDS)
(Gratton, Royer, Vicente, and Zhang 2015)

Apologies

- To everyone: Venice is so beautiful that I cannot help but get lost at the last minute.

Apologies

- To everyone: Venice is so beautiful that I cannot help but get lost at the last minute.
- To Clément: my bad title may give you a sense that your paper with Zaikun is wrong.

Apologies

- To everyone: Venice is so beautiful that I cannot help but get lost at the last minute.
- To Clément: my bad title may give you a sense that your paper with Zaikun is wrong.
- To Zaikun: recall his words in his talk “I always tell my students that DFO is vivid because of its applications.”

Apologies

- To everyone: Venice is so beautiful that I cannot help but get lost at the last minute.
  Solution: I will skip some slides to save time.
- To Clément: my bad title may give you a sense that your paper with Zaikun is wrong.
- To Zaikun: recall his words in his talk “I always tell my students that DFO is vivid because of its applications.”

Apologies

- To everyone: Venice is so beautiful that I cannot help but get lost at the last minute.
  Solution: I will skip some slides to save time.
- To Clément: my bad title may give you a sense that your paper with Zaikun is wrong.
  Solution: my talk will show that your theorem is correct and tight.
- To Zaikun: recall his words in his talk "I always tell my students that DFO is vivid because of its applications."

Apologies

- To everyone: Venice is so beautiful that I cannot help but get lost at the last minute.
  Solution: I will skip some slides to save time.
- To Clément: my bad title may give you a sense that your paper with Zaikun is wrong.
  Solution: my talk will show that your theorem is correct and tight.
- To Zaikun: recall his words in his talk "I always tell my students that DFO is vivid because of its applications."
  Solution: I will show some computation works at the end.

What is Derivative-Free Optimization and why

Derivative-Free Optimization (DFO)

- Do not use derivatives (first-order info.), only use function values
- Also called: zeroth-order/black-box/simulation-based optimization

What is Derivative-Free Optimization and why

Derivative-Free Optimization (DFO)

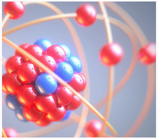
- Do not use derivatives (first-order info.), only use function values
- Also called: zeroth-order/black-box/simulation-based optimization

Derivatives are often not available in applications

Derivative-Free Optimization (DFO)

- Do not use derivatives (first-order info.), only use function values
- Also called: zeroth-order/black-box/simulation-based optimization

Derivatives are often not available in applications



Nuclear Physics

Machine Learning

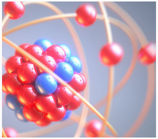
Circuit Design

What is Derivative-Free Optimization and why

Derivative-Free Optimization (DFO)

- Do not use derivatives (first-order info.), only use function values
- Also called: zeroth-order/black-box/simulation-based optimization

Derivatives are often not available in applications



Nuclear Physics

Machine Learning

Circuit Design

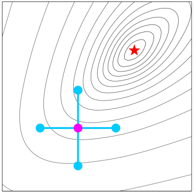
Difficulties

- Problems are often noisy (naive finite difference?)
- Each function evaluation is expensive (e.g., PDE simulation)

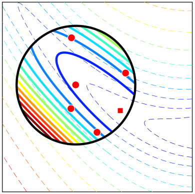
Direct-search methods and model-based methods

How to determine iterates?

- Direct-search methods: "simple" comparison of function values
- Model-based methods: build a surrogate of the objective function



Direct-search methods[1]



Model-based methods[2]

[1]Source: Kolda, Lewis, and Torczon 2003
[2]Source: Larson, Menickelly, and Wild 2019

Probabilistic Direct Search (PDS): a simplified framework

Algorithm 1: Direct Search based on sufficient decrease

Probabilistic Direct Search (PDS): a simplified framework

Algorithm 1: Direct Search based on sufficient decrease

Input: $x_0 \in \mathbb{R}^n$, $\alpha_0 \in (0, \infty)$, $0 < \theta < 1 < \gamma$.

Probabilistic Direct Search (PDS): a simplified framework

Algorithm 1: Direct Search based on sufficient decrease

Input: $x_0 \in \mathbb{R}^n$, $\alpha_0 \in (0, \infty)$, $0 < \theta < 1 < \gamma$.

for $k = 0, 1, \ldots$ **do**

Probabilistic Direct Search (PDS): a simplified framework

Algorithm 1: Direct Search based on sufficient decrease

Input: $x_0 \in \mathbb{R}^n$, $\alpha_0 \in (0, \infty)$, $0 < \theta < 1 < \gamma$.

for $k = 0, 1, \ldots$ **do**

Select a finite set of directions $\mathcal{D}_k \subset \mathbb{R}^n$.

Probabilistic Direct Search (PDS): a simplified framework

Algorithm 1: Direct Search based on sufficient decrease

Input: $x_0 \in \mathbb{R}^n$, $\alpha_0 \in (0, \infty)$, $0 < \theta < 1 < \gamma$.

for $k = 0, 1, \ldots$ **do**

Select a finite set of directions $\mathcal{D}_k \subset \mathbb{R}^n$.

(In this talk, assume $\mathcal{D}_k$ is a set of unit vectors for simplicity)

Probabilistic Direct Search (PDS): a simplified framework

Algorithm 1: Direct Search based on sufficient decrease

Input: $x_0 \in \mathbb{R}^n$, $\alpha_0 \in (0, \infty)$, $0 < \theta < 1 < \gamma$.

for $k = 0, 1, \dots$ **do**

Select a finite set of directions $\mathcal{D}_k \subset \mathbb{R}^n$.

(In this talk, assume $\mathcal{D}_k$ is a set of unit vectors for simplicity)

Set $d_k = \arg\min\{f(x_k + \alpha_k d) : d \in \mathcal{D}_k\}$. (complete polling)

Probabilistic Direct Search (PDS): a simplified framework

Algorithm 1: Direct Search based on sufficient decrease

Input: $x_0 \in \mathbb{R}^n$, $\alpha_0 \in (0, \infty)$, $0 < \theta < 1 < \gamma$.

for $k = 0, 1, \ldots$ **do**

Select a finite set of directions $\mathcal{D}_k \subset \mathbb{R}^n$.

(In this talk, assume $\mathcal{D}_k$ is a set of unit vectors for simplicity)

Set $d_k = \arg\min\{f(x_k + \alpha_k d) : d \in \mathcal{D}_k\}$. (complete polling)

if $f(x_k + \alpha_k d_k) < f(x_k) - c\alpha_k^2$ **then**

Probabilistic Direct Search (PDS): a simplified framework

Algorithm 1: Direct Search based on sufficient decrease

Input: $x_0 \in \mathbb{R}^n$, $\alpha_0 \in (0, \infty)$, $0 < \theta < 1 < \gamma$.

for $k = 0, 1, \ldots$ **do**

  Select a finite set of directions $\mathcal{D}_k \subset \mathbb{R}^n$.

  (In this talk, assume $\mathcal{D}_k$ is a set of unit vectors for simplicity)

  Set $d_k = \arg\min\{f(x_k + \alpha_k d) : d \in \mathcal{D}_k\}$. (complete polling)

  if $f(x_k + \alpha_k d_k) < f(x_k) - c\alpha_k^2$ **then**

    Set $x_{k+1} = x_k + \alpha_k d_k$ and $\alpha_{k+1} = \gamma\alpha_k$.

    (Move and expand step size)

  else

Probabilistic Direct Search (PDS): a simplified framework

Algorithm 1: Direct Search based on sufficient decrease

Input: $x_0 \in \mathbb{R}^n$, $\alpha_0 \in (0, \infty)$, $0 < \theta < 1 < \gamma$.

for $k = 0, 1, \ldots$ **do**

  Select a finite set of directions $\mathcal{D}_k \subset \mathbb{R}^n$.

  (In this talk, assume $\mathcal{D}_k$ is a set of unit vectors for simplicity)

  Set $d_k = \arg\min\{f(x_k + \alpha_k d) : d \in \mathcal{D}_k\}$. (complete polling)

  if $f(x_k + \alpha_k d_k) < f(x_k) - c\alpha_k^2$ **then**

    Set $x_{k+1} = x_k + \alpha_k d_k$ and $\alpha_{k+1} = \gamma\alpha_k$.

    (Move and expand step size)

  else

    Set $x_{k+1} = x_k$ and $\alpha_{k+1} = \theta\alpha_k$.

    (Stay and shrink step size)

Probabilistic Direct Search (PDS): a simplified framework

Algorithm 1: Probabilistic Direct Search based on sufficient decrease

Input: $x_0 \in \mathbb{R}^n$, $\alpha_0 \in (0, \infty)$, $0 < \theta < 1 < \gamma$.

for $k = 0, 1, \ldots$ **do**

  Select a finite set of directions $\mathcal{D}_k \subset \mathbb{R}^n$ randomly.

  (In this talk, assume $\mathcal{D}_k$ is a set of unit vectors for simplicity)

  Set $d_k = \arg\min\{f(x_k + \alpha_k d) : d \in \mathcal{D}_k\}$. (complete polling)

  if $f(x_k + \alpha_k d_k) < f(x_k) - c\alpha_k^2$ **then**

    Set $x_{k+1} = x_k + \alpha_k d_k$ and $\alpha_{k+1} = \gamma\alpha_k$.

    (Move and expand step size)

  else

    Set $x_{k+1} = x_k$ and $\alpha_{k+1} = \theta\alpha_k$.

    (Stay and shrink step size)

Probabilistic Direct Search (PDS): a simplified framework

Algorithm 1: Probabilistic Direct Search based on sufficient decrease

Input: $x_0 \in \mathbb{R}^n$, $\alpha_0 \in (0, \infty)$, $0 < \theta < 1 < \gamma$.

for $k = 0, 1, \ldots$ **do**

- Select a finite set of directions $\mathcal{D}_k \subset \mathbb{R}^n$ randomly.
- (In this talk, assume $\mathcal{D}_k$ is a set of unit vectors for simplicity)
- Set $d_k = \arg\min\{f(x_k + \alpha_k d) : d \in \mathcal{D}_k\}$. (complete polling)
- **if** $f(x_k + \alpha_k d_k) < f(x_k) - c\alpha_k^2$ **then**
  - Set $x_{k+1} = x_k + \alpha_k d_k$ and $\alpha_{k+1} = \gamma\alpha_k$.
  - (Move and expand step size)
- **else**
  - Set $x_{k+1} = x_k$ and $\alpha_{k+1} = \theta\alpha_k$.
  - (Stay and shrink step size)

Typical choice of $\{\mathcal{D}_k\}$ (Gratton, Royer, Vicente, and Zhang 2015):

$$\mathcal{D}_k = \{d_1, \ldots, d_m\} \text{ with } d_\ell \overset{\text{i.i.d.}}{\sim} \mathrm{U}(\mathcal{S}^{n-1})$$

Probabilistic Direct Search (PDS): a simplified framework

Algorithm 1: Probabilistic Direct Search based on sufficient decrease

Input: $x_0 \in \mathbb{R}^n$, $\alpha_0 \in (0, \infty)$, $0 < \theta < 1 < \gamma$.

for $k = 0, 1, \ldots$ **do**

Select a finite set of directions $\mathcal{D}_k \subset \mathbb{R}^n$ randomly.

(In this talk, assume $\mathcal{D}_k$ is a set of unit vectors for simplicity)

Set $d_k = \arg\min\{f(x_k + \alpha_k d) : d \in \mathcal{D}_k\}$. (complete polling)

if $f(x_k + \alpha_k d_k) < f(x_k) - c\alpha_k^2$ **then**

Set $x_{k+1} = x_k + \alpha_k d_k$ and $\alpha_{k+1} = \gamma\alpha_k$.

(Move and expand step size)

else

Set $x_{k+1} = x_k$ and $\alpha_{k+1} = \theta\alpha_k$.

(Stay and shrink step size)

Typical choice of $\{\mathcal{D}_k\}$ (Gratton, Royer, Vicente, and Zhang 2015):

$$\mathcal{D}_k = \{d_1, \ldots, d_m\} \text{ with } d_\ell \overset{\text{i.i.d.}}{\sim} \mathrm{U}(\mathcal{S}^{n-1})$$

N.B.: typical choice in the deterministic case is $\{\pm e_i\}_{i=1}^n$, Coordinate Search (CS)

Illustration of how PDS works

$\mathcal{D}_k = \{d_1, d_2\}$, where $d_\ell \overset{\text{i.i.d.}}{\sim} \mathsf{U}(\mathcal{S}^1)$

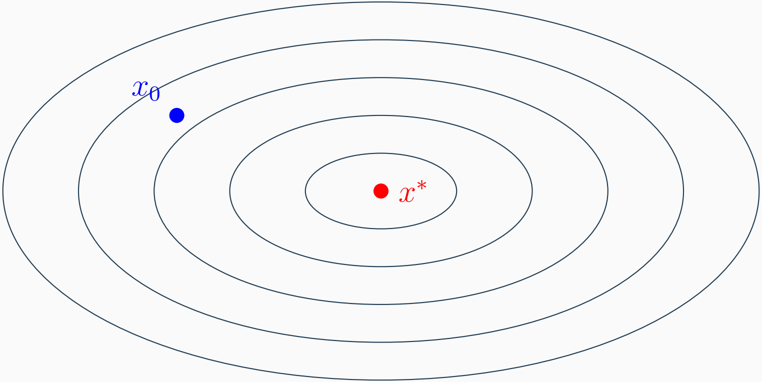


Illustration of how PDS works

$\mathcal{D}_k = \{d_1, d_2\}$, where $d_\ell \overset{\text{i.i.d.}}{\sim} \mathsf{U}(\mathcal{S}^1)$

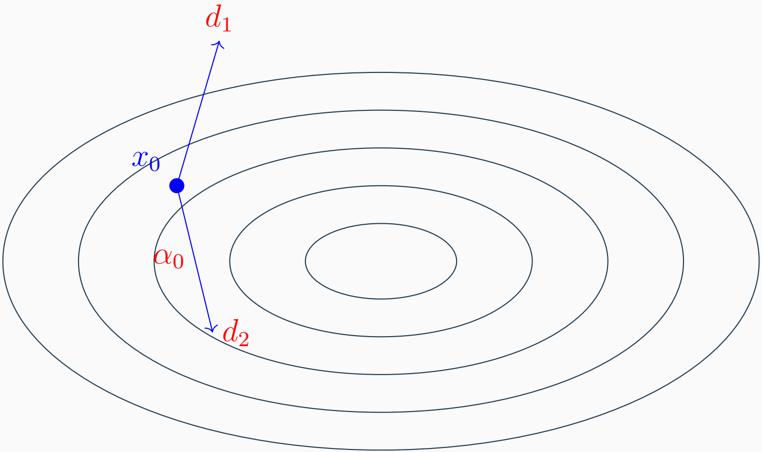


Illustration of how PDS works

$\mathcal{D}_k = \{d_1, d_2\}$, where $d_\ell \overset{\text{i.i.d.}}{\sim} \mathsf{U}(\mathcal{S}^1)$

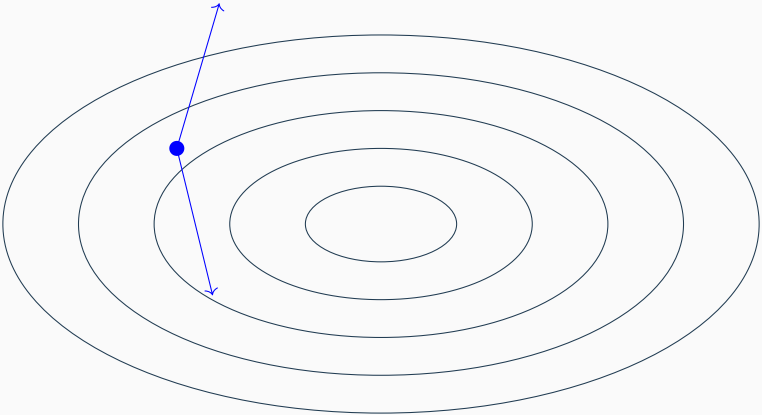


Illustration of how PDS works

$\mathcal{D}_k = \{d_1, d_2\}$, where $d_\ell \overset{\text{i.i.d.}}{\sim} \mathsf{U}(\mathcal{S}^1)$

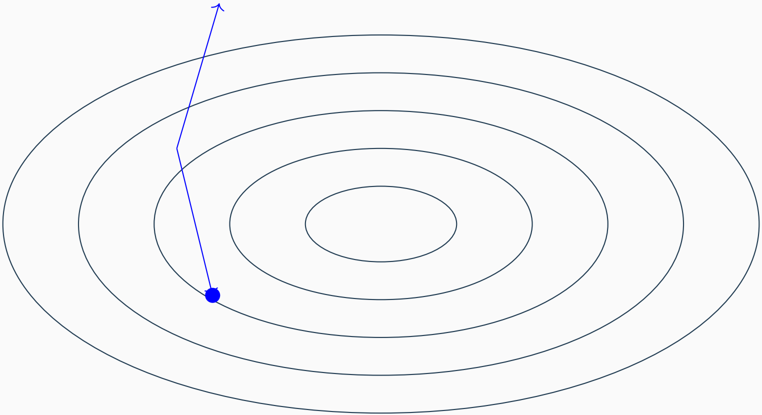


Illustration of how PDS works

$\mathcal{D}_k = \{d_1, d_2\}$, where $d_\ell \overset{\text{i.i.d.}}{\sim} \mathsf{U}(\mathcal{S}^1)$

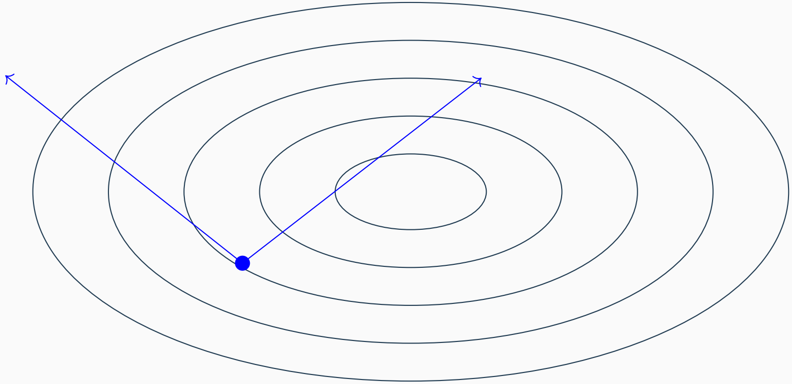


Illustration of how PDS works

$\mathcal{D}_k = \{d_1, d_2\}$, where $d_\ell \overset{\text{i.i.d.}}{\sim} \mathsf{U}(\mathcal{S}^1)$

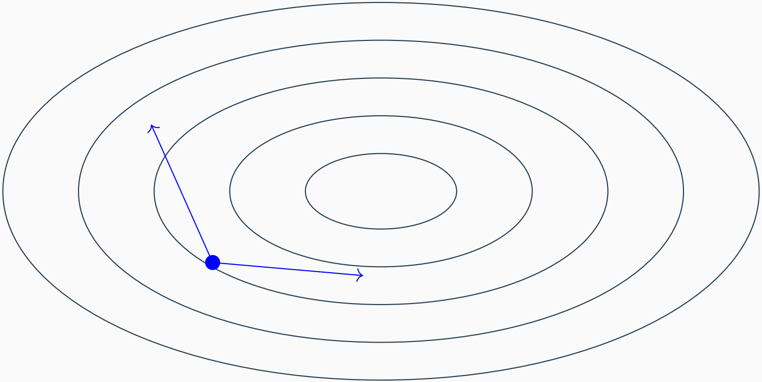


Illustration of how PDS works

$\mathcal{D}_k = \{d_1, d_2\}$, where $d_\ell \overset{\text{i.i.d.}}{\sim} \mathsf{U}(\mathcal{S}^1)$

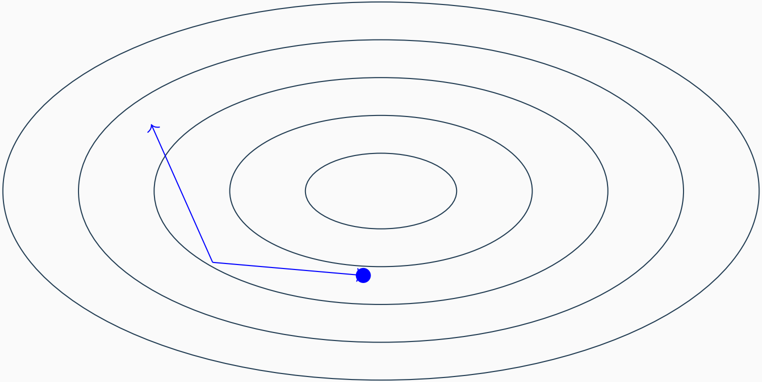


Illustration of how PDS works

$\mathcal{D}_k = \{d_1, d_2\}$, where $d_\ell \overset{\text{i.i.d.}}{\sim} \mathsf{U}(\mathcal{S}^1)$

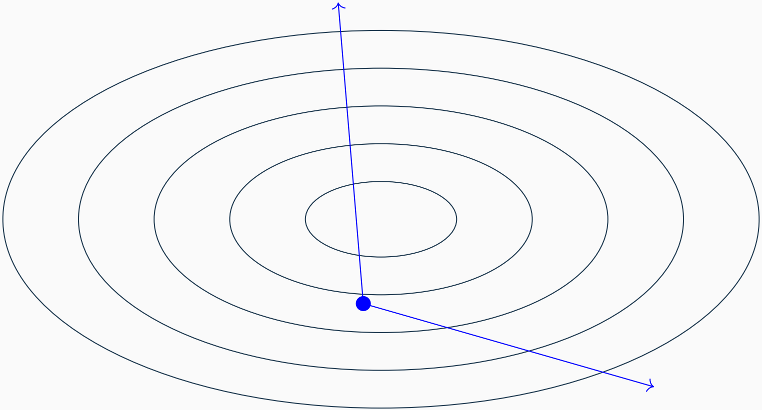


Illustration of how PDS works

$\mathcal{D}_k = \{d_1, d_2\}$, where $d_\ell \overset{\text{i.i.d.}}{\sim} \mathsf{U}(\mathcal{S}^1)$

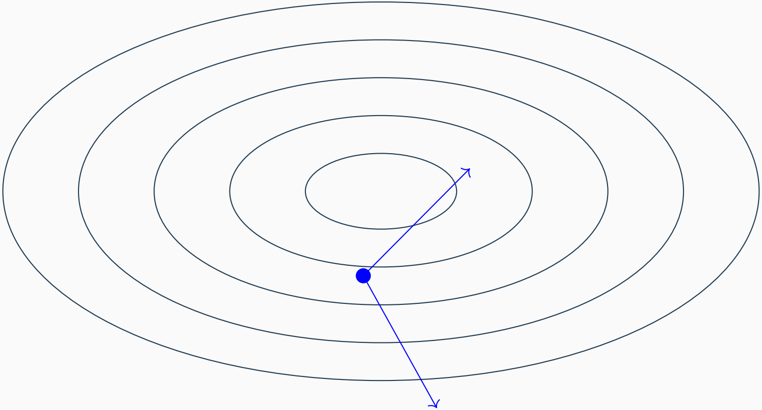


Illustration of how PDS works

$\mathcal{D}_k = \{d_1, d_2\}$, where $d_\ell \overset{\text{i.i.d.}}{\sim} \mathsf{U}(\mathcal{S}^1)$

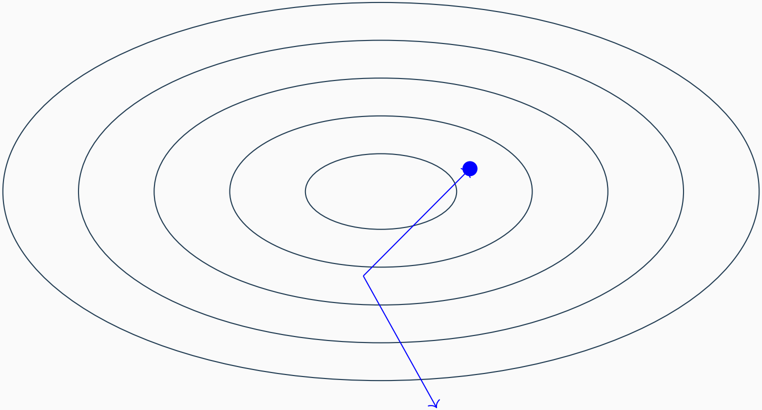


Illustration of how PDS works

$\mathcal{D}_k = \{d_1, d_2\}$, where $d_\ell \overset{\text{i.i.d.}}{\sim} \mathsf{U}(\mathcal{S}^1)$

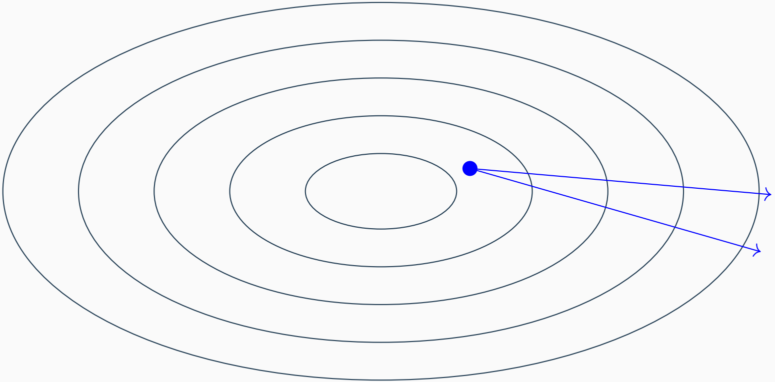


Illustration of how PDS works

$\mathcal{D}_k = \{d_1, d_2\}$, where $d_\ell \overset{\text{i.i.d.}}{\sim} \mathsf{U}(\mathcal{S}^1)$

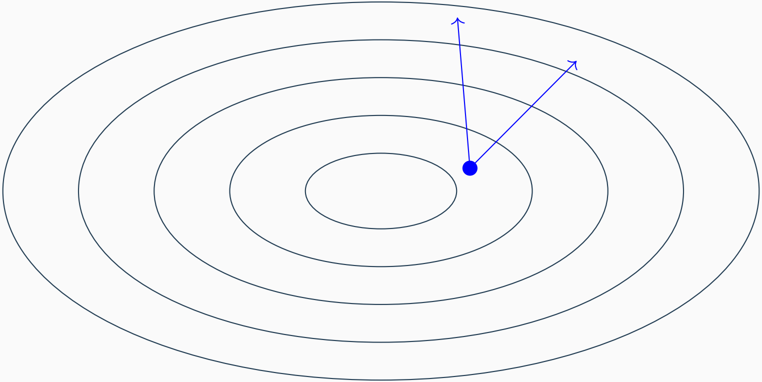
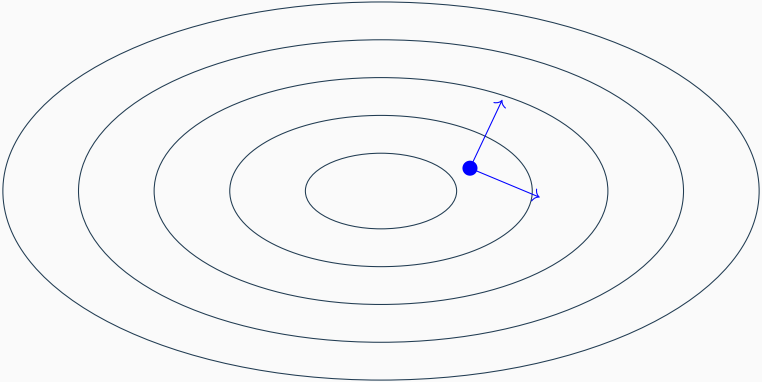


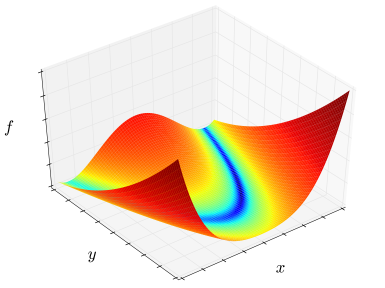
Illustration of how PDS works

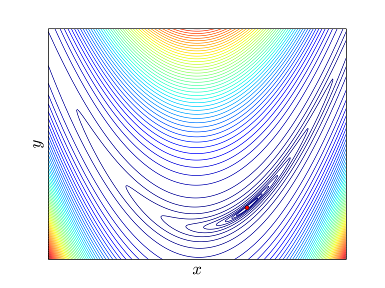
$\mathcal{D}_k = \{d_1, d_2\}$, where $d_\ell \overset{\text{i.i.d.}}{\sim} \mathsf{U}(\mathcal{S}^1)$



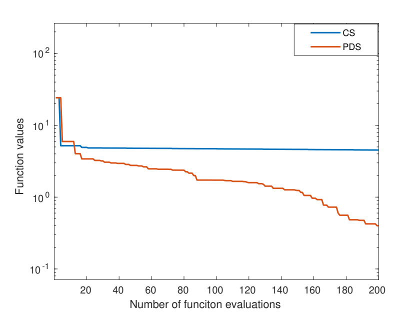
Rosenbrock "banana" function:

$$f(x) = \sum_{i=1}^{n-1} \left[(1 - x_i)^2 + 100(x_{i+1} - x_i^2)^2 \right]$$

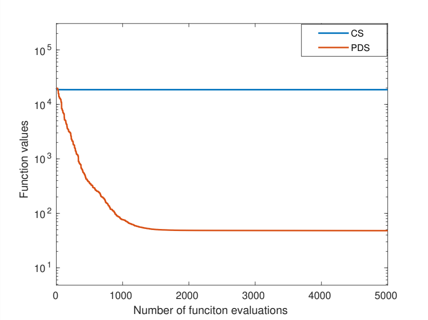




A numerical example: CS v.s. PDS with 2 directions



$n = 2$



$n = 50$

Function value v.s. number of function evaluations

Worst case complexity of function evaluations (GRVZ 2015)

$\mathcal{O}(n^2\epsilon^{-2})$ for CS while $\mathcal{O}(n\epsilon^{-2})$ for PDS

Cosine measure

Definition (Cosine measure w.r.t. a vector)

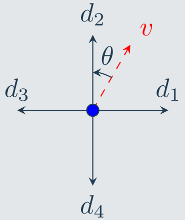
Given a finite set $\mathcal{D} \subseteq \mathbb{R}^n \backslash \{0\}$ and a vector $v \in \mathbb{R}^n \backslash \{0\}$, define

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

which is the cosine measure of $\mathcal{D}$ with respect to v.

Example

$$\mathrm{cm}(\mathcal{D}, v) \;=\; \cos\theta$$



Cosine measure

Definition (Cosine measure w.r.t. a vector)

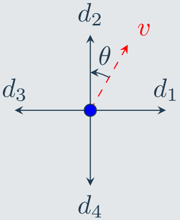
Given a finite set $\mathcal{D} \subseteq \mathbb{R}^n \backslash \{0\}$ and a vector $v \in \mathbb{R}^n \backslash \{0\}$, define

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

which is the cosine measure of $\mathcal{D}$ with respect to v.

Example

$$\mathrm{cm}(\mathcal{D}, v) \;=\; \cos\theta$$



$\mathrm{cm}(\mathcal{D}, v)$ measures the ability of $\mathcal{D}$ to “approximate” v

Definition (p-probabilistically κ-descent)

$\{\mathcal{D}_k\}$ is said to be p-probabilistically κ-descent, if

$$\mathbb{P}\left(\mathrm{cm}(\mathcal{D}_k, -g_k) \geq \kappa \mid \mathcal{D}_0, \ldots, \mathcal{D}_{k-1}\right) \;\geq\; p \quad \text{for each } k \geq 0,$$

where $g_k = \nabla f(x_k)$.

Intuition

Each $\mathcal{D}_k$ is “good enough” with probability at least p
no matter what has happened in the history

Convergence theory

Definition (p-probabilistically κ-descent)

$\{\mathcal{D}_k\}$ is said to be p-probabilistically κ-descent, if

$$\mathbb{P}\left(\mathrm{cm}(\mathcal{D}_k, -g_k) \geq \kappa \mid \mathcal{D}_0, \ldots, \mathcal{D}_{k-1}\right) \;\geq\; p \quad \text{for each } k \geq 0,$$

where $g_k = \nabla f(x_k)$.

Intuition

Each $\mathcal{D}_k$ is "good enough" with probability at least p
no matter what has happened in the history

Theorem (GRVZ, 2015)

If $\{\mathcal{D}_k\}$ is p_0-probabilistically κ-descent with $\kappa > 0$ and

$$p_0 \;=\; \frac{\log \theta}{\log(\gamma^{-1}\theta)},$$

then PDS converges w.p.1 when f is L-smooth and lower-bounded.

Practical choice and natural question

Corollary (GRVZ, 2015)

If $\mathcal{D}_k = \{d_1, \ldots, d_m\}$*, where* $d_\ell \overset{\text{i.i.d.}}{\sim} U(\mathcal{S}^{n-1})$*, then PDS converges w.p.1 if*

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

Practical choice and natural question

Corollary (GRVZ, 2015)

If $\mathcal{D}_k = \{d_1, \ldots, d_m\}$, *where* $d_\ell \overset{\text{i.i.d.}}{\sim} U(\mathcal{S}^{n-1})$, *then PDS converges w.p.1 if*

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

A natural question: what if

$$m \leq \log_2\left(1 - \frac{\log\theta}{\log\gamma}\right)?$$

Practical choice and natural question

Corollary (GRVZ, 2015)

If $\mathcal{D}_k = \{d_1, \ldots, d_m\}$, where $d_\ell \overset{\text{i.i.d.}}{\sim} U(\mathcal{S}^{n-1})$, then PDS converges w.p.1 if

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

A natural question: what if

$$m \leq \log_2\left(1 - \frac{\log\theta}{\log\gamma}\right)?$$

Moreover, are supermartingale-like assumptions essential?

$$\mathbb{P}(\text{some event} \mid \mathcal{F}) \;\geq\; p$$

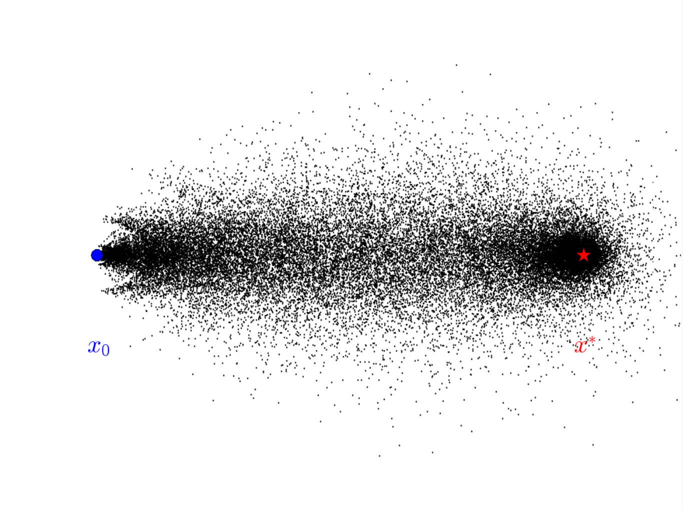
Related talks: Coralia, Kwassi Joseph, Matt, Anne, Warren, Sara, Lindon

A simple test

- Objective function: $f(x) = \|x\|^2/2$
- Initial point: $x_0 = (-10, 0)^\mathsf{T}$
- Stopping criterion: $\alpha_k \le$ machine epsilon
- Number of experiments: $100,000$
- Parameters of PDS: $\alpha_0 = 1$, $\theta = 0.25$, $\gamma = 1.5$, $m = 2$

$$m = 2 < 2.143 \approx \log_2 \left(1 - \frac{\log \theta}{\log \gamma}\right)$$

A simple test (Cont'd)



Note: each **black dot** represents the **output point** of one run of PDS.

Non-convergence study is not rare

Many well-known algorithms have non-convergence examples

- Powell, On search directions for minimization algorithms, 1973.
- Yuan, An example of non-convergence of trust region algorithms, 1998.
- Reddi, Kale, and Kumar, On the convergence of Adam and beyond, 2018.
- Chen, He, Ye, and Yuan, The direct extension of ADMM for multi-block convex minimization problems is not necessarily convergent, 2016.
- Dai, A perfect example for the BFGS method, 2013.
- Mascarenhas, The divergence of the BFGS and Gauss Newton methods, 2014.

Non-convergence study is not rare

Many well-known algorithms have non-convergence examples

- Powell, On search directions for minimization algorithms, 1973.
- Yuan, An example of non-convergence of trust region algorithms, 1998.
- Reddi, Kale, and Kumar, On the convergence of Adam and beyond, 2018.
- Chen, He, Ye, and Yuan, The direct extension of ADMM for multi-block convex minimization problems is not necessarily convergent, 2016.
- Dai, A perfect example for the BFGS method, 2013.
- Mascarenhas, The divergence of the BFGS and Gauss Newton methods, 2014.

Instead of finding a non-convergence example,
can we develop a theorem?

An overview of our theory

We assume that f is smooth and convex (explained later).

We denote the optimal solution set of f by $\mathcal{S}^*$.

We will establish the following.

Under some assumption on $\{\mathcal{D}_k\}$ and algorithmic parameters, there exist choices of x_0 such that

$$\mathbb{P}\left(\liminf_{k\to\infty} \operatorname{dist}(x_k, \mathcal{S}^*) > 0\right) > 0.$$

Differences from a non-convergence example

one function	v.s.	some function class
special parameters	v.s.	conditions for parameters
a specific initial point	v.s.	a region for initial points

Assumption on $\{\mathcal{D}_k\}$: probabilistic ascent

Recall p-probabilistically κ-descent

$$\mathbb{P}\left(\mathrm{cm}\left(\mathcal{D}_k, -g_k\right) \geq \kappa \mid \mathcal{D}_0, \ldots, \mathcal{D}_{k-1}\right) \geq p \quad \text{for each } k \geq 0.$$

Assumption on $\{\mathcal{D}_k\}$: probabilistic ascent

Recall p-probabilistically κ-descent

$$\mathbb{P}\left(\mathrm{cm}\left(\mathcal{D}_k, -g_k\right) \geq \kappa \mid \mathcal{D}_0, \ldots, \mathcal{D}_{k-1}\right) \geq p \quad \text{for each } k \geq 0.$$

q-probabilistically ascent

$$\mathbb{P}\left(\mathrm{cm}\left(\mathcal{D}_k, -g_k\right) \leq 0 \mid \mathcal{D}_0, \ldots, \mathcal{D}_{k-1}\right) \geq q \quad \text{for each } k \geq 0.$$

Assumption on $\{\mathcal{D}_k\}$: probabilistic ascent

Recall p-probabilistically κ-descent

$$\mathbb{P}\left(\mathrm{cm}\left(\mathcal{D}_k, -g_k\right) \geq \kappa \mid \mathcal{D}_0, \ldots, \mathcal{D}_{k-1}\right) \geq p \quad \text{for each } k \geq 0.$$

q-probabilistically ascent

$$\mathbb{P}\left(\mathrm{cm}\left(\mathcal{D}_k, -g_k\right) \leq 0 \mid \mathcal{D}_0, \ldots, \mathcal{D}_{k-1}\right) \geq q \quad \text{for each } k \geq 0.$$

Note

If $\mathrm{cm}\left(\mathcal{D}_k, -g_k\right) \leq 0$, then $\mathcal{D}_k$ is "bad" (no descent direction).

Why assuming convexity?

$$
\begin{tikzcd}
\operatorname{cm}(\mathcal{D}_k, -g_k) \leq 0 \arrow[r] & \text{No descent direction} \arrow[d, "f \text{ convex}"] \\
\alpha_k \text{ shrinks} & f(x_k + \alpha_k d) \geq f(x_k)\ \forall d \in \mathcal{D}_k \arrow[l]
\end{tikzcd}
$$

Why assuming convexity?

$$
\begin{tikzcd}
\mathrm{cm}\left(\mathcal{D}_k, -g_k\right) \leq 0 \arrow[r] & \text{No descent direction} \arrow[d, "f \text{ convex}"] \\
\alpha_k \text{ shrinks} & f(x_k + \alpha_k d) \geq f(x_k)\ \forall d \in \mathcal{D}_k \arrow[l]
\end{tikzcd}
$$

- Convexity connects $\mathrm{cm}\left(\mathcal{D}_k, -g_k\right) \leq 0$ and shrinking of step size

Why assuming convexity?

$$\begin{tikzcd}
\mathrm{cm}\left(\mathcal{D}_k, -g_k\right) \leq 0 \arrow[r] & \text{No descent direction} \arrow[d, "f \text{ convex}"] \\
\alpha_k \text{ shrinks} & f(x_k + \alpha_k d) \geq f(x_k)\ \forall d \in \mathcal{D}_k \arrow[l]
\end{tikzcd}$$

- Convexity connects $\mathrm{cm}\left(\mathcal{D}_k, -g_k\right) \leq 0$ and shrinking of step size
- $\{\mathcal{D}_k\}$ is probabilistic ascent implies α_k “often” shrinks

From probabilistic ascent to non-convergence: How?

$\{\mathcal{D}_k\}$ is probabilistically ascent

From probabilistic ascent to non-convergence: How?

$\{\mathcal{D}_k\}$ is probabilistically ascent

$\Downarrow$

α_k "often" shrinks

From probabilistic ascent to non-convergence: How?

$\{\mathcal{D}_k\}$ is probabilistically ascent

$\Downarrow$

α_k "often" shrinks

$\Downarrow$

$$\mathbb{P}\left(\sum_{k=0}^{\infty} \alpha_k \text{ is "bounded"}\right) > 0 \text{ ?}$$

From probabilistic ascent to non-convergence: How?

$\{\mathcal{D}_k\}$ is probabilistically ascent

$\Downarrow$

α_k "often" shrinks

$\Downarrow$

$$\mathbb{P}\left(\sum_{k=0}^{\infty} \alpha_k \text{ is "bounded"}\right) > 0 \text{ ?}$$

$\Downarrow$

$\mathbb{P}\,(\text{non-convergence}) > 0$ if $\operatorname{dist}(x_0, \mathcal{S}^*)$ is "large"?

Key ingredients of the analysis

- Define the indicator function for "bad $\mathcal{D}_k$"

$$Y_k = \mathbb{1}(\mathrm{cm}\left(\mathcal{D}_k, -g_k\right) \leq 0)$$

Key ingredients of the analysis

- Define the indicator function for "bad $\mathcal{D}_k$"

$$Y_k = \mathbb{1}(\mathrm{cm}\,(\mathcal{D}_k, -g_k) \le 0)$$

- Note the following inequality between step sizes (f is convex)

$$\alpha_{k+1} \le \begin{cases} \gamma\alpha_k, & \text{if } Y_k = 0 \\ \theta\alpha_k, & \text{if } Y_k = 1 \end{cases} = \gamma^{1-Y_k}\theta^{Y_k}\alpha_k$$

Key ingredients of the analysis

- Define the indicator function for "bad $\mathcal{D}_k$"

$$Y_k = \mathbb{1}(\mathrm{cm}\,(\mathcal{D}_k, -g_k) \leq 0)$$

- Note the following inequality between step sizes (f is convex)

$$\alpha_{k+1} \;\leq\; \begin{cases} \gamma\alpha_k, & \text{if } Y_k = 0 \\ \theta\alpha_k, & \text{if } Y_k = 1 \end{cases} \;=\; \gamma^{1-Y_k}\theta^{Y_k}\alpha_k$$

- Use the above inequality iteratively

$$\alpha_k \leq \alpha_0 \prod_{\ell=0}^{k-1} \gamma^{1-Y_\ell}\theta^{Y_\ell}$$

Key ingredients of the analysis

- Define the indicator function for "bad $\mathcal{D}_k$"

$$Y_k = \mathbb{1}(\mathrm{cm}\left(\mathcal{D}_k, -g_k\right) \leq 0)$$

- Note the following inequality between step sizes (f is convex)

$$\alpha_{k+1} \ \leq \ \begin{cases} \gamma\alpha_k, & \text{if } Y_k = 0 \\ \theta\alpha_k, & \text{if } Y_k = 1 \end{cases} \ = \ \gamma^{1-Y_k}\theta^{Y_k}\alpha_k$$

- Use the above inequality iteratively

$$\alpha_k \leq \alpha_0 \prod_{\ell=0}^{k-1} \gamma^{1-Y_\ell}\theta^{Y_\ell}$$

- Get an upper bound of series of step sizes

$$\sum_{k=1}^{\infty} \alpha_k \leq \alpha_0 \sum_{k=1}^{\infty} \prod_{\ell=0}^{k-1} \gamma^{1-Y_\ell}\theta^{Y_\ell} =: \alpha_0 S$$

- Analyze the behavior of the random series S

A closer look at the random series S

Recall that

$$S = \sum_{k=1}^{\infty} \prod_{\ell=0}^{k-1} \gamma^{1-Y_\ell} \theta^{Y_\ell},$$

where $Y_\ell = \mathbb{1}(\mathrm{cm}\,(\mathcal{D}_\ell, -g_\ell) \leq 0)$.

A closer look at the random series S

Recall that

$$S = \sum_{k=1}^{\infty} \prod_{\ell=0}^{k-1} \gamma^{1-Y_\ell} \theta^{Y_\ell},$$

where $Y_\ell = \mathbb{1}(\mathrm{cm}\,(\mathcal{D}_\ell, -g_\ell) \leq 0)$.

Two questions

- (Q1) Does there exist a constant ζ such that

$$\mathbb{P}\,(S < \zeta) > 0?$$

- (Q2) Moreover, can we specify the value of ζ?

Answer to Q1 and Q2

Proposition

If $\{\mathcal{D}_k\}$ is q-probabilistically ascent with $q > q_0$, where

$$q_0 \;=\; 1 - p_0 \;=\; \frac{\log \gamma}{\log(\theta^{-1}\gamma)},$$

then

1. $$\mathbb{P}\left(S < \infty\right) \;=\; 1,$$

2. $$\mathbb{P}\left(S < \zeta\right) \;>\; 0 \quad\Longleftrightarrow\quad \zeta > \frac{\theta}{1-\theta}.$$

Answer to Q1 and Q2

Proposition

If $\{\mathcal{D}_k\}$ is q-probabilistically ascent with $q > q_0$, where

$$q_0 \; = \; 1 - p_0 \; = \; \frac{\log \gamma}{\log(\theta^{-1}\gamma)},$$

then

1. $$\mathbb{P}\,(S < \infty) \; = \; 1,$$

2. $$\mathbb{P}\,(S < \zeta) \; > \; 0 \quad \iff \quad \zeta > \frac{\theta}{1-\theta}.$$

Note

- $\mathbb{P}(S < \infty) = 1$ implies the existence of a ζ but not its value.
- The lower bound in 2 is tight, as $S = \sum_{k=1}^{\infty} \prod_{\ell=0}^{k-1} \gamma^{1-Y_\ell} \theta^{Y_\ell} \geq \; \frac{\theta}{1-\theta}$.

Non-convergence of PDS

Theorem

Under aforementioned assumptions on f, if the sequence $\{\mathcal{D}_k\}$ in PDS is q-probabilistically ascent with $q > q_0$, then

$$\mathbb{P}\left(\liminf_{k\to\infty} \operatorname{dist}(x_k, \mathcal{S}^*) > 0\right) \; > \; 0,$$

provided that $\operatorname{dist}(x_0, \mathcal{S}^*) > \alpha_0/(1-\theta)$.

Weaker assumption than probabilistic ascent

Denote $\mathbb{P}(\mathrm{cm}(\mathcal{D}_k, -g_k) \le 0 \mid \mathcal{D}_0, \ldots, \mathcal{D}_{k-1})$ by P_k.

Recall that $\{\mathcal{D}_k\}$ is q-probabilistically ascent if $P_k \ge q$ for each $k \ge 0$.

Note that $\{P_k\}$ are random variables.

Weaker assumption than probabilistic ascent

Denote $\mathbb{P}(\mathrm{cm}(\mathcal{D}_k, -g_k) \le 0 \mid \mathcal{D}_0, \ldots, \mathcal{D}_{k-1})$ by P_k.

Recall that $\{\mathcal{D}_k\}$ is q-probabilistically ascent if $P_k \ge q$ for each $k \ge 0$.

Note that $\{P_k\}$ are random variables.

What we need is

$$\text{not} \quad \mathbb{P}(S < \infty) = 1 \qquad \text{but} \quad \mathbb{P}(S < \infty) > 0.$$

Weaker assumption than probabilistic ascent

Denote $\mathbb{P}(\mathrm{cm}(\mathcal{D}_k, -g_k) \leq 0 \mid \mathcal{D}_0, \ldots, \mathcal{D}_{k-1})$ by P_k.

Recall that $\{\mathcal{D}_k\}$ is q-probabilistically ascent if $P_k \geq q$ for each $k \geq 0$.

Note that $\{P_k\}$ are random variables.

What we need is

$$\text{not} \quad \mathbb{P}(S < \infty) = 1 \qquad \text{but} \quad \mathbb{P}(S < \infty) > 0.$$

For the latter, we can relax the assumption

$$\text{from} \quad P_k \geq q > q_0 \qquad \text{to} \quad \mathbb{P}\left(\liminf_{k \to \infty} P_k > q_0\right) > 0.$$

What happens in the typical implementation of PDS?

Let $\mathcal{D}_k = \{d_1, \ldots, d_m\}$, where $d_\ell \overset{\text{i.i.d.}}{\sim} \mathsf{U}(\mathcal{S}^{n-1})$.

Recall that PDS is convergent if

$$m \; > \; \log_2\left(1 - \frac{\log\theta}{\log\gamma}\right).$$

What happens in the typical implementation of PDS?

Let $\mathcal{D}_k = \{d_1, \ldots, d_m\}$, where $d_\ell \overset{\text{i.i.d.}}{\sim} \mathsf{U}(\mathcal{S}^{n-1})$.

Recall that PDS is convergent if

$$m \; > \; \log_2\left(1 - \frac{\log\theta}{\log\gamma}\right).$$

With our non-convergence analysis, PDS is non-convergent if

$$\mathbb{P}\left(\mathrm{cm}\left(\mathcal{D}_k, -g_k\right) \le 0 \mid \mathcal{D}_0, \ldots, \mathcal{D}_{k-1}\right) \; > \; q_0,$$

What happens in the typical implementation of PDS?

Let $\mathcal{D}_k = \{d_1, \ldots, d_m\}$, where $d_\ell \overset{\text{i.i.d.}}{\sim} \mathsf{U}(\mathcal{S}^{n-1})$.

Recall that PDS is convergent if

$$m \; > \; \log_2\left(1 - \frac{\log\theta}{\log\gamma}\right).$$

With our non-convergence analysis, PDS is non-convergent if

$$\mathbb{P}\left(\mathrm{cm}\left(\mathcal{D}_k, -g_k\right) \leq 0 \mid \mathcal{D}_0, \ldots, \mathcal{D}_{k-1}\right) \; > \; q_0,$$

which is equivalent to

$$\left(\frac{1}{2}\right)^m \; > \; \frac{\log\gamma}{\log(\theta^{-1}\gamma)},$$

What happens in the typical implementation of PDS?

Let $\mathcal{D}_k = \{d_1, \ldots, d_m\}$, where $d_\ell \overset{\text{i.i.d.}}{\sim} \mathsf{U}(\mathcal{S}^{n-1})$.

Recall that PDS is convergent if

$$m \; > \; \log_2\left(1 - \frac{\log\theta}{\log\gamma}\right).$$

With our non-convergence analysis, PDS is non-convergent if

$$\mathbb{P}\left(\mathrm{cm}\left(\mathcal{D}_k, -g_k\right) \le 0 \mid \mathcal{D}_0, \ldots, \mathcal{D}_{k-1}\right) \; > \; q_0,$$

which is equivalent to

$$\left(\frac{1}{2}\right)^m \; > \; \frac{\log\gamma}{\log(\theta^{-1}\gamma)},$$

or, equivalently,

$$m \; < \; \log_2\left(1 - \frac{\log\theta}{\log\gamma}\right).$$

Assumptions for convergence and non-convergence are essential.

Tightness of our assumption on $\{\mathcal{D}_k\}$

Our assumption on $\{\mathcal{D}_k\}$:

q-probabilistically ascent with $q > q_0$.

Natural question:

Is it sufficient to require $q \geq q_0$?

Tightness of our assumption on $\{\mathcal{D}_k\}$

Our assumption on $\{\mathcal{D}_k\}$:

q-probabilistically ascent with $q > q_0$.

Natural question:

Is it sufficient to require $q \geq q_0$?

Answer: NO!

Tightness of our assumption on $\{\mathcal{D}_k\}$

Our assumption on $\{\mathcal{D}_k\}$:

q-probabilistically ascent with $q > q_0$.

Natural question:

Is it sufficient to require $q \geq q_0$?

Answer: NO!

Example

We assume

- $\theta = 1/2$ and $\gamma = 2$, which implies $q_0 = 1/2$;
- $\mathcal{D}_k = \{g_k/\|g_k\|\}$ or $\{-g_k/\|g_k\|\}$ with probability $1/2$, respectively.

Then PDS converges w.p.1.

Convergence result inspired by non-convergence analysis

Define a series

$$S(\kappa) = \sum_{k=1}^{\infty} \prod_{\ell=0}^{k-1} \gamma^{Z_\ell(\kappa)} \theta^{1-Z_\ell(\kappa)},$$

where $Z_\ell(\kappa) = \mathbb{1}(\mathrm{cm}\,(\mathcal{D}_\ell, -g_\ell) \geq \kappa)$.

Convergence result inspired by non-convergence analysis

Define a series

$$S(\kappa) \ = \ \sum_{k=1}^{\infty} \prod_{\ell=0}^{k-1} \gamma^{Z_\ell(\kappa)} \theta^{1-Z_\ell(\kappa)},$$

where $Z_\ell(\kappa) = \mathbb{1}(\mathrm{cm}\,(\mathcal{D}_\ell, -g_\ell) \geq \kappa)$.

Roughly speaking, $\mathbb{P}(S(0) < \infty) > 0$ implies non-convergence of PDS.

Convergence result inspired by non-convergence analysis

Define a series

$$S(\kappa) \;=\; \sum_{k=1}^{\infty} \prod_{\ell=0}^{k-1} \gamma^{Z_\ell(\kappa)} \theta^{1-Z_\ell(\kappa)},$$

where $Z_\ell(\kappa) = \mathbb{1}(\mathrm{cm}\,(\mathcal{D}_\ell, -g_\ell) \geq \kappa)$.

Roughly speaking, $\mathbb{P}(S(0) < \infty) > 0$ implies non-convergence of PDS.

Theorem

If there exists a $\kappa > 0$ such that $S(\kappa) = \infty$, then DS converges.

Convergence result inspired by non-convergence analysis

Define a series

$$S(\kappa) \; = \sum_{k=1}^{\infty} \prod_{\ell=0}^{k-1} \gamma^{Z_\ell(\kappa)} \theta^{1-Z_\ell(\kappa)},$$

where $Z_\ell(\kappa) = \mathbb{1}(\mathrm{cm}\,(\mathcal{D}_\ell, -g_\ell) \geq \kappa)$.

Roughly speaking, $\mathbb{P}(S(0) < \infty) > 0$ implies non-convergence of PDS.

Theorem

If there exists a $\kappa > 0$ such that $S(\kappa) = \infty$, then DS converges.

Relation with existing result (GRVZ, 2015)

$$p_0\text{-probabilistically } \kappa\text{-descent} \quad \Longrightarrow \quad S(\kappa) = \infty \text{ w.p.1}$$

Take away

In this talk, we

- theoretically explain the non-convergence phenomenon of PDS,
- find out the behavior of PDS is closely related to the random series

$$S = \sum_{k=1}^{\infty} \prod_{\ell=0}^{k-1} \gamma^{1-Y_\ell} \theta^{Y_\ell}.$$

Non-convergence analysis can

- verify whether your assumption for convergence is essential,
- deepen our understanding of mathematical tools we use,
- provide new perspectives on convergence analysis,
- guide the choice of algorithmic parameters,

One more thing: OptiProfiler

OptiProfiler (joint work with Tom M. Ragonneau and Zaikun Zhang) is

a benchmarking platform for DFO solvers.

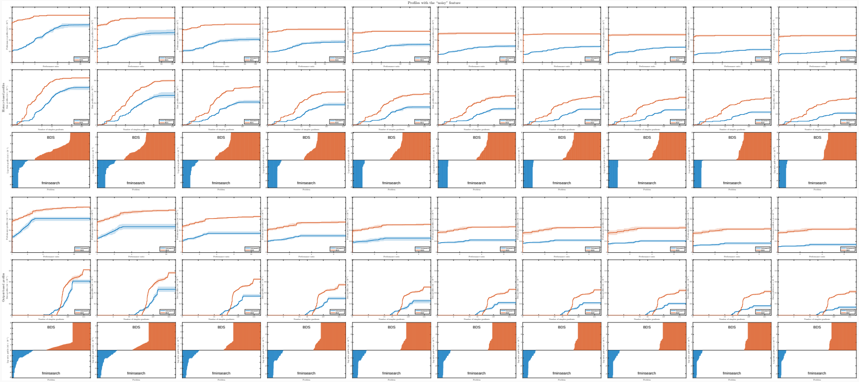
Our goal: fair, convenient, and uniform benchmarking.

- Creating performance profiles, data profiles, and log-ratio profiles.
  [Moré, Wild, 2009; Shi, Xuan, Oztoprak, and Nocedal, 2023]
  Thanks for Nikolaus's nice talk: runtime distributions and COCO!
- Providing multiple types of tests
  noisy function, unrelaxable constraints, randomized initial point...
- Implemented in Python and MATLAB

One more thing: OptiProfiler

Just one line MATLAB code:

```
benchmark({@bds, @fminsearch}, "noisy")
```



GitHub repository: https://github.com/optiprofiler

Acknowledgement

- Thanks to the organizers!
- Thanks to all the speakers!
- Thanks Giampaolo and Geovani for saving my life!
- Thanks Zaikun for giving me this great opportunity!

Grazie mille!

Photo taken by Lindon.

References I

- Chen, C. et al. (2016). "The direct extension of ADMM for multi-block convex minimization problems is not necessarily convergent". *Math. Program.* 155, pp. 57–79.
- Conn, A. R., Scheinberg, K., and Vicente, L. N. (2009). *Introduction to Derivative-Free Optimization*. Vol. 8. MOS-SIAM Ser. Optim. Philadelphia: SIAM.
- Durrett, R. (2010). *Probability: Theory and Examples*. Fourth. Camb. Ser. Stat. Probab. Math. Cambridge: Cambridge University Press.
- Fermi, E. and Metropolis, N. (1952). *Numerical solution of a minimum problem*. Tech. rep. Alamos National Laboratory, Los Alamos, USA.

References II

- Ghanbari, H. and Scheinberg, K. (2017). “Black-box optimization in machine learning with trust region based derivative free algorithm”. *arXiv:1703.06925*.
- Gratton, S. et al. (2015). “Direct search based on probabilistic descent”. *SIAM J. Optim.* 25, pp. 1515–1541.
- Kolda, T. G., Lewis, R. M., and Torczon, V. (2003). “Optimization by direct search: New perspectives on some classical and modern methods”. *SIAM Rev.* 45, pp. 385–482.
- Larson, J., Menickelly, M., and Wild, S. M. (2019). “Derivative-free optimization methods”. *Acta Numer.* 28, pp. 287–404.
- Mascarenhas, W. (2014). “The divergence of the BFGS and Gauss Newton methods”. *Math. Program.* 147, pp. 253–276.

References III

- Powell, M. J. D. (1973). “On search directions for minimization algorithms”. *Math. Program.* 4, pp. 193–201.
- Yuan, Y. (1998). “An example of non-convergence of trust region algorithms”. In: *Advances in Nonlinear Programming*. Ed. by Y. Yuan. Dordrecht: Kluwer Academic Publishers, pp. 205–215.