# Non-convergence Analysis of Probabilistic Direct Search

The 25th International Symposium on Mathematical Programming

---

Cunxin Huang

Joint work with Zaikun Zhang

Montréal, Canada July 26, 2024

The Hong Kong Polytechnic University

# Non-convergence Analysis of Probabilistic Direct Search

The 25th International Symposium on Mathematical Programming

---

Zaikun Zhang (replacing Cunxin Huang)

Joint work with Cunxin Huang

Montréal, Canada July 26, 2024

The Hong Kong Polytechnic University

# PRIMA and my gratitude

libprima.net

PRIMA is an acronym for

"Reference Implementation for Powell's Methods
with Modernization and Amelioration".

- Number of lines: > 100,000.
- The total time I spent on PRIMA:

$$\geq\ 3 \text{ years} \times 300 \text{ days per year} \times 10 \text{ hours per day} \ = \ 9,000 \text{ hours}.$$

In the past years, due to the gap in my publication record while working on PRIMA, I needed a lot of support from the community. **Thank you** for the help and support, explicit or implicit, known or unknown to me. Without your support, I would not have survived.

# Non-convergence analysis: What?

Consider an algorithm

$$\mathscr{A} \ : \ \Xi \times \mathbb{F} \times \mathcal{X} \to \mathcal{X}^{\infty}, \quad (\xi,\ f,\ x_0) \mapsto \{x_k\}.$$

- $\xi$ represents algorithmic parameters.
- $f$ is the objective function.
- $x_0$ is the starting point.

When $\mathscr{A}$ is deterministic:

- (Global) Convergence analysis: For all $(\xi, f, x_0) \in \hat{\Xi} \times \hat{\mathbb{F}} \times \mathcal{X}$, prove

$$\{x_k\} \text{ achieves stationarity asymptotically.}$$

- Non-convergence analysis: For all $(\xi, f, x_0) \in \tilde{\Xi} \times \tilde{\mathbb{F}} \times \tilde{\mathcal{X}}$, prove

$$\{x_k\} \text{ fails to achieve stationarity asymptotically.}$$

# Non-convergence analysis: What?

Consider an algorithm

$$\mathscr{A} \;:\; \Xi \times \mathbb{F} \times \mathcal{X} \to \mathcal{X}^{\infty}, \quad (\xi,\ f,\ x_0) \mapsto \{x_k\}.$$

- $\xi$ represents algorithmic parameters.
- $f$ is the objective function.
- $x_0$ is the starting point.

When $\mathscr{A}$ is random:

- (Global) Convergence analysis: For all $(\xi, f, x_0) \in \hat{\Xi} \times \hat{\mathbb{F}} \times \mathcal{X}$, prove

$$\mathbb{P}\left(\{x_k\} \text{ achieves stationarity asymptotically}\right) \;=\; 1.$$

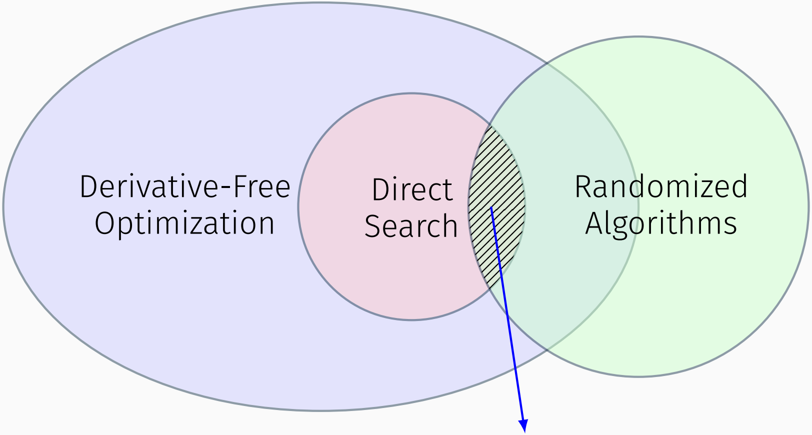
- Non-convergence analysis: For all $(\xi, f, x_0) \in \tilde{\Xi} \times \tilde{\mathbb{F}} \times \tilde{\mathcal{X}}$, prove

$$\mathbb{P}\left(\{x_k\} \text{ fails to achieve stationarity asymptotically}\right) \;>\; 0.$$

# Non-convergence analysis: Why?

- Sharpen our knowledge about the algorithm.
- Deepen our understanding about the convergence analysis.
- Guide the selection of algorithmic parameters.
- Provide new perspectives on convergence analysis.

# Probabilistic Direct Search (PDS)




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

# Derivative-Free Optimization (DFO)

## Derivative-Free Optimization

- Do not use derivatives (first-order info.), only use function values
- Closely related: zeroth-order/black-box optimization ...

Derivatives are often not available in applications

Quantum Computing

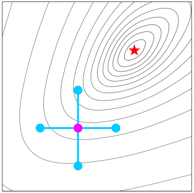

Machine Learning

Circuit Design

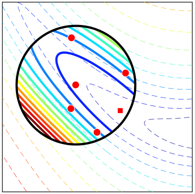
# Direct-search methods and model-based methods

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

# PDS: a simplified framework

**Algorithm 1:** Direct Search based on sufficient decrease

## PDS: a simplified framework

**Algorithm 1:** Direct Search based on sufficient decrease

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

# PDS: a simplified framework

**Algorithm 1:** Direct Search based on sufficient decrease

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

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

## PDS: a simplified framework

**Algorithm 1:** Direct Search based on sufficient decrease

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

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

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

# PDS: a simplified framework

**Algorithm 1:** Direct Search based on sufficient decrease

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

**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)

# PDS: a simplified framework

**Algorithm 1:** Direct Search based on sufficient decrease

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

**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 for simplicity)

# PDS: a simplified framework

**Algorithm 1:** Direct Search based on sufficient decrease

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

**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 for simplicity)

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

# PDS: a simplified framework

**Algorithm 1:** Direct Search based on sufficient decrease

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

**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 for simplicity)

**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**

# PDS: a simplified framework

**Algorithm 1:** Direct Search based on sufficient decrease

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

**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 for simplicity)
>
> **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)

# PDS: a simplified framework

**Algorithm 1:** Probabilistic Direct Search based on sufficient decrease

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

**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 for simplicity)

**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)

# PDS: a simplified framework

**Algorithm 1:** Probabilistic Direct Search based on sufficient decrease

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

**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 for simplicity)

  **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 $\{\mathcal{D}_k\}$ (GRVZ 2015): $\mathcal{D}_k = \{d_1, \ldots, d_m\}$ with $d_\ell \overset{\text{i.i.d.}}{\sim} \mathrm{U}(\mathcal{S}^{n-1})$

# PDS: a simplified framework

**Algorithm 1:** Probabilistic Direct Search based on sufficient decrease

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

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

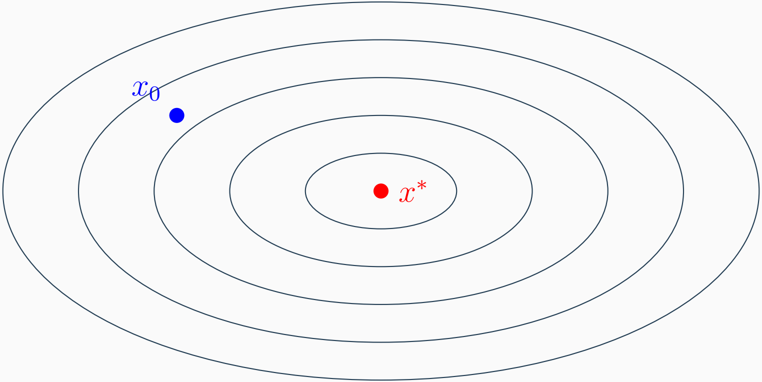
- Select a finite set of directions $\mathcal{D}_k \subset \mathbb{R}^n$ randomly.
- (In this talk, assume $\mathcal{D}_k$ is a set of unit vectors for simplicity)
- Set $d_k = \arg\min\{f(x_k + \alpha_k d) : d \in \mathcal{D}_k\}$.
- (complete polling for simplicity)
- **if** $f(x_k + \alpha_k d_k) < f(x_k) - c\alpha_k^2$ **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 $\{\mathcal{D}_k\}$ (GRVZ 2015): $\mathcal{D}_k = \{d_1, \ldots, d_m\}$ 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)

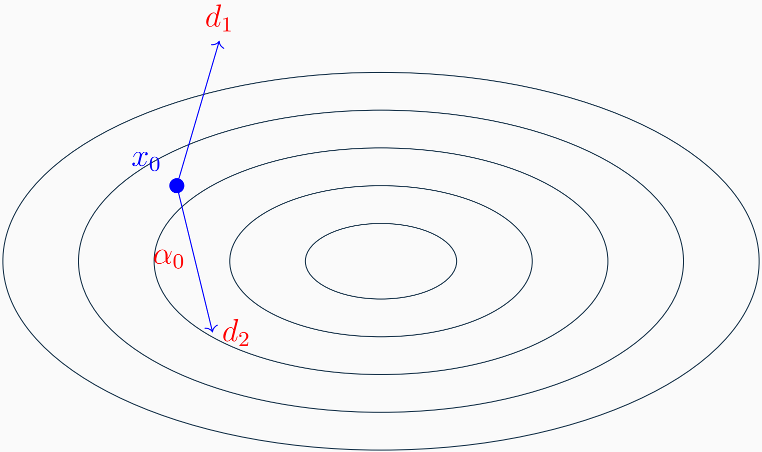
# Illustration of how PDS works

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



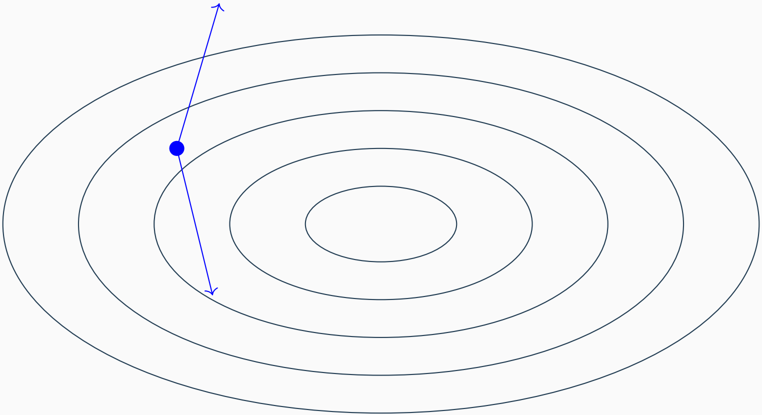
# Illustration of how PDS works

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



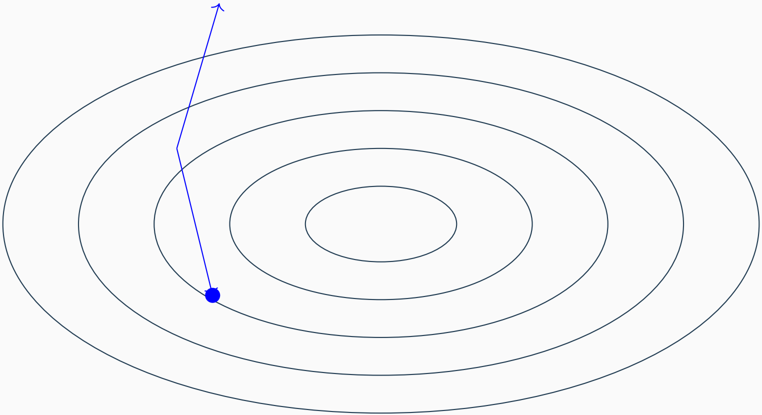
# Illustration of how PDS works

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



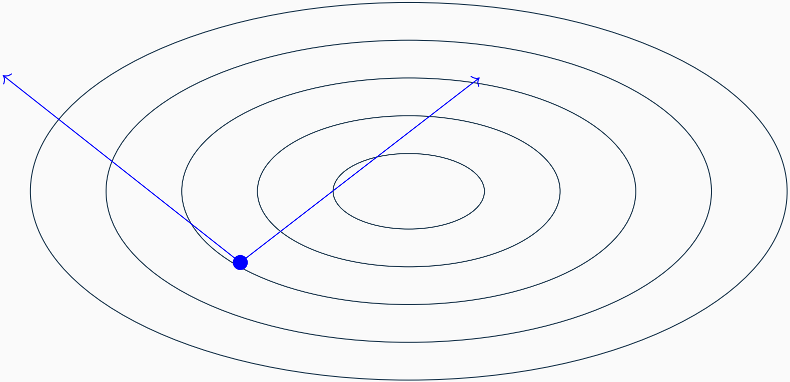
# Illustration of how PDS works

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



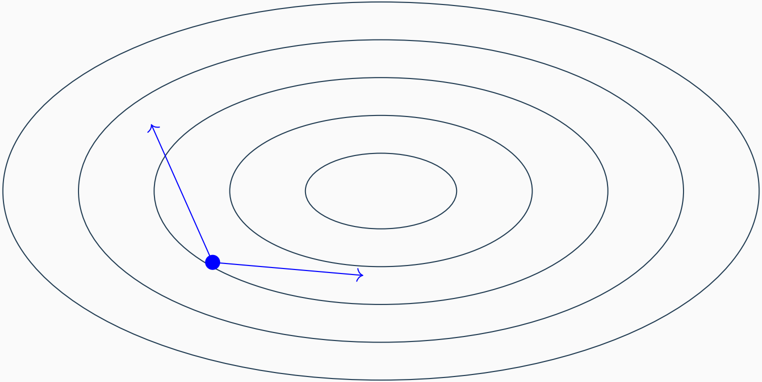
# Illustration of how PDS works

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



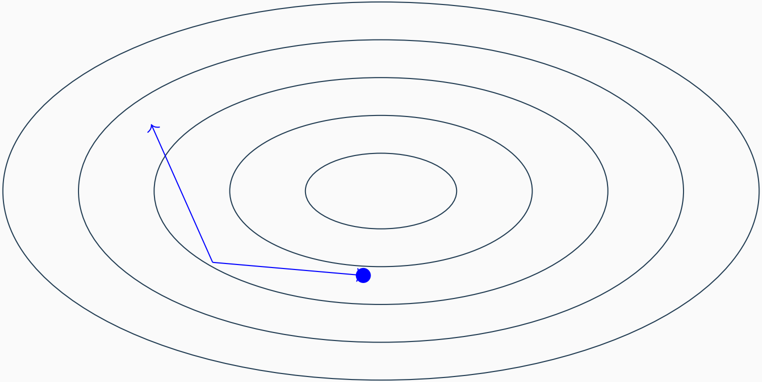
# Illustration of how PDS works

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



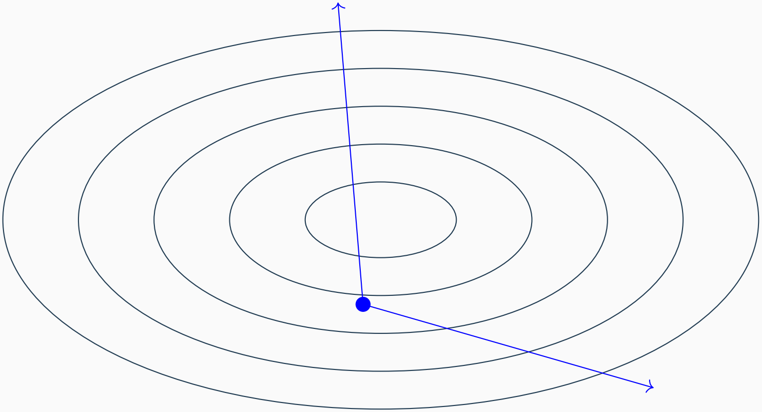
# Illustration of how PDS works

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



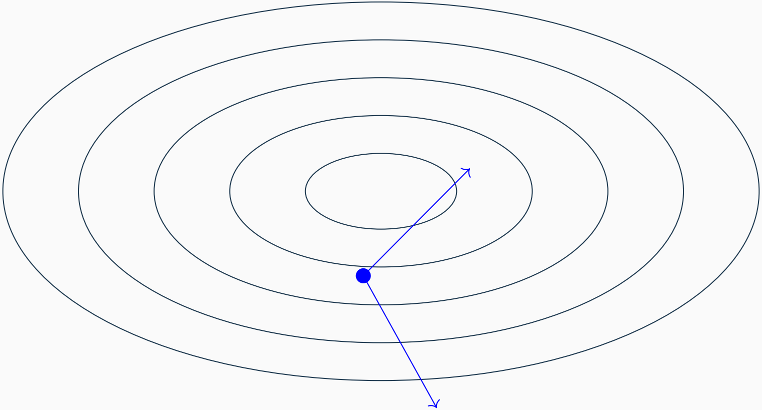
# Illustration of how PDS works

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



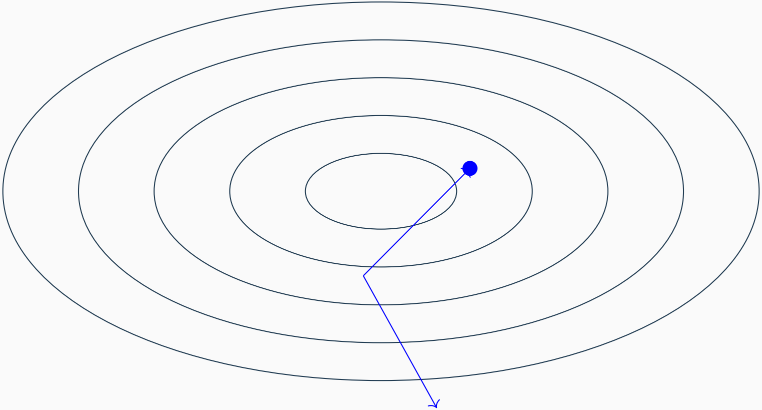
# Illustration of how PDS works

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



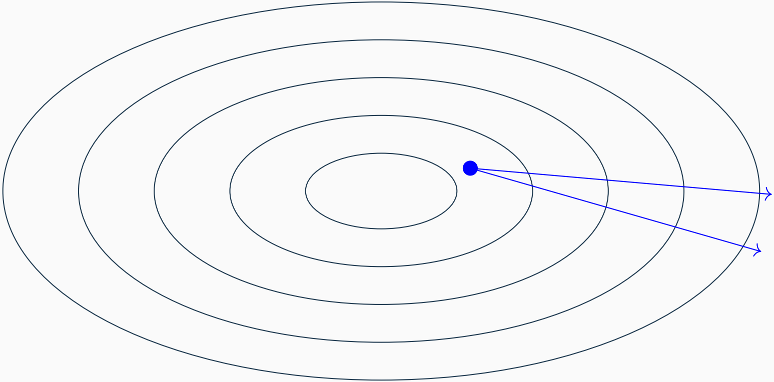
# Illustration of how PDS works

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



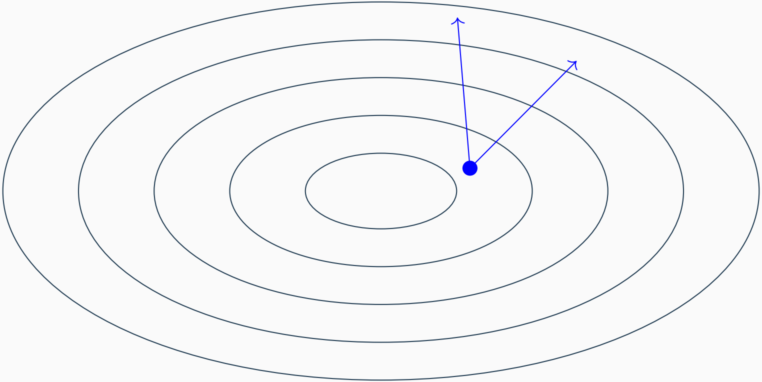
# Illustration of how PDS works

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



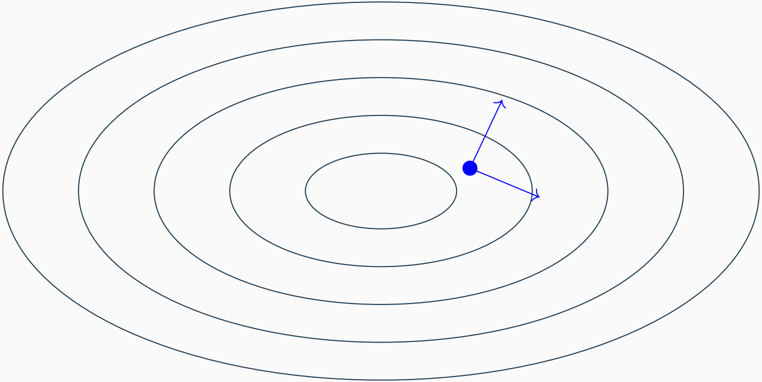
# Illustration of how PDS works

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



# Illustration of how PDS works

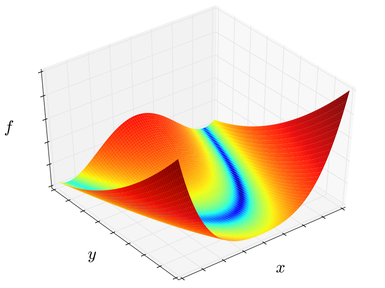
$\mathcal{D}_k = \{d_1, d_2\}$, where $d_\ell \overset{\text{i.i.d.}}{\sim} \mathsf{U}(\mathcal{S}^1)$

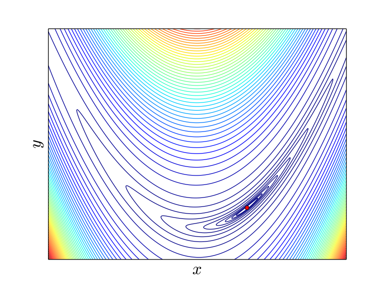


# A numerical example: CS v.s. PDS with $2$ directions

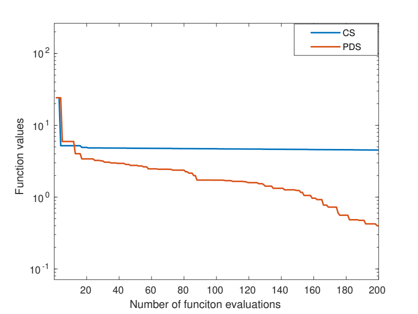
Rosenbrock "banana" function:

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

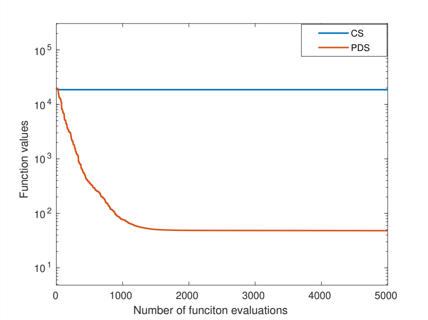


# A numerical example: CS v.s. PDS with $2$ directions




$n = 2$




$n = 50$

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)

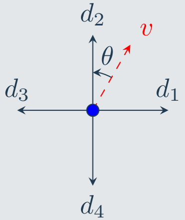
Given a finite set $\mathcal{D} \subseteq \mathbb{R}^n \setminus \{0\}$ and a vector $v \in \mathbb{R}^n \setminus \{0\}$, define

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

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)

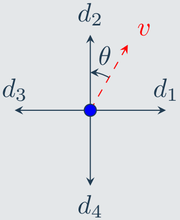
Given a finite set $\mathcal{D} \subseteq \mathbb{R}^n \backslash \{0\}$ and a vector $v \in \mathbb{R}^n \backslash \{0\}$, define

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

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” $v$

# Convergence theory

## Definition ($p$-probabilistically $\kappa$-descent)

$\{\mathcal{D}_k\}$ is said to be $p$-probabilistically $\kappa$-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 $g_k = \nabla f(x_k)$.

## Intuitive meaning of $p$-probabilistically $\kappa$-descent

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 $\kappa$-descent)**

$\{\mathcal{D}_k\}$ is said to be $p$-probabilistically $\kappa$-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 $g_k = \nabla f(x_k)$.

**Intuitive meaning of $p$-probabilistically $\kappa$-descent**

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 $\kappa$-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 questions

**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 questions

**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*

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

## Questions:

- Is $p_0$-probabilistically $\kappa$-descent an essential assumption or a technical one? (Such supermartingale-like assumptions are ubiquitous in the convergence analysis of randomized methods!)
- What will happen if

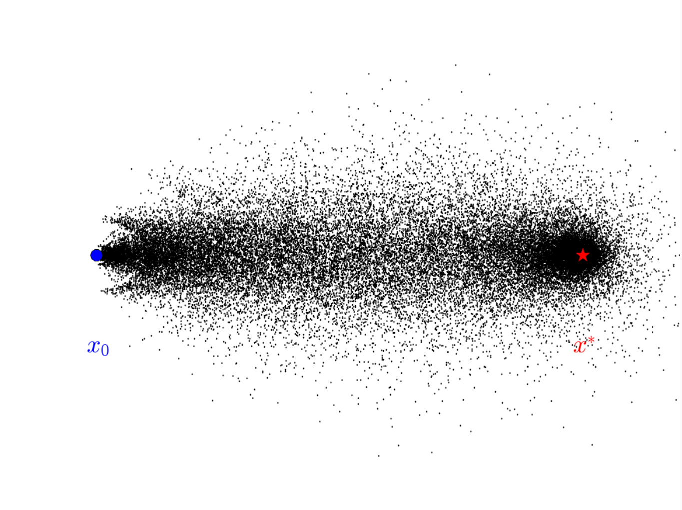
$$m \leq \log_2\left(1 - \frac{\log\theta}{\log\gamma}\right)?$$

# A simple test

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

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

# 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 $\kappa$-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 $\kappa$-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 $\kappa$-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?

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

# Why assuming convexity?

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

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

# Why assuming convexity?

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

- Convexity connects $\operatorname{cm}\left(\mathcal{D}_k, -g_k\right) \le 0$ and shrinking of step size
- $\{\mathcal{D}_k\}$ is probabilistic ascent implies $\alpha_k$ “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

$\Downarrow$

$\alpha_k$ "often" shrinks

# From probabilistic ascent to non-convergence: How?

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

$\Downarrow$

$\alpha_k$ "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

$\Downarrow$

$\alpha_k$ "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 $\mathrm{dist}(x_0, \mathcal{S}^*)$ 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 $\mathcal{D}_k$":

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

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

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

- Use the above inequality iteratively:

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

- Get an upper bound of the series of step sizes:

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

- 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 $\zeta$ such that

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

- (Q2) Moreover, can we specify the value of $\zeta$?

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

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

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

## Note

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

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

# 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

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

# 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

$$\left(\frac{1}{2}\right)^m \; > \; q_0 \; = \; \frac{\log \gamma}{\log(\theta^{-1}\gamma)},$$

# 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

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

which is equivalent to

$$\left(\frac{1}{2}\right)^m \ > \ q_0 \ = \ \frac{\log\gamma}{\log(\theta^{-1}\gamma)},$$

or, equivalently,

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

# 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 $\{\mathcal{D}_k\}$:

$$q\text{-probabilistically ascent with } q > q_0.$$

Natural question:

$$\text{Is it sufficient to require } q \geq q_0?$$

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 results inspired by non-convergence analysis

Consider the series

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

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

# Convergence results inspired by non-convergence analysis

Consider the series

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

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

Roughly speaking, $S(0) < \infty$ implies non-convergence of PDS.

What can we say about convergence using $S(\kappa)$?

# Convergence results inspired by non-convergence analysis

Consider the series

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

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

Roughly speaking, $S(0) < \infty$ implies non-convergence of PDS.

What can we say about convergence using $S(\kappa)$?

**Theorem**

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

# Convergence results inspired by non-convergence analysis

Consider the series

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

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

Roughly speaking, $S(0) < \infty$ implies non-convergence of PDS.

What can we say about convergence using $S(\kappa)$?

**Theorem**

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

**Relation with existing result in [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

- sharpen our knowledge about the algorithm,
- deepen our understanding about the convergence analysis,
- guide the selection of algorithmic parameters, and
- provide new perspectives on convergence analysis.

Thank you!

# One more thing: OptiProfiler

github.com/optiprofiler

OptiProfiler (joint work with Cunxin Huang and Tom M. Ragonneau) 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]
- Providing multiple types of tests
  noisy function, unrelaxable constraints, randomized initial point ...
- Implemented in Python and MATLAB
- Default problem set: S2MPJ [Gratton, Toint 2024]

# One more thing: OptiProfiler

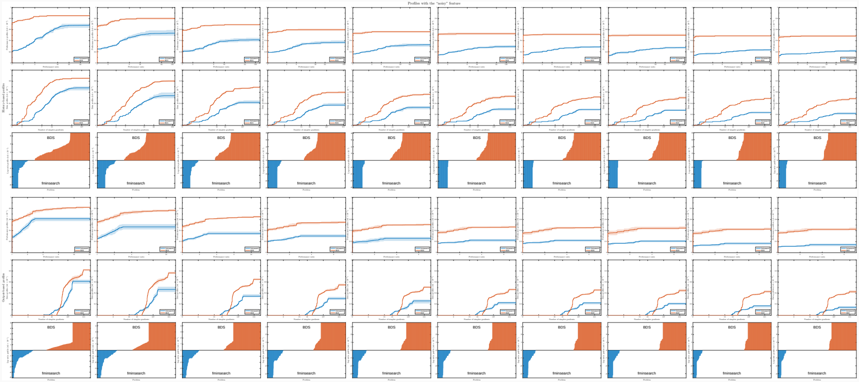
Example (MATLAB):

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

# One more thing: OptiProfiler

Example (MATLAB):

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




N.B.: Separate profiles can also be generated.

# 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.
- Ghanbari, H. and Scheinberg, K. (2017). "Black-box optimization in machine learning with trust region based derivative free algorithm". *arXiv:1703.06925*.

# References II


- 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.
- Powell, M. J. D. (1973). **"On search directions for minimization algorithms".** *Math. Program.* 4, pp. 193–201.

# References III


- 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.