Non-convergence Analysis of Probabilistic Direct Search

Cunxin Huang
Co-supervised by Dr. Zaikun Zhang and Prof. Xiaojun Chen

Graduate Student Forum of ORSC Mathematical Programming Branch
Chongqing, China, 2023

Department of Applied Mathematics
The Hong Kong Polytechnic University

Table of contents

Preliminaries

What is derivative-free optimization and why

Derivative-free optimization (DFO)

- A branch of optimization
- Do not use derivatives (only use function evaluations)

What is derivative-free optimization and why

Derivative-free optimization (DFO)

- A branch of optimization
- Do not use derivatives (only use function evaluations)

Why DFO?

- Derivatives are not available
- Problems are often noisy (finite difference?)
- Function evaluations are expensive (e.g.: PDE simulation)

What is derivative-free optimization and why

Derivative-free optimization (DFO)

- A branch of optimization
- Do not use derivatives (only use function evaluations)

Why DFO?

- Derivatives are not available
- Problems are often noisy (finite difference?)
- Function evaluations are expensive (e.g.: PDE simulation)

Typical situation: black box

$$x \in \mathbb{R}^n \longrightarrow \boxed{f : \mathbb{R}^n \to \mathbb{R}} \longrightarrow f(x)$$

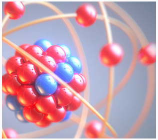
What is derivative-free optimization and why

Derivative-free optimization (DFO)

- A branch of optimization
- Do not use derivatives (only use function evaluations)

Why DFO?

- Derivatives are not available
- Problems are often noisy (finite difference?)
- Function evaluations are expensive (e.g.: PDE simulation)



Nuclear Physics

Machine Learning

Cosmology

Basic assumptions

In this talk, we consider the unconstrained problem

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

where

- ∇f is Lipschitz continuous, although it cannot be evaluated,
- f is bounded below.

Probabilistic direct search (PDS)

Algorithm 1: Probabilistic direct search based on sufficient decrease

Probabilistic direct search (PDS)

Algorithm 1: Probabilistic direct search based on sufficient decrease

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

Probabilistic direct search (PDS)

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**

Probabilistic direct search (PDS)

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 $D_k \subset \mathbb{R}^n$ randomly.

Probabilistic direct search (PDS)

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 $D_k \subset \mathbb{R}^n$ randomly.

(In this talk, assume D_k is a set of unit vectors for simplicity.)

Probabilistic direct search (PDS)

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 $D_k \subset \mathbb{R}^n$ randomly.

(In this talk, assume D_k is a set of unit vectors for simplicity.)

if $f(x_k + \alpha_k d) < f(x_k) - \alpha_k^2$ for some $d \in D_k$ **then**

Probabilistic direct search (PDS)

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 $D_k \subset \mathbb{R}^n$ randomly.

  (In this talk, assume D_k is a set of unit vectors for simplicity.)

  if $f(x_k + \alpha_k d) < f(x_k) - \alpha_k^2$ for some $d \in D_k$ **then**

    Expand step size, and move to that point

Probabilistic direct search (PDS)

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 $D_k \subset \mathbb{R}^n$ randomly.

  (In this talk, assume D_k is a set of unit vectors for simplicity.)

  if $f(x_k + \alpha_k d) < f(x_k) - \alpha_k^2$ for some $d \in D_k$ **then**

    Expand step size, and move to that point

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

  else

Probabilistic direct search (PDS)

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 $D_k \subset \mathbb{R}^n$ randomly.

  (In this talk, assume D_k is a set of unit vectors for simplicity.)

  if $f(x_k + \alpha_k d) < f(x_k) - \alpha_k^2$ for some $d \in D_k$ **then**

    Expand step size, and move to that point

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

  else

    Shrink step size, and stand still

Probabilistic direct search (PDS)

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 $D_k \subset \mathbb{R}^n$ randomly.

  (In this talk, assume D_k is a set of unit vectors for simplicity.)

  if $f(x_k + \alpha_k d) < f(x_k) - \alpha_k^2$ for some $d \in D_k$ **then**

    Expand step size, and move to that point

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

  else

    Shrink step size, and stand still

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

Probabilistic direct search (PDS)

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 $D_k \subset \mathbb{R}^n$ randomly. How to select?

  (In this talk, assume D_k is a set of unit vectors for simplicity.)

  if $f(x_k + \alpha_k d) < f(x_k) - \alpha_k^2$ for some $d \in D_k$ **then**

    Expand step size, and move to that point

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

  else

    Shrink step size, and stand still

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

Probabilistic direct search (PDS)

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 $D_k \subset \mathbb{R}^n$ randomly. How to select?
- (In this talk, assume D_k is a set of unit vectors for simplicity.)
- **if** $f(x_k + \alpha_k d) < f(x_k) - \alpha_k^2$ for some $d \in D_k$ **then**
  - Expand step size, and move to that point
  - Set $\alpha_{k+1} = \gamma \alpha_k$ and $x_{k+1} = x_k + \alpha_k d$.
- **else**
  - Shrink step size, and stand still
  - Set $\alpha_{k+1} = \theta \alpha_k$ and $x_{k+1} = x_k$.

Typical choice: $D_k = \{d_1, \ldots, d_m\}$, where $d_i \overset{\text{i.i.d.}}{\sim} \text{Unif}(\mathcal{S}^{n-1})$

Illustration of how PDS works

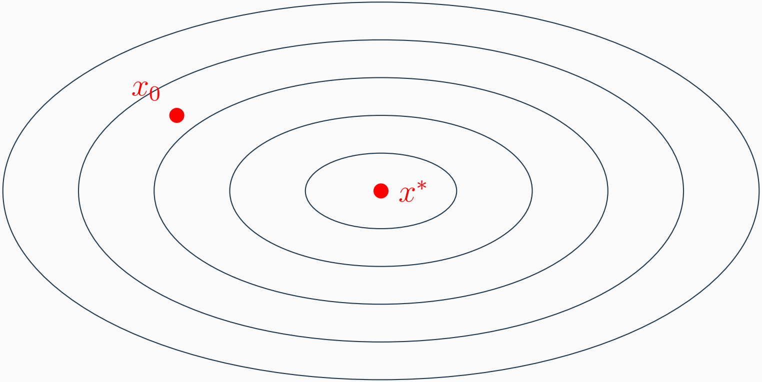


Illustration of how PDS works

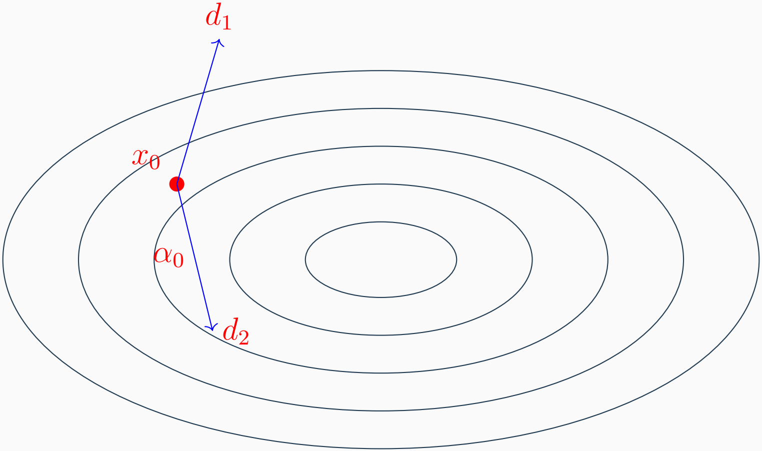


Illustration of how PDS works

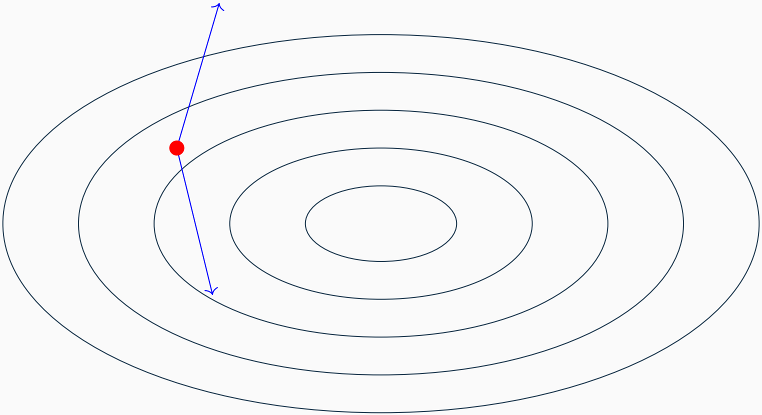


Illustration of how PDS works

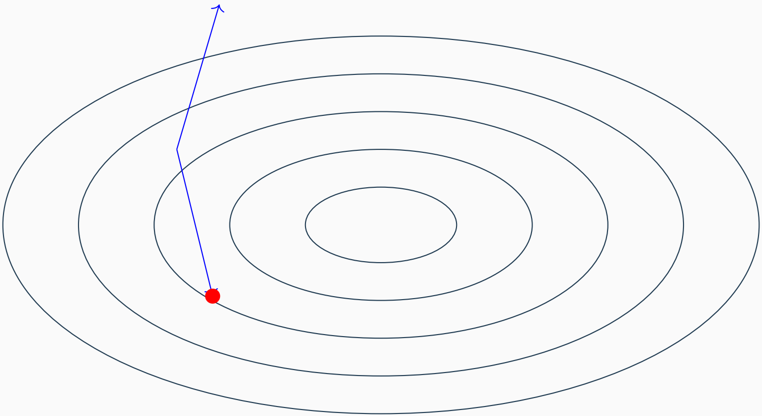


Illustration of how PDS works

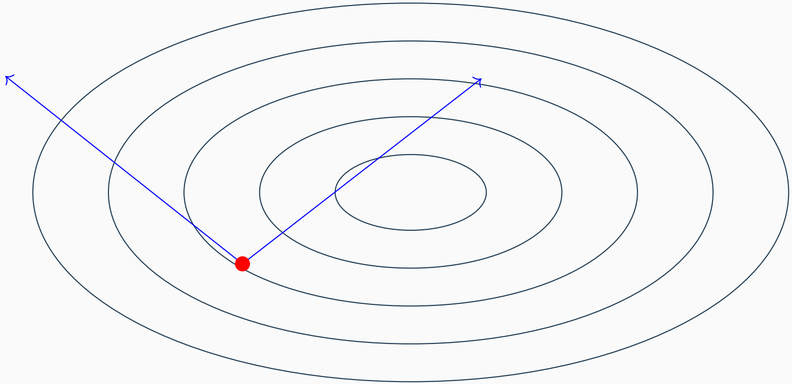


Illustration of how PDS works

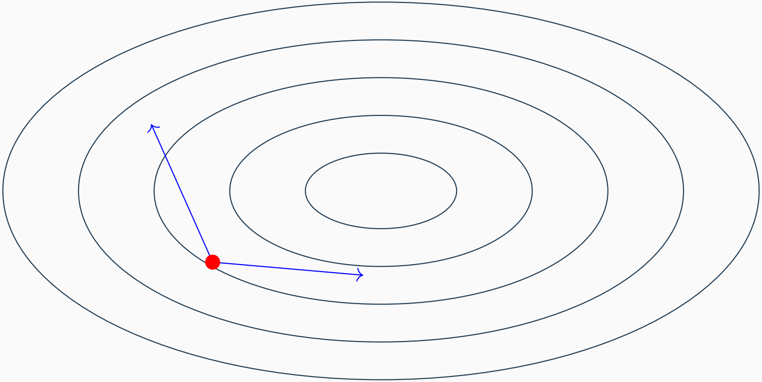


Illustration of how PDS works

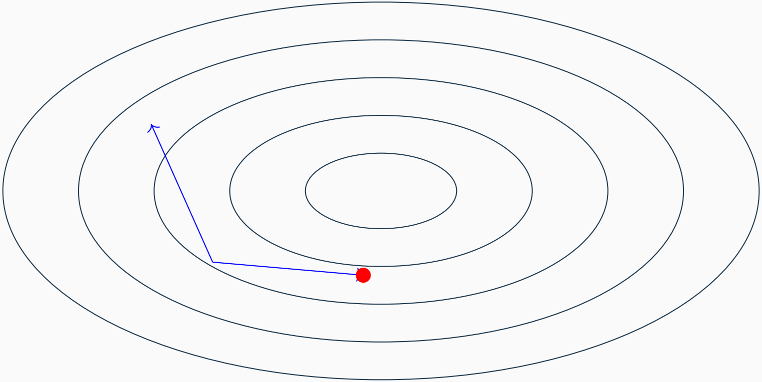


Illustration of how PDS works

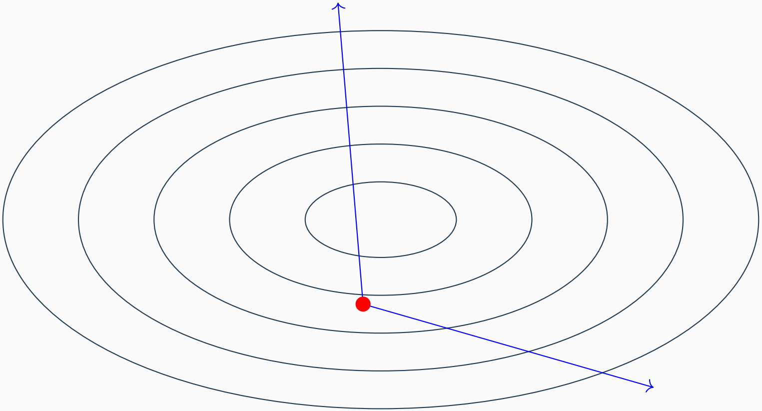


Illustration of how PDS works

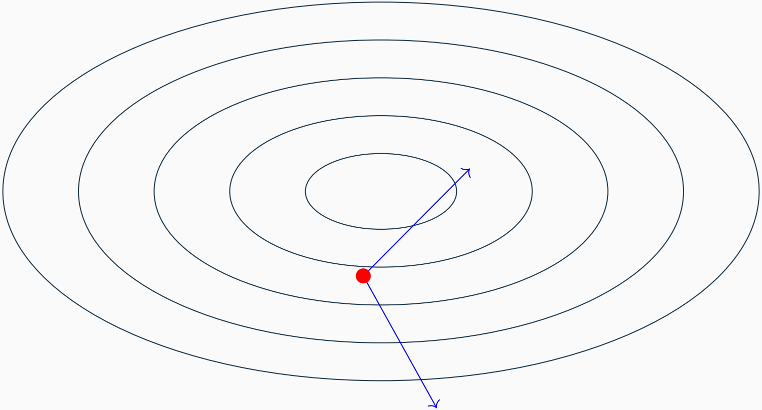


Illustration of how PDS works

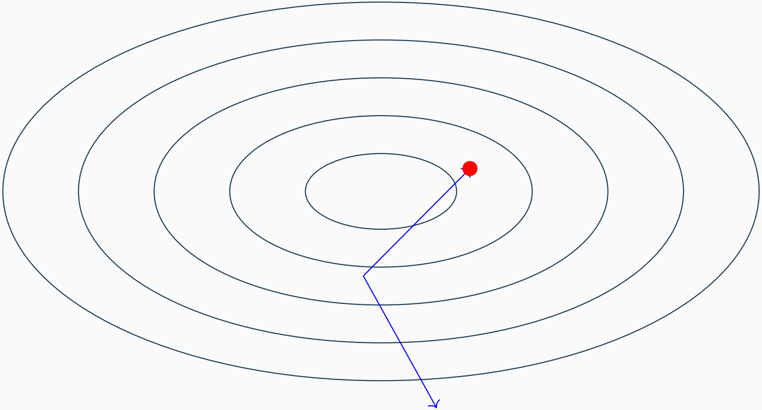


Illustration of how PDS works

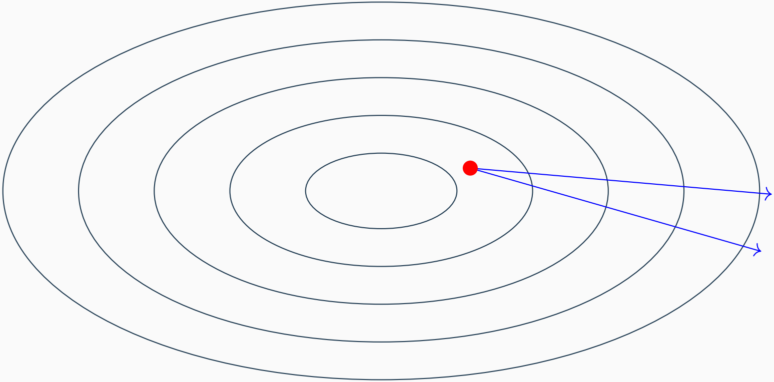


Illustration of how PDS works

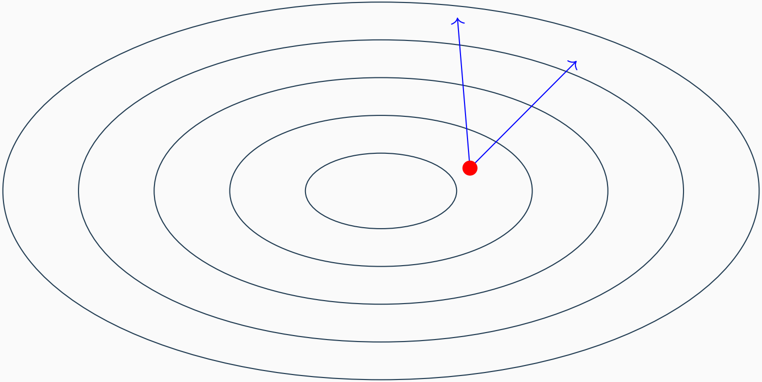
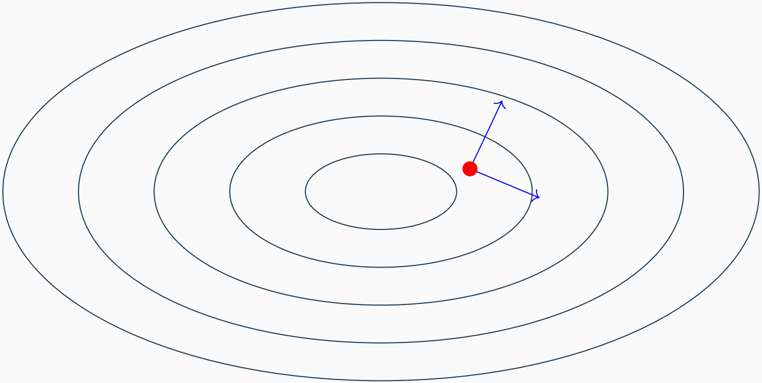


Illustration of how PDS works



Definition (Cosine measure)

Cosine measure for a finite set of nonzero vectors $D \subseteq \mathbb{R}^n$ w.r.t. a given vector $v \in \mathbb{R}^n$:

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

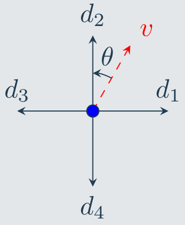
Definition (Cosine measure)

Cosine measure for a finite set of nonzero vectors $D \subseteq \mathbb{R}^n$ w.r.t. a given vector $v \in \mathbb{R}^n$:

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

Example

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



Convergence theory

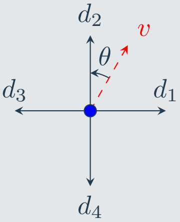
Definition (Cosine measure)

Cosine measure for a finite set of nonzero vectors $D \subseteq \mathbb{R}^n$ w.r.t. a given vector $v \in \mathbb{R}^n$:

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

Example

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



Measure the ability that "D approximates v"

Convergence theory

Definition (p-probabilistically κ-descent)

$(D_k)_{k\geq 0}$ is said to be p-probabilistically κ-descent, if

$$\mathbb{P}\left(\mathrm{cm}(D_k, -g_k) \geq \kappa \mid D_0, \ldots, D_{k-1}\right) \;\geq\; p \quad \text{for each } k \geq 0,$$

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

Convergence theory

Definition (p-probabilistically κ-descent)

$(D_k)_{k\geq 0}$ is said to be p-probabilistically κ-descent, if

$$\mathbb{P}\left(\mathrm{cm}(D_k, -g_k) \geq \kappa \mid D_0, \dots, D_{k-1}\right) \;\geq\; p \quad \text{for each } k \geq 0,$$

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

Intuition:

each D_k is “good enough with lower-bounded probability”,
no matter what happened before

Convergence theory

Definition (p-probabilistically κ-descent)

$(D_k)_{k\geq 0}$ is said to be p-probabilistically κ-descent, if

$$\mathbb{P}\left(\mathrm{cm}(D_k, -g_k) \geq \kappa \mid D_0, \ldots, D_{k-1}\right) \;\geq\; p \quad \text{for each } k \geq 0,$$

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

Intuition:

each D_k is “good enough with lower-bounded probability”, no matter what happened before

Theorem (Gratton et al. 2015)

If $(D_k)_{k\geq 0}$ is p-probabilistically κ-descent with $\kappa > 0$ and

$$p = \log\theta / \log(\gamma^{-1}\theta),$$

then PDS converges with probability 1.

Practical choice and natural question

Corollary (Gratton et al. 2015)

If $D_k = \{d_1, \ldots, d_m\}$, where $d_i \overset{\text{i.i.d.}}{\sim}$ Unif$(\mathcal{S}^{n-1})$, then PDS is convergent if

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

Practical choice and natural question

Corollary (Gratton et al. 2015)

If $D_k = \{d_1, \ldots, d_m\}$, where $d_i \overset{\text{i.i.d.}}{\sim}$ Unif$(\mathcal{S}^{n-1})$, then PDS is convergent if

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

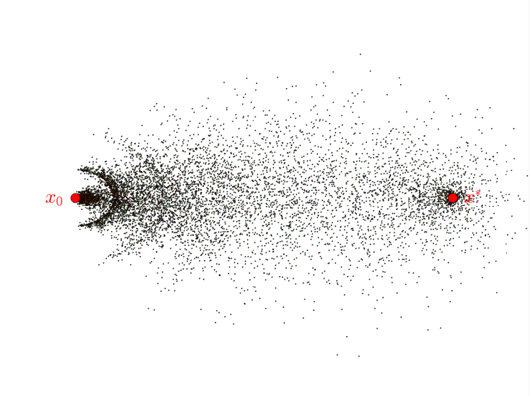
A nature question: what if

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

Practical choice and natural question

A nature question: what if

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



Non-convergence analysis

Motivation: non-convergence analysis matters

Many well-known algorithms have non-convergence analysis.

- S. Reddi, S. Kale, and S. Kumar. On the convergence of Adam and beyond. In Y. Bengio, Y. LeCun, T. Sainath, I. Murray, M. Ranzato, and O. Vinyals, editors, *International Conference on Learning Representations (ICLR 2018)*. Curran Associates, Inc., 2018.
- C. Chen, B. He, Y. Ye, and X. Yuan. The direct extension of ADMM for multi-block convex minimization problems is not necessarily convergent. *Math. Program.*, 155:57-79, 2016.
- W. Mascarenhas. The divergence of the BFGS and Gauss Newton methods. *Math. Program.*, 147:253-276, 2014.
- ...

Naive idea of non-convergence

Recall p-probabilistically κ-descent:

$$\mathbb{P}\left(\mathrm{cm}\left(D_k, -g_k\right) \geq \kappa \mid D_0, \ldots, D_{k-1}\right) \geq p \quad \text{for each } k \geq 0.$$

Naive idea of non-convergence

Recall p-probabilistically κ-descent:

$$\mathbb{P}\left(\mathrm{cm}\left(D_k, -g_k\right) \geq \kappa \mid D_0, \ldots, D_{k-1}\right) \geq p \quad \text{for each } k \geq 0.$$

q-probabilistically ascent

$$\mathbb{P}\left(\mathrm{cm}\left(D_k, -g_k\right) > 0 \mid D_0, \ldots, D_{k-1}\right) \leq q \quad \text{for each } k \geq 0.$$

Naive idea of non-convergence

q-probabilistically ascent

$$\mathbb{P}\left(\mathrm{cm}\left(D_k, -g_k\right) > 0 \mid D_0, \ldots, D_{k-1}\right) \leq q \quad \text{for each } k \geq 0.$$

$$\mathrm{cm}\left(D_k, -g_k\right) \leq 0 \longrightarrow \text{No descent direction}$$

$$\text{No descent direction} \xrightarrow{f \text{ convex}} f(x_k + \alpha_k d_k) \geq f(x_k) \; \forall d_k \in D_k$$

$$f(x_k + \alpha_k d_k) \geq f(x_k) \; \forall d_k \in D_k \longrightarrow \alpha_k \text{ shrinks}$$

Naive idea of non-convergence

q-probabilistically ascent

$$\mathbb{P}\left(\operatorname{cm}\left(D_k, -g_k\right) > 0 \mid D_0, \ldots, D_{k-1}\right) \leq q \quad \text{for each } k \geq 0.$$

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

non-convergence for convex functions

$\Downarrow$

non-convergence in general

Establishment of non-convergence

$(D_k)_{k\in\mathbb{N}}$ is
q-probabilistically ascent

Establishment of non-convergence

$(D_k)_{k\in\mathbb{N}}$ is
q-probabilistically ascent

$$\Downarrow$$

α_k shrinks with high probability

Establishment of non-convergence

$(D_k)_{k\in\mathbb{N}}$ is
q-probabilistically ascent

⬇

α_k shrinks with high probability

⬇

$$\sum_{k=0}^{\infty} \alpha_k \quad < \quad \infty \qquad \text{a.s. ?}$$

Establishment of non-convergence

$(D_k)_{k\in\mathbb{N}}$ is
q-probabilistically ascent

⬇

α_k shrinks with high probability

⬇

$$\sum_{k=0}^{\infty} \alpha_k \; < \; \infty \quad \text{a.s. ?}$$

⬇

$$x_0 \; \notin \; \bar{\mathcal{B}}(x^*, \textstyle\sum_{k=0}^{\infty} \alpha_k)$$

Establishment of non-convergence

$(D_k)_{k\in\mathbb{N}}$ is
q-probabilistically ascent

$\Downarrow$

α_k shrinks with high probability

$\Downarrow$

$$\sum_{k=0}^{\infty} \alpha_k \;<\; \infty \quad \text{a.s. ?}$$

$\Downarrow$

$$x_0 \;\notin\; \bar{\mathcal{B}}(x^*, \textstyle\sum_{k=0}^{\infty} \alpha_k)$$

$\Downarrow$

$\mathbb{P}\,$(Convergence) $\;<\; 1$

Key to analysis

- Define indicator function $Y_k = \mathbb{1}_{\{\mathrm{cm}(D_k, -g_k) > 0\}}$
  Indicator for "good" event

Key to analysis

- Define indicator function $Y_k = \mathbb{1}_{\{\mathrm{cm}(D_k, -g_k) > 0\}}$
  Indicator for "good" event
- $\alpha_{k+1} \leq \gamma^{Y_k} \theta^{1-Y_k} \alpha_k$, when f is convex

Key to analysis

- Define indicator function $Y_k = \mathbb{1}_{\{\mathrm{cm}(D_k, -g_k) > 0\}}$
  Indicator for "good" event
- $\alpha_{k+1} \leq \gamma^{Y_k} \theta^{1-Y_k} \alpha_k$, when f is convex
- $\alpha_k \leq \alpha_0 \prod_{i=0}^{k-1} \gamma^{Y_i} \theta^{1-Y_i} =: \alpha_0 U_k$

Key to analysis

- Define indicator function $Y_k = \mathbb{1}_{\{\mathrm{cm}(D_k, -g_k) > 0\}}$
  Indicator for "good" event
- $\alpha_{k+1} \leq \gamma^{Y_k} \theta^{1-Y_k} \alpha_k$, when f is convex
- $\alpha_k \leq \alpha_0 \prod_{i=0}^{k-1} \gamma^{Y_i} \theta^{1-Y_i} =: \alpha_0 U_k$
- $\sum_{k=1}^{\infty} \alpha_k \leq \alpha_0 \sum_{k=1}^{\infty} U_k$

Key to analysis

- Define indicator function $Y_k = \mathbb{1}_{\{\mathrm{cm}(D_k, -g_k) > 0\}}$
  Indicator for "good" event
- $\alpha_{k+1} \leq \gamma^{Y_k} \theta^{1-Y_k} \alpha_k$, when f is convex
- $\alpha_k \leq \alpha_0 \prod_{i=0}^{k-1} \gamma^{Y_i} \theta^{1-Y_i} =: \alpha_0 U_k$
- $\sum_{k=1}^{\infty} \alpha_k \leq \alpha_0 \sum_{k=1}^{\infty} U_k < \infty$ a.s. ?

Key to analysis

- Define indicator function $Y_k = \mathbb{1}_{\{\mathrm{cm}(D_k, -g_k) > 0\}}$
  Indicator for "good" event
- $\alpha_{k+1} \le \gamma^{Y_k}\theta^{1-Y_k}\alpha_k$, when f is convex
- $\alpha_k \le \alpha_0 \prod_{i=0}^{k-1} \gamma^{Y_i}\theta^{1-Y_i} =: \alpha_0 U_k$
- $\sum_{k=1}^{\infty} \alpha_k \le \alpha_0 \sum_{k=1}^{\infty} U_k < \infty$ a.s. ?

Under q-probabilistically ascent assumption, can we find a constant ζ such that

$$\mathbb{P}\left(\sum_{k=1}^{\infty} U_k < \zeta\right) \; > \; 0?$$

Main results

Assumption

$$\mathbb{P}\left(\mathrm{cm}\left(D_k, -g_k\right) \le 0 \mid D_0, \ldots, D_{k-1}\right) \;\ge\; q \;>\; q_0 \quad \text{for each } k \ge 0,$$

where $q_0 = 1 - p_0 = \log\gamma / \log(\theta^{-1}\gamma)$.

Main results

Assumption

$$\mathbb{P}\left(Y_k = 0 \mid Y_0, \ldots, Y_{k-1}\right) \;\geq\; q \;>\; q_0 \quad \text{for each } k \geq 0,$$

where $q_0 = 1 - p_0 = \log\gamma / \log(\theta^{-1}\gamma)$.

Main results

Assumption

$$\mathbb{P}\left(Y_k = 0 \mid Y_0, \dots, Y_{k-1}\right) \;\geq\; q \;>\; q_0 \quad \text{for each } k \geq 0,$$

where $q_0 = 1 - p_0 = \log\gamma / \log(\theta^{-1}\gamma)$.

Result

1. $$\mathbb{P}\left(\sum_{k=1}^{\infty} U_k < \infty\right) \;=\; 1$$

2. $$\mathbb{P}\left(\sum_{k=1}^{\infty} U_k < \zeta\right) \;>\; 0 \quad\Longleftrightarrow\quad \zeta > \frac{\theta}{1-\theta}$$

Main results

Assumption

$$\mathbb{P}\left(Y_k = 0 \mid Y_0, \ldots, Y_{k-1}\right) \;\geq\; q \;>\; q_0 \quad \text{for each } k \geq 0,$$

where $q_0 = 1 - p_0 = \log\gamma / \log(\theta^{-1}\gamma)$.

Result

1. $$\mathbb{P}\left(\sum_{k=1}^{\infty} U_k < \infty\right) \;=\; 1$$
2. $$\mathbb{P}\left(\sum_{k=1}^{\infty} U_k < \zeta\right) \;>\; 0 \quad \iff \quad \zeta > \frac{\theta}{1-\theta}$$

Note that $\sum_{k=1}^{\infty} U_k = \sum_{k=1}^{\infty} \prod_{i=0}^{k-1} \gamma^{Y_i}\theta^{1-Y_i} \geq \theta/(1-\theta)$

Main results

Assumption

$$\mathbb{P}\left(Y_k = 0 \mid Y_0, \ldots, Y_{k-1}\right) \;\geq\; q \;>\; q_0 \quad \text{for each } k \geq 0,$$

where $q_0 = 1 - p_0 = \log\gamma / \log(\theta^{-1}\gamma)$.

Result

1. $$\mathbb{P}\left(\sum_{k=1}^{\infty} U_k < \infty\right) \;=\; 1$$

2. $$\cancel{\mathbb{P}\left(\sum_{k=1}^{\infty} U_k < \zeta\right) \;>\; 0 \quad \Longleftrightarrow \quad \zeta > \frac{\theta}{1-\theta}}$$

Main results

Assumption

$$\liminf_{k\to\infty} \mathbb{P}\left(Y_k = 0 \mid Y_0, \ldots, Y_{k-1}\right) \; > \; q_0,$$

where $q_0 = 1 - p_0 = \log\gamma / \log(\theta^{-1}\gamma)$.

Result

1. $$\mathbb{P}\left(\sum_{k=1}^{\infty} U_k < \infty\right) \; = \; 1$$

2. $$\cancel{\mathbb{P}\left(\sum_{k=1}^{\infty} U_k < \zeta\right) \; > \; 0 \quad \Longleftrightarrow \quad \zeta > \frac{\theta}{1-\theta}}$$

Tightness of analysis

Almost zero gap

Let $D_k = \{d_1, \ldots, d_m\}$, where $d_i \overset{\text{i.i.d.}}{\sim} \text{Unif}(\mathcal{S}^{n-1})$.

Recall that PDS is convergent if

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

Almost zero gap

Let $D_k = \{d_1, \ldots, d_m\}$, where $d_i \overset{\text{i.i.d.}}{\sim}$ Unif$(\mathcal{S}^{n-1})$.

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

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

Tightness of assumption

Natural question:

$$\mathbb{P}\left(\mathrm{cm}\left(D_k, -g_k\right) \leq 0 \mid D_0, \ldots, D_{k-1}\right) \;\geq\; q \cancel{>} \geq q_0,$$

Tightness of assumption

Natural question:

$$\mathbb{P}\left(\mathrm{cm}\left(D_k, -g_k\right) \leq 0 \mid D_0, \ldots, D_{k-1}\right) \;\geq\; q \cancel{>} \geq q_0,$$

Answer: NO

Tightness of assumption

Natural question:

$$\mathbb{P}\left(\mathrm{cm}\left(D_k, -g_k\right) \leq 0 \mid D_0, \ldots, D_{k-1}\right) \; \geq \; q \cancel{>} \geq q_0,$$

Answer: NO

Example

We assume

- $f : \mathbb{R}^n \to \mathbb{R}$ be gradient Lipschitz and strongly convex,
- $\theta = 1/2$ and $\gamma = 2$, $\Rightarrow$ $q_0 = 1/2$
- $D_k = \{g_k/\|g_k\|\}$ or $D_k = \{-g_k/\|g_k\|\}$ with probability $1/2$, respectively,

then we have

$$\mathbb{P}\left(\lim_{k\to\infty} \|g_k\| = 0\right) \; = \; 1.$$

Conclusions

Conclusions

In this talk

- Non-convergence analysis for probabilistic direct search
- Tightness of non-convergence analysis

Conclusions

In this talk

- Non-convergence analysis for probabilistic direct search
- Tightness of non-convergence analysis

Future work

- Estimate the non-convergence probability
- Conduct non-convergence analysis for other frameworks

Thank you!

References I

- Biviano, A. et al. (2013). "CLASH-VLT: the mass, velocity-anisotropy, and pseudo-phase-space density profiles of the $z = 0.44$ galaxy cluster MACS J1206.2-0847". *A&A* 558, A1:1–A1:22.
- 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.