Non-convergence of Probabilistic direct search

ICNONLA, taiyuan, China

Cunxin Huang, joint work with Zaikun Zhang

July 19, 2023

Department of Applied Mathematics
The Hong Kong Polytechnic University

Derivative-free optimization

Derivative-free optimization (DFO) studies optimization algorithms that do not use derivatives (only use function evaluations).

Derivative-free optimization

Derivative-free optimization

Derivative-free optimization (DFO) studies optimization algorithms that do not use derivatives (only use function evaluations).

Example (Typical case: f is a blackbox)

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

Derivative-free optimization

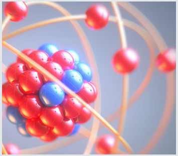
Derivative-free optimization

Derivative-free optimization (DFO) studies optimization algorithms that do not use derivatives (only use function evaluations).

Example (Typical case: f is a blackbox)

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

Applications



Nuclear Physics

Machine Learning

Cosmology

In this talk, we solve the unconstrained problem

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

where we follow standard setup that

- ∇f is Lipschitz continuous with constant ν, cannot be evaluated,
- f is bounded below,
- the evaluation of f is expensive.

Direct-search methods

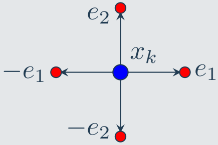
Sample points based on a finite direction set and make decisions by comparing values.

Direct-search methods

Sample points based on a finite direction set and make decisions by comparing values.

Example (typical direction set in $\mathbb{R}^2$)

$$D = \{e_1, -e_1, e_2, -e_2\}$$



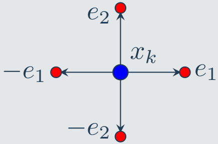
Direct-search methods

Direct-search methods

Sample points based on a finite direction set and make decisions by comparing values.

Example (typical direction set in $\mathbb{R}^2$)

$$D = \{e_1, -e_1, e_2, -e_2\}$$



In this talk, we assume direction set always be a set of unit vectors in $\mathbb{R}^n$.

Algorithm 1: Direct Search with sufficient decrease

Algorithm 1: Direct Search with sufficient decrease

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

Algorithm 1: Direct Search with sufficient decrease

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

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

Algorithm 1: Direct Search with sufficient decrease

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

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

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

Algorithm 1: Direct Search with sufficient decrease

Input: $x_0 \in \mathbb{R}^n$, $\alpha_0 \in (0, \infty)$, $0 < \theta < 1 \leq \gamma$.
for $k = 0, 1, \ldots$ **do**
  Select $D_k \subset \mathbb{R}^n$.
  if $f(x_k + \alpha_k d) < f(x_k) - \alpha_k^2$ for some $d \in \mathcal{D}_k$ **then**

Algorithm 1: Direct Search with sufficient decrease

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

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

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

if $f(x_k + \alpha_k d) < f(x_k) - \alpha_k^2$ for some $d \in \mathcal{D}_k$ **then**

Expand step size, and move to that point

Algorithm 1: Direct Search with sufficient decrease

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

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

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

  if $f(x_k + \alpha_k d) < f(x_k) - \alpha_k^2$ for some $d \in \mathcal{D}_k$ **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

Algorithm 1: Direct Search with sufficient decrease

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

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

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

  if $f(x_k + \alpha_k d) < f(x_k) - \alpha_k^2$ for some $d \in \mathcal{D}_k$ **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

Algorithm 1: Direct Search with sufficient decrease

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

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

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

  if $f(x_k + \alpha_k d) < f(x_k) - \alpha_k^2$ for some $d \in \mathcal{D}_k$ **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 $x_{k+1} = x_k$.

Algorithm 1: Direct Search with sufficient decrease

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

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

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

  if $f(x_k + \alpha_k d) < f(x_k) - \alpha_k^2$ for some $d \in \mathcal{D}_k$ **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 $x_{k+1} = x_k$.

Everything almost clear, except "select D_k".

"Good" direction set

Question 1: what is a "good" D_k?

“Good” direction set

Question 1: what is a “good” D_k?

Include at least one descent direction, i.e. exists a $d \in D_k$ s.t., $-g_k^\top d > 0$.

“Good” direction set

Question 1: what is a “good” D_k?

Include at least one descent direction, i.e. exists a $d \in D_k$ s.t., $-g_k^\top d > 0$.

Question 2: how to choose a “good” D_k when we don’t know g_k?

“Good” direction set

Question 1: what is a “good” D_k?

Include at least one descent direction, i.e. exists a $d \in D_k$ s.t., $-g_k^\top d > 0$.

Question 2: how to choose a “good” D_k when we don’t know g_k?

Natural idea: choose a D_k s.t., for any $v \in \mathbb{R}^n$, there exists at least a $d \in D_k$ satisfying $d^\top v > 0$.

Question 1: what is a “good” D_k?

Include at least one descent direction, i.e. exists a $d \in D_k$ s.t., $-g_k^\top d > 0$.

Question 2: how to choose a “good” D_k when we don’t know g_k?

Natural idea: choose a D_k s.t., for any $v \in \mathbb{R}^n$, there exists at least a $d \in D_k$ satisfying $d^\top v > 0$.

Positive Spanning Set (PSS)

Question 1: what is a “good” D_k?

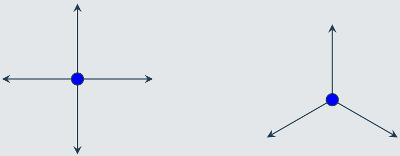
Include at least one descent direction, i.e. exists a $d \in D_k$ s.t., $-g_k^\top d > 0$.

Question 2: how to choose a “good” D_k when we don't know g_k?

Natural idea: choose a D_k s.t., for any $v \in \mathbb{R}^n$, there exists at least a $d \in D_k$ satisfying $d^\top v > 0$.

Positive Spanning Set (PSS)

Example (typical PSS in $\mathbb{R}^2$)



"Good" direction set

Question 3: how "good" is a PSS?

“Good” direction set

Question 3: how “good” is a PSS?

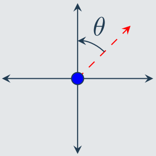
Cosine measure

Define the cosine measure, $\mathrm{cm}(D)$, for a finite set of nonzero vectors, $D \in \mathbb{R}^n$, as

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

Example

$$\mathrm{cm}(D) = \cos\theta = \frac{\sqrt{2}}{2}$$



Theorem

Under standard assumptions, if in Algorithm 1, there exists a positive constant κ such that

$$\mathrm{cm}(D_k) \;\geq\; \kappa \qquad \textit{for all } k \geq 0,$$

then we have

$$\liminf_{k\to\infty} \|g_k\| \;=\; 0.$$

Theorem

Under standard assumptions, if in Algorithm 1, there exists a positive constant κ such that

$$\mathrm{cm}(D_k) \;\geq\; \kappa \qquad \textit{for all } k \geq 0,$$

then we have

$$\liminf_{k \to \infty} \|g_k\| \;=\; 0.$$

Each direction set D_k should be a “good enough” PSS.

Randomized method

Bad news: if D is a PSS in $\mathbb{R}^n$, then $|D| \geq n+1$.
Complexity: $\mathcal{O}(n^2\varepsilon^{-2})$

Randomized method

Bad news: if D is a PSS in $\mathbb{R}^n$, then $|D| \geq n + 1$.
Complexity: $\mathcal{O}(n^2 \varepsilon^{-2})$

Goal: use randomized methods to reduce $|D|$ from $\mathcal{O}(n)$ to $\mathcal{O}(1)$.

Bad news: if D is a PSS in $\mathbb{R}^n$, then $|D| \geq n+1$.
Complexity: $\mathcal{O}(n^2\varepsilon^{-2})$

Goal: use randomized methods to reduce $|D|$ from $\mathcal{O}(n)$ to $\mathcal{O}(1)$.

Algorithm 2: Probabilistic Direct Search

Input: $x_0 \in \mathbb{R}^n$, $\alpha_0 \in (0,\infty)$, $0 < \theta < 1 \leq \gamma$.
for $k = 0, 1, \ldots$ **do**
- Randomly generate $\mathfrak{D}_k \subset \mathbb{R}^n$.
- **if** $f(x_k + \alpha_k d) < f(x_k) - \rho(\alpha_k)$ for some $d \in \mathfrak{D}_k$ **then**
  - Set $\alpha_{k+1} = \gamma\alpha_k$ and $x_{k+1} = x_k + \alpha_k d$.
- **else**
  - Set $\alpha_{k+1} = \theta\alpha_k$ and $x_{k+1} = x_k$.

Bad news: if D is a PSS in $\mathbb{R}^n$, then $|D| \geq n+1$.
Complexity: $\mathcal{O}(n^2\varepsilon^{-2})$

Goal: use randomized methods to reduce $|D|$ from $\mathcal{O}(n)$ to $\mathcal{O}(1)$.

Algorithm 2: Probabilistic Direct Search

Input: $x_0 \in \mathbb{R}^n$, $\alpha_0 \in (0, \infty)$, $0 < \theta < 1 \leq \gamma$.
for $k = 0, 1, \ldots$ **do**
  Randomly generate $\mathfrak{D}_k \subset \mathbb{R}^n$.
  if $f(x_k + \alpha_k d) < f(x_k) - \rho(\alpha_k)$ for some $d \in \mathfrak{D}_k$ **then**
    Set $\alpha_{k+1} = \gamma\alpha_k$ and $x_{k+1} = x_k + \alpha_k d$.
  else
    Set $\alpha_{k+1} = \theta\alpha_k$ and $x_{k+1} = x_k$.

All the randomness comes from $\mathfrak{D}_k$

Bad news: if D is a PSS in $\mathbb{R}^n$, then $|D| \geq n+1$.
Complexity: $\mathcal{O}(n^2\varepsilon^{-2})$

Goal: use randomized methods to reduce $|D|$ from $\mathcal{O}(n)$ to $\mathcal{O}(1)$.

Algorithm 2: Probabilistic Direct Search

Input: $x_0 \in \mathbb{R}^n$, $\alpha_0 \in (0, \infty)$, $0 < \theta < 1 \leq \gamma$.
for $k = 0, 1, \dots$ **do**
  Randomly generate $\mathfrak{D}_k \subset \mathbb{R}^n$.
  if $f(x_k + \alpha_k d) < f(x_k) - \rho(\alpha_k)$ for some $d \in \mathfrak{D}_k$ **then**
    Set $\alpha_{k+1} = \gamma\alpha_k$ and $x_{k+1} = x_k + \alpha_k d$.
  else
    Set $\alpha_{k+1} = \theta\alpha_k$ and $x_{k+1} = x_k$.

All the randomness comes from $\mathfrak{D}_k$

Notations for random variables or random vectors

$D_k \Rightarrow \mathfrak{D}_k, \quad d \Rightarrow \mathfrak{d}, \quad x_k \Rightarrow X_k, \quad \alpha_k \Rightarrow A_k, \quad g_k \Rightarrow G_k$

Actually, what we need is not $\mathrm{cm}(D_k) \geq \kappa$, but $\mathrm{cm}(D_k, -G_k) \geq \kappa$.

Actually, what we need is not $\mathrm{cm}(D_k) \geq \kappa$, but $\mathrm{cm}(D_k, -G_k) \geq \kappa$.

Definition (p-probabilistically κ-descent)

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

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

Intuition: high probability that $\mathfrak{D}_k$ is “always good enough”, no matter what happened before.

Theorem (Gratton et al. 2015)

Under standard assumptions, if in Algorithm 2, there exists a positive constant κ such that $(\mathfrak{D}_k)_{k\geq 0}$ is p-probabilistically κ-descent with $p = \log\theta / \log(\gamma^{-1}\theta)$, then we have

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

Complexity: $\mathcal{O}(mn\varepsilon^{-2})$ with extremely high probability, where $m \ll n$.

Table of contents

Why and how of “non-convergence”: ideas and main results

Recall "Converge or Diverge? A Story of Adam" by Prof. Luo

Recall "Converge or Diverge? A Story of Adam" by Prof. Luo

What will happen if $(\mathfrak{D}_k)_{k\geq 0}$ fails to meet p-probabilistically κ-descent? Or worse $\mathfrak{D}_k$ is always "bad"?

Can we obtain some kind of "non-convergence"?

“Bad” direction set

p-probabilistically ascent

$(\mathfrak{D}_k)_{k\in\mathbb{N}}$ is said to be p-probabilistically ascent if it satisfies

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

Observation

For any convex function, if $\mathrm{cm}\left(D_k, -g_k\right) \leq 0$, then no descent direction in D_k, leading to shrinking of step size.

"Bad" direction set

p-probabilistically ascent

$(\mathfrak{D}_k)_{k\in\mathbb{N}}$ is said to be p-probabilistically ascent if it satisfies

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

Observation

For any convex function, if $\mathrm{cm}\left(D_k, -g_k\right) \leq 0$, then no descent direction in D_k, leading to shrinking of step size.

If we show non-convergence for convex functions, then we cannot expect convergence for others.

Naive idea of non-convergence

$\mathfrak{D}_k$ always bad

Naive idea of non-convergence

$\mathfrak{D}_k$ always bad

$\Downarrow$

A_k always shrinks

Naive idea of non-convergence

$\mathfrak{D}_k$ always bad

$\Downarrow$

A_k always shrinks

$\Downarrow$

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

Naive idea of non-convergence

$\mathfrak{D}_k$ always bad

$\Downarrow$

A_k always shrinks

$\Downarrow$

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

$\Downarrow$

$$x_o \quad \notin \quad \bar{\mathcal{B}}(x^*, \textstyle\sum_{k=0}^{\infty} A_k)$$

Naive idea of non-convergence

$\mathfrak{D}_k$ always bad

$\Downarrow$

A_k always shrinks

$\Downarrow$

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

$\Downarrow$

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

$\Downarrow$

Non-convergence:

$$\mathbb{P}\left(\liminf_k \operatorname{dist}\left(X_k, x^*\right) = 0\right) < 1$$

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

- Define indicator function $Y_k = \mathbb{1}_{\{\mathrm{cm}(\mathfrak{D}_k, -G_k) > 0\}}$
  Indicator for "good" event
- $A_{k+1} \leq \gamma^{Y_k} \theta^{1-Y_k} A_k$, when f is convex

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

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

Assumption

$$\mathbb{P}\left(Y_k = 0 \mid Y_0, \ldots, Y_{k-1}\right) \;\geq\; p \;>\; p_0 \quad \text{for each } k \geq 0,$$

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

Quantitative analysis

Assumption

$$\mathbb{P}\left(Y_k = 0 \mid Y_0, \ldots, Y_{k-1}\right) \;\geq\; p \;>\; p_0 \quad \text{for each } k \geq 0,$$

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

Result

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

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

Assumption

$$\mathbb{P}\left(Y_k = 0 \mid Y_0, \ldots, Y_{k-1}\right) \;\geq\; p \;>\; p_0 \quad \text{for each } k \geq 0,$$

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

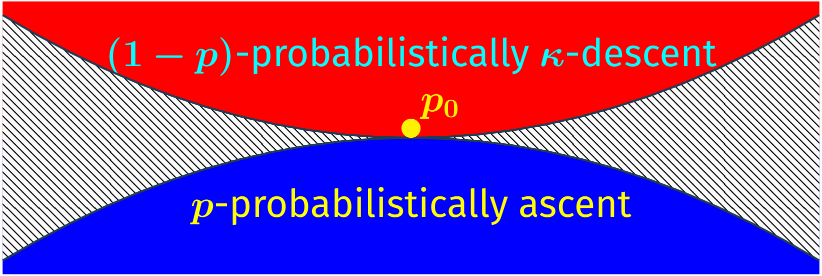
Result

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

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

Observation

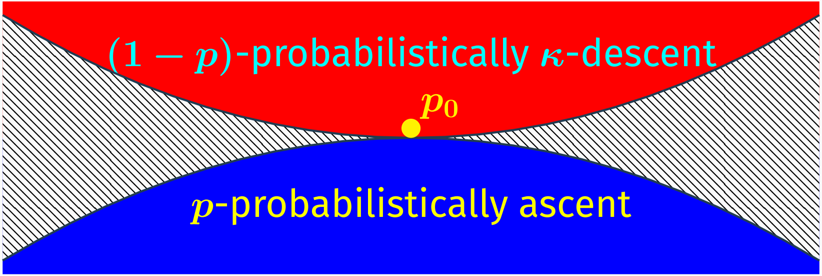
- Easy to achieve non-convergence: $\theta/(1-\theta)$ is fixed number.
- Cannot further improve: $\sum_{k=1}^{\infty} U_k \geq \theta/(1-\theta)$.



Space of $(\mathfrak{D}_k)_{k\geq 0}$

$\mathfrak{D}_k$ is measurable map from Ω to $2^{\mathbb{R}^n}$

Red: Convergence Blue: Non-convergence



Space of $(\mathfrak{D}_k)_{k\geq 0}$

$\mathfrak{D}_k$ is measurable map from Ω to $2^{\mathbb{R}^n}$

Red: Convergence   Blue: Non-convergence

Too much “shadow area”, not satisfied!

Question: what if we only need $\sum_{k=1}^{\infty} U_k < \infty$?

Question: what if we only need $\sum_{k=1}^{\infty} U_k < \infty$?
Only need $\mathfrak{D}_k$ to be “bad” eventually.

Assumption

$$\liminf_{k\to\infty} \mathbb{P}\left(Y_k = 0 \mid Y_0, \ldots, Y_{k-1}\right) \;>\; p_0.$$

Question: what if we only need $\sum_{k=1}^{\infty} U_k < \infty$?
Only need $\mathfrak{D}_k$ to be "bad" eventually.

Assumption

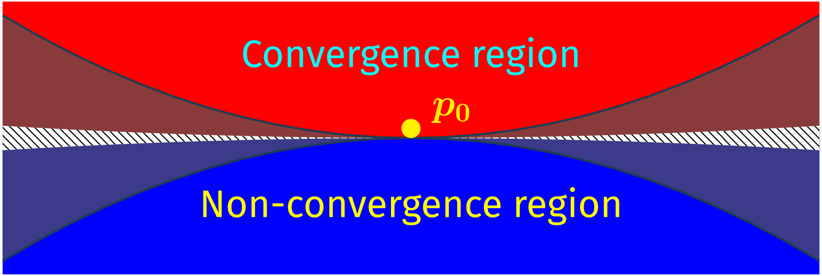
$$\liminf_{k\to\infty} \mathbb{P}\left(Y_k = 0 \mid Y_0, \ldots, Y_{k-1}\right) \; > \; p_0.$$

Result

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

Also use “liminf”-type assumption in convergence theory

Also use “liminf”-type assumption in convergence theory



Tightness of non-convergence: counterexample

Natural question:
can p_0 be included in the non-convergence region?

Can p_0 be included?

Natural question:

can p_0 be included in the non-convergence region?

Answer: NO

Can p_0 be included?

Natural question:
can p_0 be included in the non-convergence region?
Answer: NO

Example

We assume

- $f : \mathbb{R}^n \to \mathbb{R}$ be gradient Lipschitz and strongly convex,
- $\theta = 1/2$ and $\gamma = 2$,
- $\mathbb{P}(\mathfrak{d}_k = G_k/\|G_k\|) = \mathbb{P}(\mathfrak{d}_k = -G_k/\|G_k\|) = 1/2$,

then we have

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

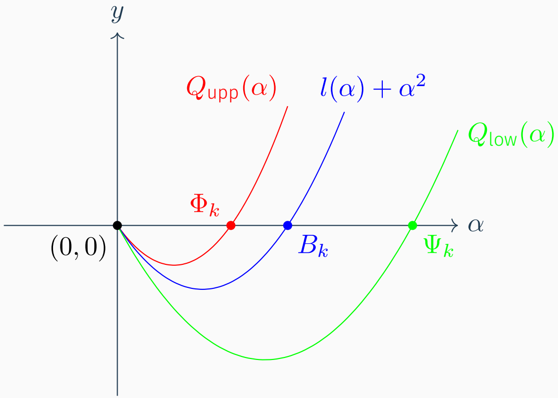
- $l(\alpha) = f(X_k - \alpha G_k / \|G_k\|) - f(X_k)$

Sketch of proof

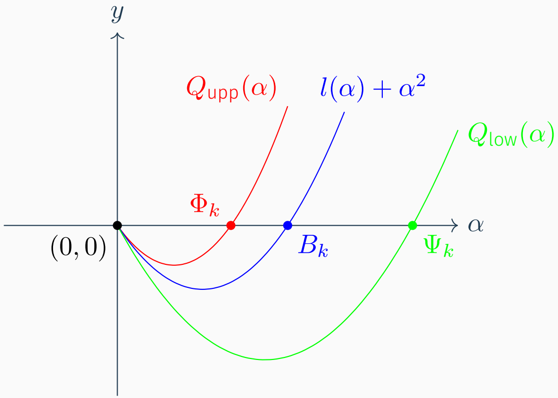
- $l(\alpha) = f(X_k - \alpha G_k / \|G_k\|) - f(X_k)$
- Sufficient decrease condition becomes $l(\alpha) + \alpha^2 < 0$

Sketch of proof

- $l(\alpha) = f(X_k - \alpha G_k/\|G_k\|) - f(X_k)$
- Sufficient decrease condition becomes $l(\alpha) + \alpha^2 < 0$
- $l(\alpha) + \alpha^2$ is both upper and lower bounded by quadratic functions

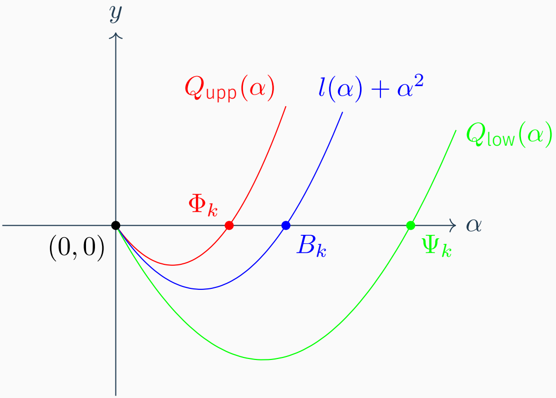

y
$Q_{\mathsf{upp}}(\alpha)$
$l(\alpha) + \alpha^2$
$Q_{\mathsf{low}}(\alpha)$
Φ_k
α
$(0, 0)$
B_k
Ψ_k

Sketch of proof

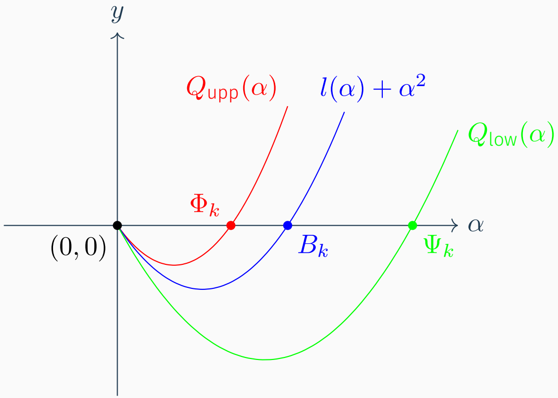


- $C_1\|G_k\| = \Phi_k \le B_k \le \Psi_k = C_2\|G_k\|$

Sketch of proof



- $C_1\|G_k\| = \Phi_k \leq B_k \leq \Psi_k = C_2\|G_k\|$
- Just to prove $\mathbb{P}(\liminf_k B_k = 0) = 1$



- $C_1\|G_k\| = \Phi_k \le B_k \le \Psi_k = C_2\|G_k\|$
- Just to prove $\mathbb{P}(\liminf_k B_k = 0) = 1$
- Known that $A_k(\omega) \to 0 \; \forall \omega \in \Omega$, just to prove $\mathbb{P}(A_k \ge B_k \text{ i.o.}) = 1$

Let $S_k = \log_2(A_k/\alpha_0)$ and $R_k = \log_2(B_k/\alpha_0)$. Then we abstract the following problem.

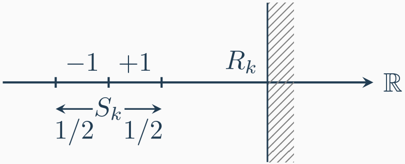
Let $S_k = \log_2(A_k/\alpha_0)$ and $R_k = \log_2(B_k/\alpha_0)$. Then we abstract the following problem.

Assume that $(S_k)_{k\geq 0}$ and $(R_k)_{k\geq 0}$ are two stochastic processes such that

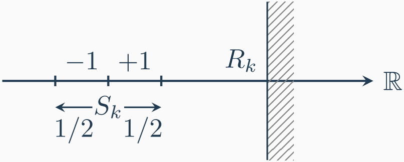
- $S_0 = 0$ and $(R_k)_{k\geq 0}$ uniformly bounded,
- $$\begin{aligned}\mathbb{P}\left(S_{k+1} = S_k - 1 \mid S_1, \ldots, S_k\right) &= \frac{1}{2}\mathbb{1}_{\{S_k < R_k\}} + \mathbb{1}_{\{S_k \geq R_k\}},\\ \mathbb{P}\left(S_{k+1} = S_k + 1 \mid S_1, \ldots, S_k\right) &= \frac{1}{2}\mathbb{1}_{\{S_k < R_k\}}.\end{aligned}$$

Then, $\mathbb{P}(S_k \geq R_k \text{ i.o.}) = 1$?

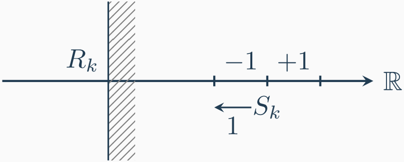
When $S_k < R_k$,



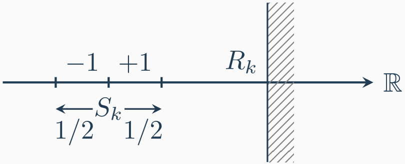
When $S_k < R_k$,



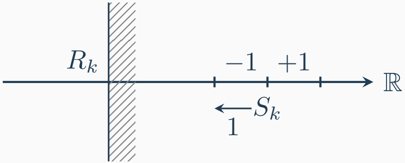
When $S_k \geq R_k$,



When $S_k < R_k$,



When $S_k \geq R_k$,



Does S_k go beyond the "wall" R_k i.o. with probability 1?

- Construct $\widetilde{S}_k = S_k + \sum_{i=0}^{k-1} T_i \mathbb{1}_{\{S_k \geq R_k\}}$,
  where $T_0, T_1, \ldots$ are i.i.d., $\mathbb{P}(T_i = 0) = \mathbb{P}(T_i = 2) = 1/2$

Sketch of proof

- Construct $\widetilde{S}_k = S_k + \sum_{i=0}^{k-1} T_i \mathbb{1}_{\{S_k \geq R_k\}}$,
  where $T_0, T_1, \ldots$ are i.i.d., $\mathbb{P}(T_i = 0) = \mathbb{P}(T_i = 2) = 1/2$
  Intuition: Add the probability of S_k going right in the case $S_k \geq R_k$

Sketch of proof

- Construct $\widetilde{S}_k = S_k + \sum_{i=0}^{k-1} T_i \mathbb{1}_{\{S_k \geq R_k\}}$,
  where $T_0, T_1, \ldots$ are i.i.d., $\mathbb{P}(T_i = 0) = \mathbb{P}(T_i = 2) = 1/2$
  Intuition: Add the probability of S_k going right in the case $S_k \geq R_k$
- $(\widetilde{S}_k)_{k \geq 0}$ is simple random walk $\Rightarrow \limsup_k \widetilde{S}_k = \infty$ w.p. 1

- Construct $\widetilde{S}_k = S_k + \sum_{i=0}^{k-1} T_i \mathbb{1}_{\{S_k \geq R_k\}}$,
  where $T_0, T_1, \dots$ are i.i.d., $\mathbb{P}(T_i = 0) = \mathbb{P}(T_i = 2) = 1/2$
  Intuition: Add the probability of S_k going right in the case $S_k \geq R_k$
- $(\widetilde{S}_k)_{k\geq 0}$ is simple random walk $\Rightarrow \limsup_k \widetilde{S}_k = \infty$ w.p. 1
- $(R_k)_{k\geq 0}$ uniformly bounded $\Rightarrow$ so does $(S_k)_{k\geq 0}$

Sketch of proof

- Construct $\widetilde{S}_k = S_k + \sum_{i=0}^{k-1} T_i \mathbb{1}_{\{S_k \geq R_k\}}$,
  where $T_0, T_1, \ldots$ are i.i.d., $\mathbb{P}(T_i = 0) = \mathbb{P}(T_i = 2) = 1/2$
  Intuition: Add the probability of S_k going right in the case $S_k \geq R_k$
- $(\widetilde{S}_k)_{k \geq 0}$ is simple random walk $\Rightarrow \limsup_k \widetilde{S}_k = \infty$ w.p. 1
- $(R_k)_{k \geq 0}$ uniformly bounded $\Rightarrow$ so does $(S_k)_{k \geq 0}$
- $\sum_{i=0}^{\infty} T_i \mathbb{1}_{\{S_k \geq R_k\}} = \infty$ w.p. 1

Conclusion

Conclusion

In this talk:

Conclusion

In this talk:

- establish non-convergence theory for probabilistic direct search,

Conclusion

In this talk:

- establish non-convergence theory for probabilistic direct search,
- distinguish convergence region and non-convergence region,

Conclusion

In this talk:

- establish non-convergence theory for probabilistic direct search,
- distinguish convergence region and non-convergence region,
- construct one counterexample to show the tightness of non-convergence region boundary.

Future work:

- find estimation or lower bound for the probability of non-convergence,
- establish non-convergence theory for other models.

Thank you!

- 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.
- 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*.
- Gratton, S. et al. (2015). “Direct search based on probabilistic descent”. *SIAM J. Optim.* 25, pp. 1515–1541.