Global Rates for Zero-Order Methods

Luis Nunes Vicente
University of Coimbra

July 25, 2013 — RTRA STAE Conférence, CERFACS

http//www.mat.uc.pt/~lnv

Analysis of algorithms for Nonlinear Optimization

- Global convergence: convergence to some form of stationarity independently of the starting point.

Analysis of algorithms for Nonlinear Optimization

- Global convergence: convergence to some form of stationarity independently of the starting point.
- Global rates (worst case complexity): no assumption on the starting point, amount of work needed to reach some threshold of stationarity.

Analysis of algorithms for Nonlinear Optimization

- Global convergence: convergence to some form of stationarity independently of the starting point.
- Global rates (worst case complexity): no assumption on the starting point, amount of work needed to reach some threshold of stationarity.
- Local rates of convergence: rates of convergence, like superlinear or quadratic, in a neighborhood of a minimizer.

Analysis of algorithms for Nonlinear Optimization

- Global convergence: convergence to some form of stationarity independently of the starting point.
- Global rates (worst case complexity): no assumption on the starting point, amount of work needed to reach some threshold of stationarity.
- Local rates of convergence: rates of convergence, like superlinear or quadratic, in a neighborhood of a minimizer.

My talk will focus on: Derivative-Free Optimization (DFO), zero-order methods.

There are two main classes of rigorous methods in DFO

- Directional methods, like direct search.

There are two main classes of rigorous methods in DFO

- Directional methods, like direct search.

- Model-based methods, like trust-region methods.

Direct-search methods

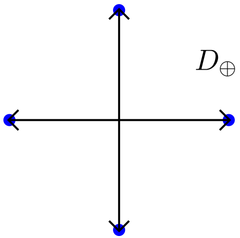
Definition

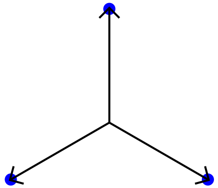
- *Sample the objective function at a finite number of points at each iteration.*
- *Achieve descent by moving in directions of potential descent.*
- *In the smooth case, these directions lie in positive spanning sets (PSS):*

Direct-search methods

Definition

- *Sample the objective function at a finite number of points at each iteration.*
- *Achieve descent by moving in directions of potential descent.*
- *In the smooth case, these directions lie in positive spanning sets (PSS):*

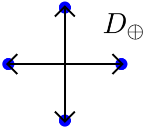


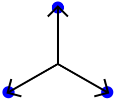


Direct-search methods

Definition

- *Sample the objective function at a finite number of points at each iteration.*
- *Achieve descent by moving in directions of potential descent.*
- *In the smooth case, these directions lie in positive spanning sets (PSS):*





Our problem setting

Unconstrained optimization

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

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

f is at least locally Lipschitz continuous

Our problem setting

Unconstrained optimization

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

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

f is at least locally Lipschitz continuous

A forcing function $\rho(\cdot)$ is a positive and monotonically nondecreasing function such that

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

Our problem setting

Unconstrained optimization

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

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

f is at least locally Lipschitz continuous

A forcing function $\rho(\cdot)$ is a positive and monotonically nondecreasing function such that

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

We will consider $\rho(\alpha) = \alpha^p$, with $p > 1$.

Our problem setting

Unconstrained optimization

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

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

f is at least locally Lipschitz continuous

A forcing function $\rho(\cdot)$ is a positive and monotonically nondecreasing function such that

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

We will consider $\rho(\alpha) = \alpha^p$, with $p > 1$.

In most of the talk, we take $p = 2$: $\rho(\alpha) = \alpha^2$.

A class of direct-search methods (f smooth, for now)

Choose: x_0 and α_0.

A class of direct-search methods (f smooth, for now)

Choose: x_0 and α_0.

For $k = 0, 1, 2, \ldots$ (Until α_k is suff. small)

- **Search step (optional)**

A class of direct-search methods (f smooth, for now)

Choose: x_0 and α_0.

For $k = 0, 1, 2, \ldots$ (Until α_k is suff. small)

- **Search step (optional)**
- **Poll step:** Select D_k PSS and find $x_k + \alpha_k d_k$ ($d_k \in D_k$):

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

A class of direct-search methods (f smooth, for now)

Choose: x_0 and α_0.

For $k = 0, 1, 2, \ldots$ (Until α_k is suff. small)

- **Search step (optional)**
- **Poll step:** Select D_k PSS and find $x_k + \alpha_k d_k$ ($d_k \in D_k$):

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

- Update the new iterate x_{k+1} (stay at x_k is unsuccessful).

A class of direct-search methods (f smooth, for now)

Choose: x_0 and α_0.

For $k = 0, 1, 2, \ldots$ (Until α_k is suff. small)

- **Search step (optional)**
- **Poll step:** Select D_k PSS and find $x_k + \alpha_k d_k$ ($d_k \in D_k$):

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

- Update the new iterate x_{k+1} (stay at x_k is unsuccessful).
- Update the step size α_{k+1}.
  Possible increase if iteration is successful. Decrease otherwise.

Behavior of the step size parameter

Assumption

The level set $L(x_0) = \{x \in \mathbb{R}^n : \; f(x) \leq f(x_0)\}$ *is bounded.*

Behavior of the step size parameter

Assumption

The level set $L(x_0) = \{x \in \mathbb{R}^n : \; f(x) \leq f(x_0)\}$ *is bounded.*

Lemma (IDFO book or SIAM Review 2003 survey on DS)

There exists a point x_* *and a subsequence* K *of unsuccessful iterations:*

$$\lim_{k \in K} x_k = x_* \quad \textit{and} \quad \lim_{k \in K} \alpha_k = 0.$$

Behavior of the step size parameter

Assumption

The level set $L(x_0) = \{x \in \mathbb{R}^n : \; f(x) \leq f(x_0)\}$ *is bounded.*

Lemma (IDFO book or SIAM Review 2003 survey on DS)

There exists a point x_* *and a subsequence* K *of unsuccessful iterations:*

$$\lim_{k \in K} x_k = x_* \quad \textit{and} \quad \lim_{k \in K} \alpha_k = 0.$$

Assumption

The directions in D_k *are bounded above and away from zero.*

The cosine measure of D_k *is bounded away from zero.*

Behavior of unsuccessful iterations

Theorem (Lewis, Tolda, and Torczon 2003)

Let D_k be a PSS.

Assume $f \in \mathcal{C}^1_\nu$.

If the iterate k is unsuccessful, i.e.,

$$f(x_k + \alpha_k d) \geq f(x_k) - \rho(\alpha_k), \qquad \text{for all } d \in D_k,$$

then

$$\|\nabla f(x_k)\| \leq \frac{C(\nu) \times \alpha_k}{\mathrm{cm}(D_k)} \qquad \text{... since } \rho(\alpha) = \alpha^2.$$

Behavior of unsuccessful iterations

Theorem (Lewis, Tolda, and Torczon 2003)

Let D_k be a PSS.

Assume $f \in \mathcal{C}^1_\nu$.

If the iterate k is unsuccessful, i.e.,

$$f(x_k + \alpha_k d) \geq f(x_k) - \rho(\alpha_k), \qquad \text{for all } d \in D_k,$$

then

$$\|\nabla f(x_k)\| \;\leq\; \frac{C(\nu) \times \alpha_k}{\mathrm{cm}(D_k)} \qquad \text{... since } \rho(\alpha) = \alpha^2.$$

Note that global convergence is deduced from here: $\|\nabla f(x_k)\| \underset{K}{\longrightarrow} 0$.

The question that interests us (smooth case)

Question

Given $\epsilon \in (0,1)$, *how many iterations* $\bar{k}$ *are needed to reach*

$$\|\nabla f(x_{\bar{k}})\| \leq \epsilon \quad ?$$

WCC of direct search (smooth case)

Let k_0 be the index of the first unsuccessful iteration.

WCC of direct search (smooth case)

Let k_0 be the index of the first unsuccessful iteration.

Let $\bar{k}$ be the first index after k_0 such that $\|\nabla f(x_{\bar{k}})\| \leq \epsilon$.

WCC of direct search (smooth case)

Let k_0 be the index of the first unsuccessful iteration.

Let $\bar{k}$ be the first index after k_0 such that $\|\nabla f(x_{\bar{k}})\| \leq \epsilon$.

For an unsuccessful iteration $k < \bar{k}$,

$$\epsilon \; < \; \|\nabla f(x_k)\| \; \leq \; C(\nu)\alpha_k.$$

WCC of direct search (smooth case)

Let k_0 be the index of the first unsuccessful iteration.

Let $\bar{k}$ be the first index after k_0 such that $\|\nabla f(x_{\bar{k}})\| \leq \epsilon$.

For an unsuccessful iteration $k < \bar{k}$,

$$\epsilon \; < \; \|\nabla f(x_k)\| \; \leq \; C(\nu)\alpha_k.$$

One can backtrack from any successful to the previous unsuccessful one using

$$f(x_k) - f(x_{k+1}) \; \geq \; \alpha_k^2$$

WCC of direct search (smooth case)

Let k_0 be the index of the first unsuccessful iteration.

Let $\bar{k}$ be the first index after k_0 such that $\|\nabla f(x_{\bar{k}})\| \leq \epsilon$.

For an unsuccessful iteration $k < \bar{k}$,

$$\epsilon \ < \ \|\nabla f(x_k)\| \ \leq \ C(\nu)\alpha_k.$$

One can backtrack from any successful to the previous unsuccessful one using

$$f(x_k) - f(x_{k+1}) \ \geq \ \alpha_k^2 \ \geq \ \frac{1}{C(\nu^2)}\epsilon^2.$$

WCC of direct search (smooth case)

Theorem

The number of successful iterations between k_0 and $\bar{k}$ is

$$\mathcal{S}(k_0, \bar{k}) \;\leq\; \left\lceil C(\nu^2)(f(x_{k_0}) - f_*)\,\frac{1}{\epsilon^2} \right\rceil .$$

WCC of direct search (smooth case)

Theorem

The number of successful iterations between k_0 and $\bar{k}$ is

$$\mathcal{S}(k_0, \bar{k}) \;\leq\; \left\lceil C(\nu^2)(f(x_{k_0}) - f_*)\,\frac{1}{\epsilon^2} \right\rceil.$$

Since $\alpha_{k+1} \leq C\alpha_k$ ($C \geq 1$ or $C < 1$), one obtains by induction

WCC of direct search (smooth case)

Theorem

The number of successful iterations between k_0 and $\bar{k}$ is

$$\mathcal{S}(k_0, \bar{k}) \;\leq\; \left\lceil C(\nu^2)(f(x_{k_0}) - f_*)\,\frac{1}{\epsilon^2} \right\rceil.$$

Since $\alpha_{k+1} \leq C\alpha_k$ ($C \geq 1$ or $C < 1$), one obtains by induction

Theorem

The number of successful iterations between k_0 and $\bar{k}$ is

$$\mathcal{U}(k_0, \bar{k}) \;=\; \mathcal{O}(\mathcal{S}(k_0, \bar{k})).$$

WCC of direct search (smooth case)

Theorem

The number of successful iterations between k_0 and $\bar{k}$ is

$$\mathcal{S}(k_0, \bar{k}) \;\leq\; \left\lceil C(\nu^2)(f(x_{k_0}) - f_*)\,\frac{1}{\epsilon^2} \right\rceil.$$

Since $\alpha_{k+1} \leq C\alpha_k$ ($C \geq 1$ or $C < 1$), one obtains by induction

Theorem

The number of successful iterations between k_0 and $\bar{k}$ is

$$\mathcal{U}(k_0, \bar{k}) \;=\; \mathcal{O}(\mathcal{S}(k_0, \bar{k})).$$

The number of suc. iterations until k_0 is at most

$$\left\lceil \frac{f(x_0) - f_*}{\alpha_0^2} \right\rceil.$$

WCC of direct search (smooth case)

Theorem

Any direct-search method (based on sufficient decrease) takes at most

$$\mathcal{O}\left(n\,\nu^2\,\epsilon^{-2}\right)$$

iterations to reduce the gradient below $\epsilon \in (0,1)$.

WCC of direct search (smooth case)

Theorem

Any direct-search method (based on sufficient decrease) takes at most

$$\mathcal{O}\left(n\,\nu^2\,\epsilon^{-2}\right)$$

iterations to reduce the gradient below $\epsilon \in (0,1)$.

- The number of function evaluations must be multiplied by n:

$$\mathcal{O}\left(n^2\nu^2\,\epsilon^{-2}\right).$$

WCC of direct search (smooth case)

Theorem

Any direct-search method (based on sufficient decrease) takes at most

$$\mathcal{O}\left(n\,\nu^2\,\epsilon^{-2}\right)$$

iterations to reduce the gradient below $\epsilon \in (0,1)$.

- The number of function evaluations must be multiplied by n:

$$\mathcal{O}\left(n^2\nu^2\,\epsilon^{-2}\right).$$

- L. N. Vicente, Worst case complexity of direct search, to appear in EURO J. on Computational Optimization, Vol. 1, Num. 1, 2013.

Assumption in the smooth, convex case

Assumption

There exists a positive constant R *such that*

$$\sup_{k \in \mathcal{U}} \operatorname{dist}(x_k, X_*^f) \leq R$$

where $X_*^f = \{x \in \mathbb{R}^n : \ x$ *is a minimizer of* $f\}$.

Assumption in the smooth, convex case

Assumption

There exists a positive constant R *such that*

$$\sup_{k\in\mathcal{U}} \operatorname{dist}(x_k, X_*^f) \;\leq\; R$$

where $X_*^f \;=\; \{x \in \mathbb{R}^n :\; x$ *is a minimizer of* $f\}$.

One needs this assumption because

$$\|x_k - x_*\| \;\leq\; \|x_0 - x_*\|, \qquad \forall k \geq 0,$$

does NOT hold as in the gradient method (because $d_k \neq -\nabla f(x_k)$).

WCC of direct search (smooth, convex case)

Lemma (Decrease rate for f)

Any direct-search method (based on sufficient decrease) generates a sequence $\{x_k\}_{k \geq k_0}$ *such that*

$$f(x_k) - f_* \ < \ \frac{C(\nu^2) R^2}{k - k_0 - m - 1}$$

WCC of direct search (smooth, convex case)

Lemma (Decrease rate for f)

Any direct-search method (based on sufficient decrease) generates a sequence $\{x_k\}_{k \geq k_0}$ such that

$$f(x_k) - f_* \; < \; \frac{C(\nu^2) R^2}{k - k_0 - m - 1}$$

where $m = m(k_0, k)$ is the # of unsucc. iter. between k_0 (the first unsucc. iter.) and k.

WCC of direct search (smooth, convex case)

Lemma (Decrease rate for f)

Any direct-search method (based on sufficient decrease) generates a sequence $\{x_k\}_{k \geq k_0}$ such that

$$f(x_k) - f_* \; < \; \frac{C(\nu^2) R^2}{k - k_0 - m - 1}$$

where $m = m(k_0, k)$ is the $\#$ of unsucc. iter. between k_0 (the first unsucc. iter.) and k.

Omitting constants, as in the gradient method,

$$\frac{C(\nu^2) R^2}{2k - \cdots}$$

WCC of direct search (smooth, convex case)

Lemma (Decrease rate for f)

Any direct-search method (based on sufficient decrease) generates a sequence $\{x_k\}_{k \geq k_0}$ such that

$$f(x_k) - f_* \; < \; \frac{C(\nu^2) R^2}{k - k_0 - m - 1}$$

where $m = m(k_0, k)$ is the # of unsucc. iter. between k_0 (the first unsucc. iter.) and k.

Omitting constants, as in the gradient method,

$$\frac{C(\nu^2) R^2}{2k - \cdots} \; \geq \; \sum_{l=k+1}^{2k} f(x_l) - f(x_{l+1})$$

WCC of direct search (smooth, convex case)

Lemma (Decrease rate for f)

Any direct-search method (based on sufficient decrease) generates a sequence $\{x_k\}_{k \geq k_0}$ such that

$$f(x_k) - f_* \; < \; \frac{C(\nu^2) R^2}{k - k_0 - m - 1}$$

where $m = m(k_0, k)$ is the # of unsucc. iter. between k_0 (the first unsucc. iter.) and k.

Omitting constants, as in the gradient method,

$$\frac{C(\nu^2)R^2}{2k - \cdots} \; \geq \; \sum_{l=k+1}^{2k} f(x_l) - f(x_{l+1}) \; \geq \; \sum_{l=k+1}^{2k} \alpha_l^2$$

WCC of direct search (smooth, convex case)

Lemma (Decrease rate for f)

Any direct-search method (based on sufficient decrease) generates a sequence $\{x_k\}_{k \geq k_0}$ such that

$$f(x_k) - f_* \; < \; \frac{C(\nu^2) R^2}{k - k_0 - m - 1}$$

where $m = m(k_0, k)$ is the # of unsucc. iter. between k_0 (the first unsucc. iter.) and k.

Omitting constants, as in the gradient method,

$$\frac{C(\nu^2)R^2}{2k - \cdots} \; \geq \; \sum_{l=k+1}^{2k} f(x_l) - f(x_{l+1}) \; \geq \; \sum_{l=k+1}^{2k} \alpha_l^2 \; \geq \; k \times \epsilon^2.$$

WCC of DS (smooth, convex case)

Theorem

Any direct-search method (based on sufficient decrease) takes at most

$$\mathcal{O}\left(n\,\nu^2 R\,\epsilon^{-1}\right)$$

iterations to reduce the gradient below $\epsilon \in (0,1)$.

WCC of DS (smooth, convex case)

Theorem

Any direct-search method (based on sufficient decrease) takes at most

$$\mathcal{O}\left(n\,\nu^2 R\,\epsilon^{-1}\right)$$

iterations to reduce the gradient below $\epsilon \in (0,1)$.

- The number of function evaluations must be multiplied by n:

$$\mathcal{O}\left(n^2\nu^2 R\,\epsilon^{-1}\right)$$

WCC of DS (smooth, convex case)

Theorem

Any direct-search method (based on sufficient decrease) takes at most

$$\mathcal{O}\left(n\,\nu^2 R\,\epsilon^{-1}\right)$$

iterations to reduce the gradient below $\epsilon \in (0,1)$.

- The number of function evaluations must be multiplied by n:

$$\mathcal{O}\left(n^2\nu^2 R\,\epsilon^{-1}\right)$$

Reference:

- M. Dodangeh and L. N. Vicente, Worst case complexity of direct search under convexity, 2013.

How to bound R

Proposition

Let f be continuous and strongly convex with constant μ. Then

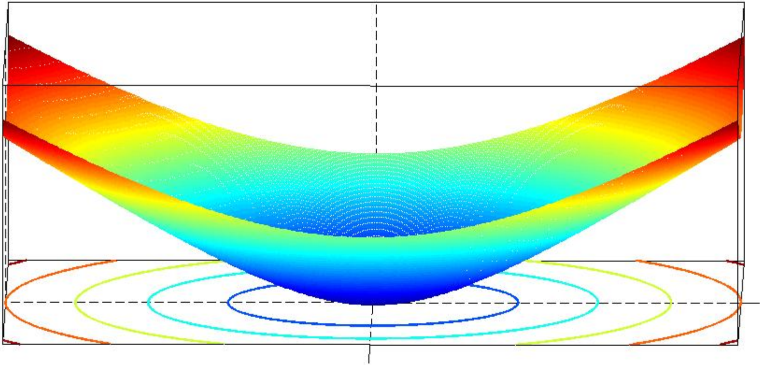
$$\sup_{y \in L(x_0)} \operatorname{dist}(y, X_*^f) \leq \sqrt{\frac{2}{\mu}(f(x_0) - f_*)}.$$

How to bound R

Proposition

Let f be continuous and strongly convex with constant μ. Then

$$\sup_{y \in L(x_0)} \operatorname{dist}(y, X_*^f) \leq \sqrt{\frac{2}{\mu}(f(x_0) - f_*)}.$$



An example where R is UNBOUNDED

Objective function

Let $\epsilon \in (0,\ \frac{1}{2})$ and f be strongly convex function and parameterized by ϵ

$$f(x, y) \ = \ y^2 + \frac{1}{2}(\epsilon x)^2 + \epsilon x.$$

An example where R is UNBOUNDED

Objective function

Let $\epsilon \in (0, \frac{1}{2})$ and f be strongly convex function and parameterized by ϵ

$$f(x, y) = y^2 + \frac{1}{2}(\epsilon x)^2 + \epsilon x.$$

- The unique minimizer of f is $x_* = (-\epsilon^{-1}, 0)$.

An example where R is UNBOUNDED

Objective function

Let $\epsilon \in (0, \frac{1}{2})$ and f be strongly convex function and parameterized by ϵ

$$f(x, y) \ = \ y^2 + \frac{1}{2}(\epsilon x)^2 + \epsilon x.$$

- The unique minimizer of f is $x_* = (-\epsilon^{-1}, 0)$.
- The Lipschitz constant ν of the gradient of f is at most 2.

An example where R is UNBOUNDED

Objective function

Let $\epsilon \in (0,\ \frac{1}{2})$ and f be strongly convex function and parameterized by ϵ

$$f(x, y) \ = \ y^2 + \frac{1}{2}(\epsilon x)^2 + \epsilon x.$$

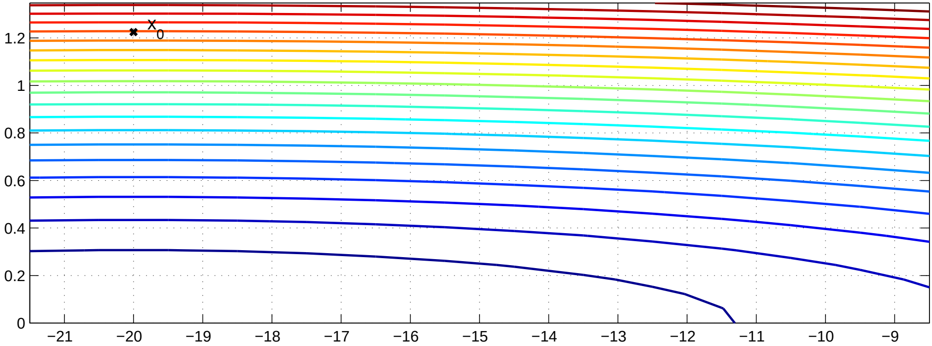
- The unique minimizer of f is $x_* = (-\epsilon^{-1}, 0)$.
- The Lipschitz constant ν of the gradient of f is at most 2.

Algorithmic choices

Using $\gamma = 1$ (suc. iterates), $x_0 = (-\epsilon^{-1}, \frac{\sqrt{6}}{2})$, $\alpha_0 = 1$, $\rho(\alpha) = \epsilon\alpha^2$, and

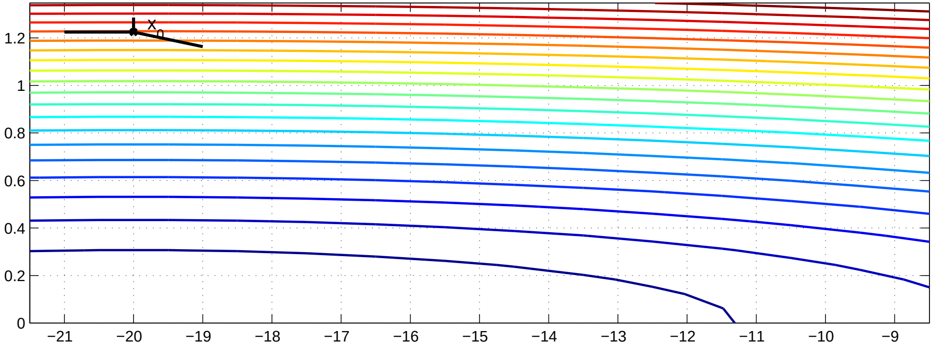
$$D \ = \begin{bmatrix} 1 & 0 & -1 \\ -\frac{\sqrt{6}}{2}\epsilon & \frac{\sqrt{6}}{2}\epsilon & 0 \end{bmatrix}.$$

An example where R is UNBOUNDED



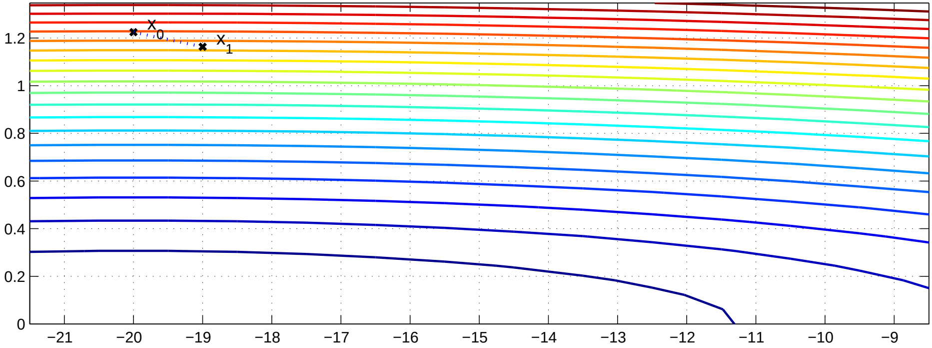
$$\begin{cases} \epsilon = 0.05, & \|x_0 - x_*\| = \frac{\sqrt{6}}{2}, \\ \|\nabla f(x_k)\| > \epsilon, & \|x_k - x_*\| \geq k. \end{cases}$$

An example where R is UNBOUNDED



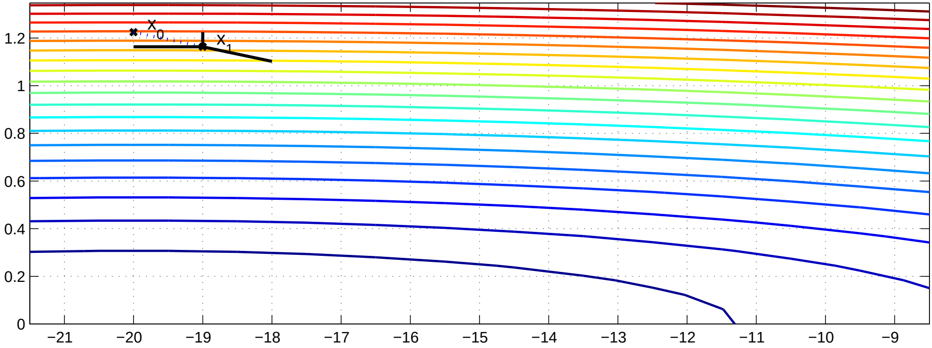
$$\begin{cases} \epsilon = 0.05, & \|x_0 - x_*\| = \frac{\sqrt{6}}{2}, \\ \|\nabla f(x_k)\| > \epsilon, & \|x_k - x_*\| \geq k. \end{cases}$$

An example where R is UNBOUNDED



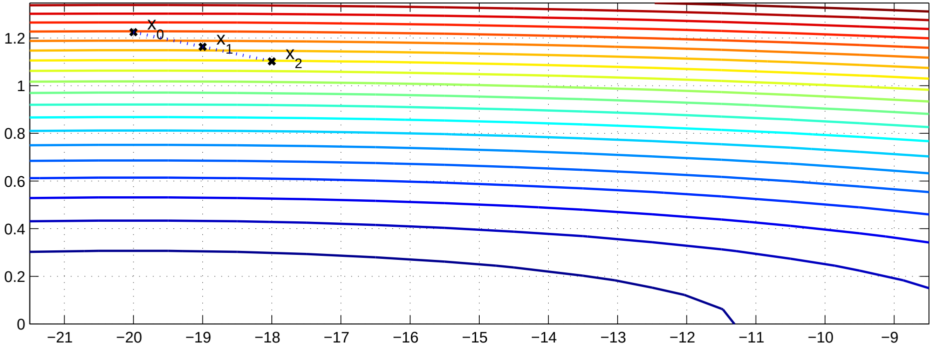
$$\begin{cases} \epsilon = 0.05, & \|x_0 - x_*\| = \frac{\sqrt{6}}{2}, \\ \|\nabla f(x_k)\| > \epsilon, & \|x_k - x_*\| \geq k. \end{cases}$$

An example where R is UNBOUNDED



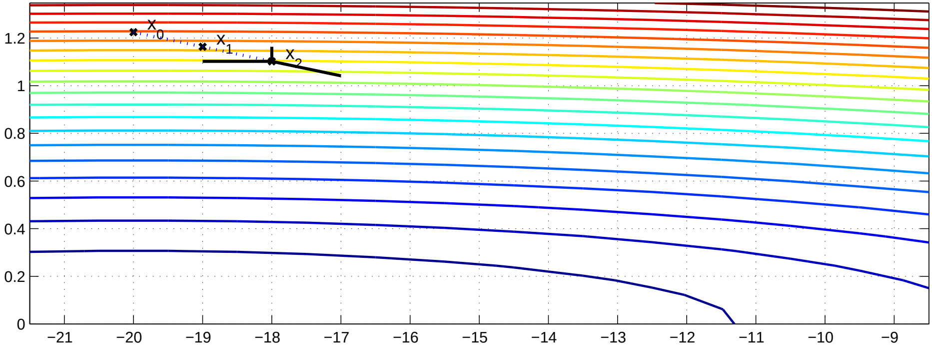
$$\begin{cases} \epsilon = 0.05, & \|x_0 - x_*\| = \frac{\sqrt{6}}{2}, \\ \|\nabla f(x_k)\| > \epsilon, & \|x_k - x_*\| \geq k. \end{cases}$$

An example where R is UNBOUNDED



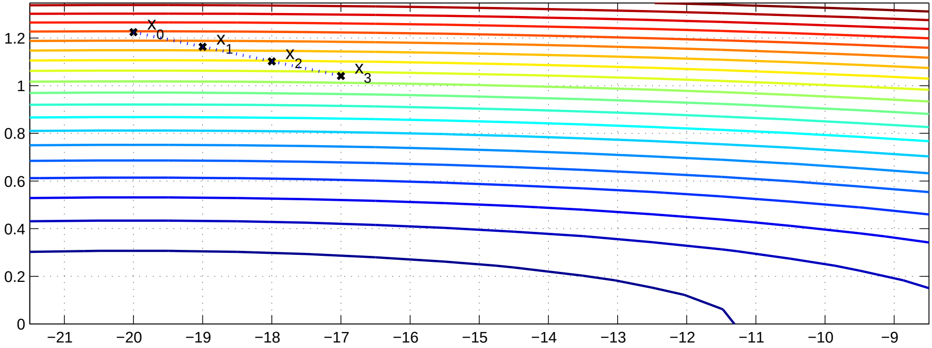
$$\begin{cases} \epsilon = 0.05, & \|x_0 - x_*\| = \frac{\sqrt{6}}{2}, \\ \|\nabla f(x_k)\| > \epsilon, & \|x_k - x_*\| \geq k. \end{cases}$$

An example where R is UNBOUNDED



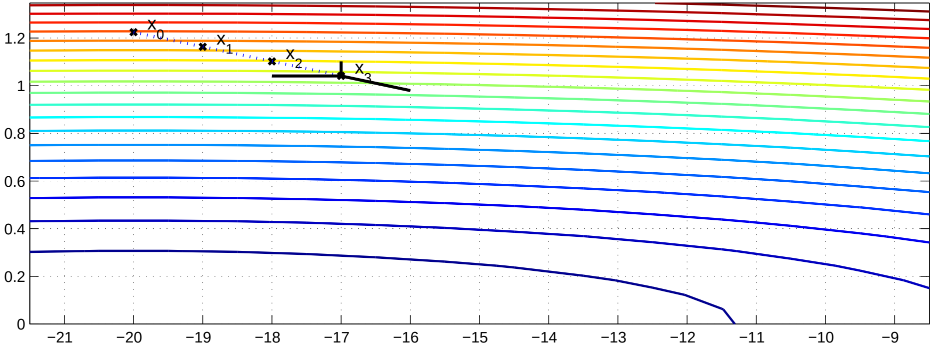
$$\begin{cases} \epsilon = 0.05, & \|x_0 - x_*\| = \frac{\sqrt{6}}{2}, \\ \|\nabla f(x_k)\| > \epsilon, & \|x_k - x_*\| \geq k. \end{cases}$$

An example where R is UNBOUNDED



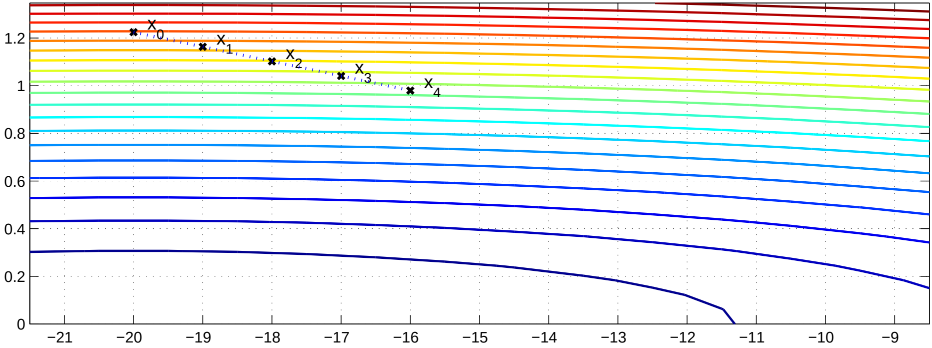
$$\begin{cases} \epsilon = 0.05, & \|x_0 - x_*\| = \frac{\sqrt{6}}{2}, \\ \|\nabla f(x_k)\| > \epsilon, & \|x_k - x_*\| \geq k. \end{cases}$$

An example where R is UNBOUNDED



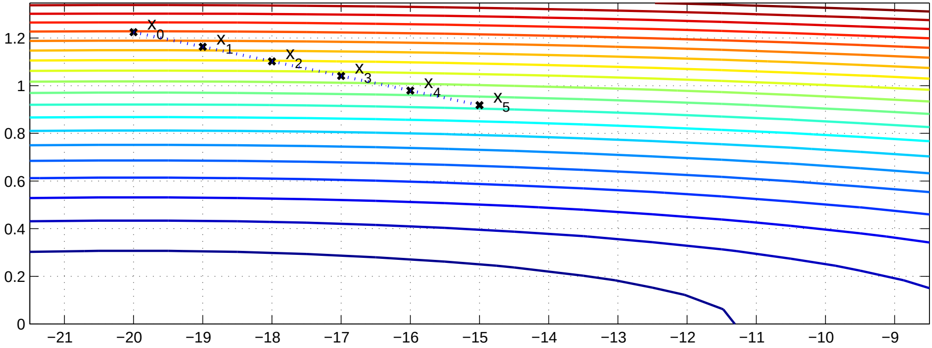
$$\begin{cases} \epsilon = 0.05, & \|x_0 - x_*\| = \frac{\sqrt{6}}{2}, \\ \|\nabla f(x_k)\| > \epsilon, & \|x_k - x_*\| \geq k. \end{cases}$$

An example where R is UNBOUNDED



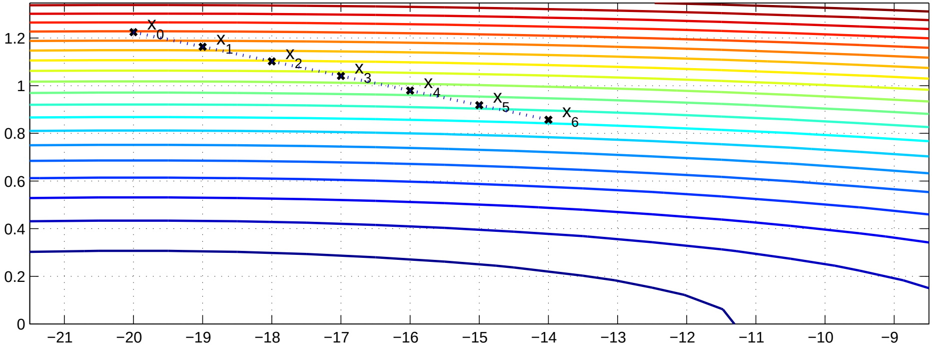
$$\begin{cases} \epsilon = 0.05, & \|x_0 - x_*\| = \frac{\sqrt{6}}{2}, \\ \|\nabla f(x_k)\| > \epsilon, & \|x_k - x_*\| \geq k. \end{cases}$$

An example where R is UNBOUNDED



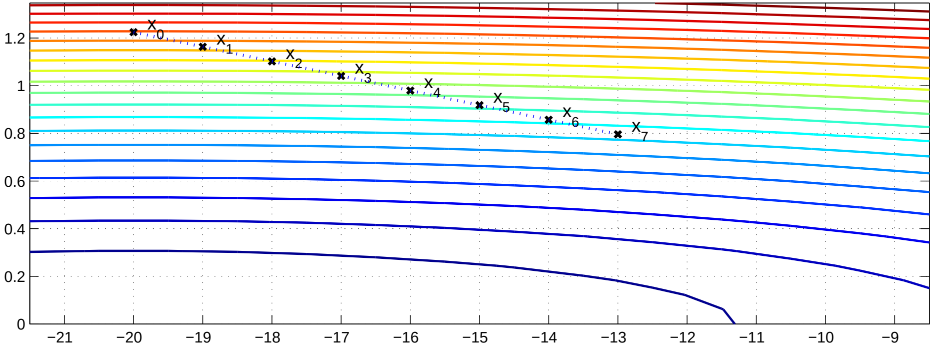
$$\begin{cases} \epsilon = 0.05, & \|x_0 - x_*\| = \frac{\sqrt{6}}{2}, \\ \|\nabla f(x_k)\| > \epsilon, & \|x_k - x_*\| \geq k. \end{cases}$$

An example where R is UNBOUNDED



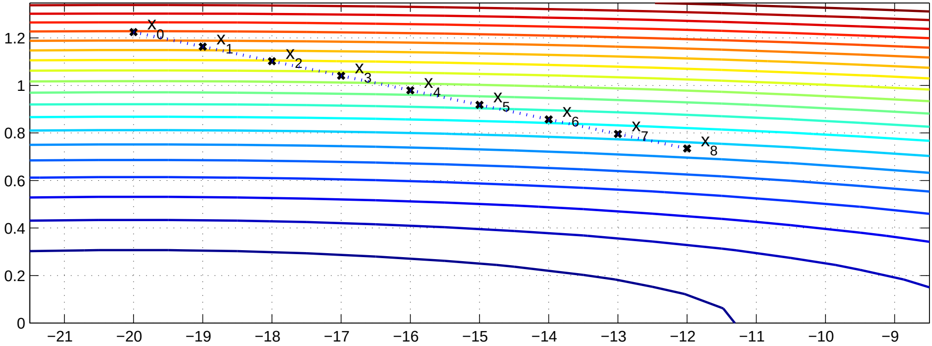
$$\begin{cases} \epsilon = 0.05, & \|x_0 - x_*\| = \frac{\sqrt{6}}{2}, \\ \|\nabla f(x_k)\| > \epsilon, & \|x_k - x_*\| \geq k. \end{cases}$$

An example where R is UNBOUNDED



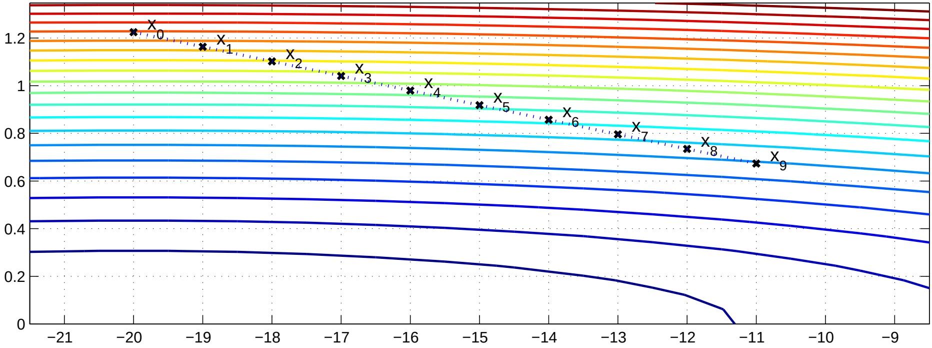
$$\begin{cases} \epsilon = 0.05, & \|x_0 - x_*\| = \frac{\sqrt{6}}{2}, \\ \|\nabla f(x_k)\| > \epsilon, & \|x_k - x_*\| \geq k. \end{cases}$$

An example where R is UNBOUNDED



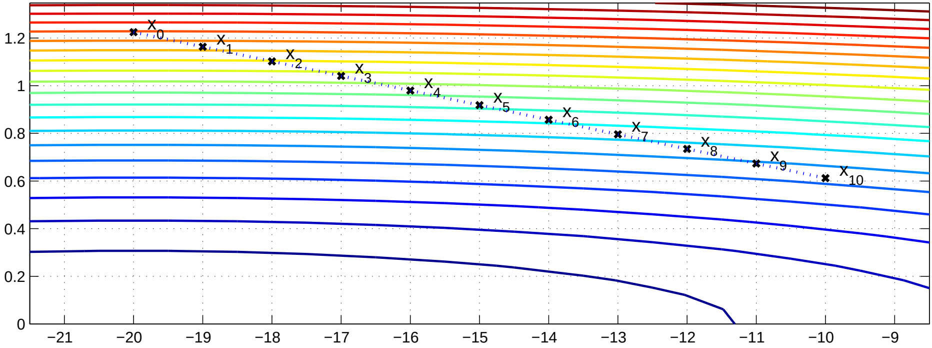
$$\begin{cases} \epsilon = 0.05, & \|x_0 - x_*\| = \frac{\sqrt{6}}{2}, \\ \|\nabla f(x_k)\| > \epsilon, & \|x_k - x_*\| \geq k. \end{cases}$$

An example where R is UNBOUNDED



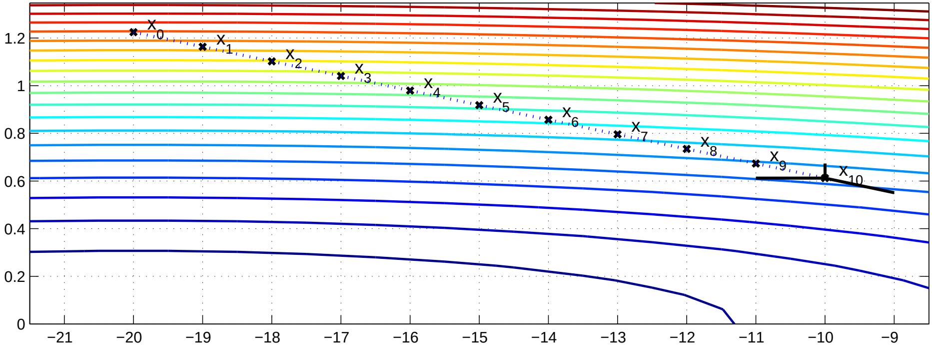
$$\begin{cases} \epsilon = 0.05, & \|x_0 - x_*\| = \frac{\sqrt{6}}{2}, \\ \|\nabla f(x_k)\| > \epsilon, & \|x_k - x_*\| \geq k. \end{cases}$$

An example where R is UNBOUNDED

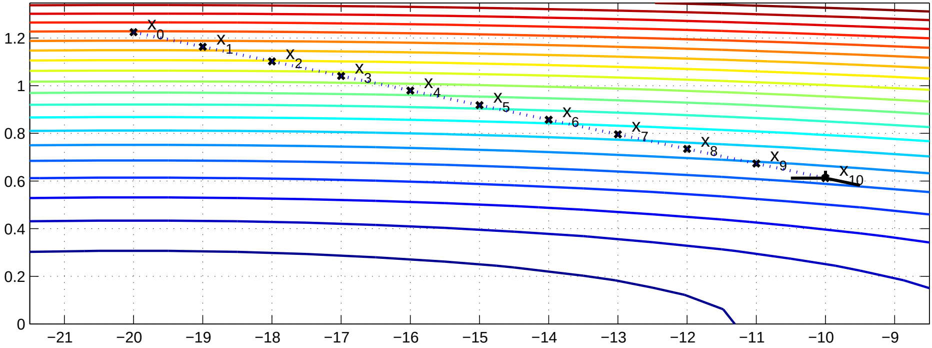


$$\begin{cases} \epsilon = 0.05, & \|x_0 - x_*\| = \frac{\sqrt{6}}{2}, \\ \|\nabla f(x_k)\| > \epsilon, & \|x_k - x_*\| \geq k. \end{cases}$$

An example where R is UNBOUNDED

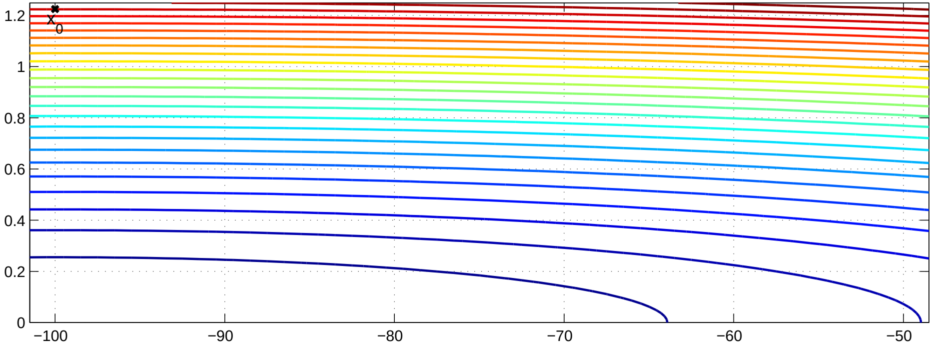


$$\begin{cases} \epsilon = 0.05, & \|x_0 - x_*\| = \frac{\sqrt{6}}{2}, \\ \|\nabla f(x_k)\| > \epsilon, & \|x_k - x_*\| \geq k. \end{cases}$$



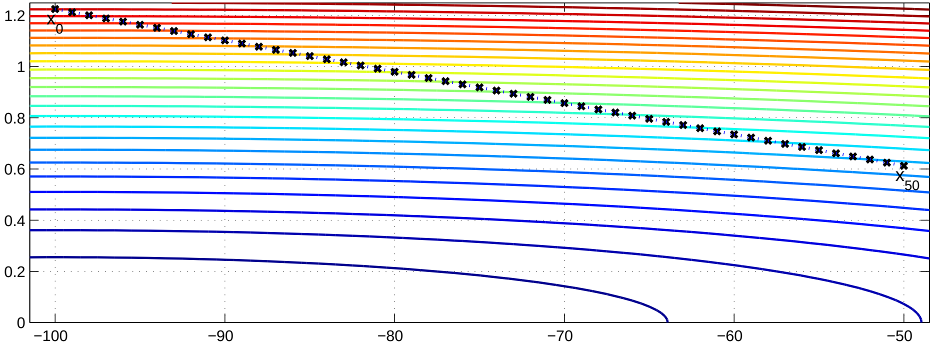
$$\begin{cases} \epsilon = 0.05, & \|x_0 - x_*\| = \frac{\sqrt{6}}{2}, \\ k_0 = 10 \geq \frac{\epsilon^{-1}-1}{2}, & \|x_k - x_*\| \geq k. \end{cases}$$

An example where R is UNBOUNDED



$$\begin{cases} \epsilon = 0.01, & \|x_0 - x_*\| = \frac{\sqrt{6}}{2}, \\ \|\nabla f(x_k)\| > \epsilon, & \|x_k - x_*\| \geq k. \end{cases}$$

An example where R is UNBOUNDED



$$\begin{cases} \epsilon = 0.01, & \|x_0 - x_*\| = \frac{\sqrt{6}}{2}, \\ k_0 = 50 \geq \frac{\epsilon^{-1}-1}{2}, & \|x_k - x_*\| \geq k. \end{cases}$$

An example where R is UNBOUNDED

The distance from the unsuccessful iterate x_{k_0} to X_*^f is arbitrarily large,

$$\|x_{k_0} - x_*\| \geq \frac{\epsilon^{-1} - 1}{2}$$

An example where R is UNBOUNDED

The distance from the unsuccessful iterate x_{k_0} to X_*^f is arbitrarily large,

$$\|x_{k_0} - x_*\| \geq \frac{\epsilon^{-1} - 1}{2}$$

and so one has that

$$R = \mathcal{O}(\epsilon^{-1}).$$

An example where R is UNBOUNDED

The distance from the unsuccessful iterate x_{k_0} to X_*^f is arbitrarily large,

$$\|x_{k_0} - x_*\| \geq \frac{\epsilon^{-1} - 1}{2}$$

and so one has that

$$R = \mathcal{O}(\epsilon^{-1}).$$

One sees immediately that our theory cannot predict better than $\mathcal{O}(\epsilon^{-2})$.

An example where R is UNBOUNDED

The distance from the unsuccessful iterate x_{k_0} to X_*^f is arbitrarily large,

$$\|x_{k_0} - x_*\| \geq \frac{\epsilon^{-1} - 1}{2}$$

and so one has that

$$R = \mathcal{O}(\epsilon^{-1}).$$

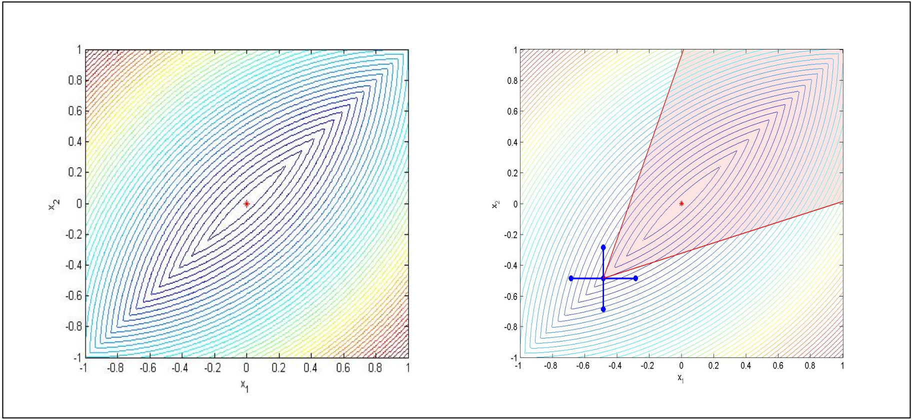
One sees immediately that our theory cannot predict better than $\mathcal{O}(\epsilon^{-2})$.

However, one has

$$\nu\|x_0 - x_*\| \leq \sqrt{6}$$

and one expects the global rate $\mathcal{O}(\epsilon^{-1})$ to hold for gradient methods.

Difficulties in the nonsmooth case



The cone of descent directions at the poll center is shaded.

One possible fix: Infinite number of directions

One possibility is to use an infinite number of polling directions.

One possible fix: Infinite number of directions

One possibility is to use an infinite number of polling directions.

This does not pose a problem to global convergence, which can be guaranteed a.e. in the unit sphere:

One possible fix: Infinite number of directions

One possibility is to use an infinite number of polling directions.

This does not pose a problem to global convergence, which can be guaranteed a.e. in the unit sphere:

- C. Audet and J. E. Dennis Jr., Mesh adaptive direct search algorithms for constrained optimization, SIAM J. on Optimization, 17 (2006) 188-217.

  LTMADS, ORTHOMADS: ways of dense generation guaranteeing the integer lattice requirement.

One possible fix: Infinite number of directions

One possibility is to use an infinite number of polling directions.

This does not pose a problem to global convergence, which can be guaranteed a.e. in the unit sphere:

- C. Audet and J. E. Dennis Jr., Mesh adaptive direct search algorithms for constrained optimization, SIAM J. on Optimization, 17 (2006) 188-217.

  LTMADS, ORTHOMADS: ways of dense generation guaranteeing the integer lattice requirement.

- L. N. Vicente and A. L. Custódio Analysis of direct searches for discontinuous functions, Math. Programming, 133 (2012) 299-325.

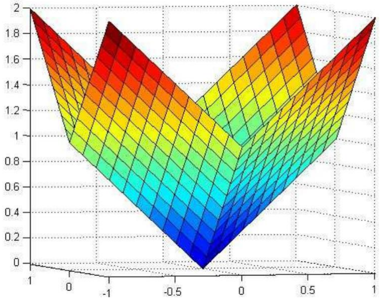
  Dense generation is waived of rules when imposing sufficient decrease.

Another possible fix: Smoothing functions

Definition

We call $\tilde{f} : \mathbb{R}^n \times [0, +\infty) \to \mathbb{R}$ *a* *smoothing function* *of* f *if,* $\forall \mu \in (0, +\infty)$, $\tilde{f}(\cdot, \mu)$ *is* $\mathcal{C}^1$ *and,* $\forall x \in \mathbb{R}^n$,

$$\lim_{z \to x, \mu \downarrow 0} \tilde{f}(z, \mu) \;=\; f(x).$$

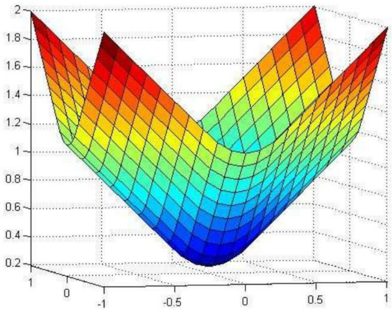


One possible fix: Smoothing functions

Definition

We call $\tilde{f} : \mathbb{R}^n \times [0, +\infty) \to \mathbb{R}$ *a smoothing function of* f *if,* $\forall \mu \in (0, +\infty)$, $\tilde{f}(\cdot, \mu)$ *is* $\mathcal{C}^1$ *and,* $\forall x \in \mathbb{R}^n$,

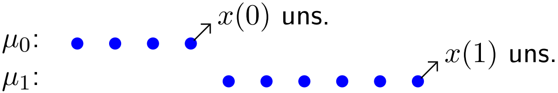
$$\lim_{z \to x, \mu \downarrow 0} \tilde{f}(z, \mu) \;=\; f(x).$$



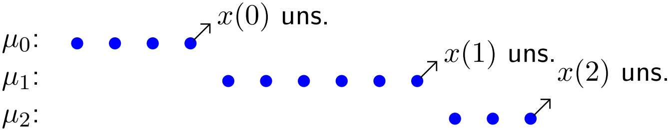
A class of smoothing DS methods

μ_0:  $x(0)$ uns.

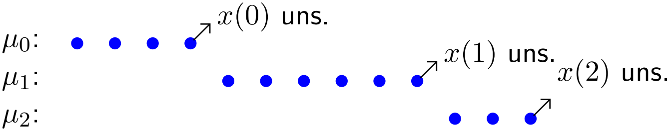
A class of smoothing DS methods



A class of smoothing DS methods



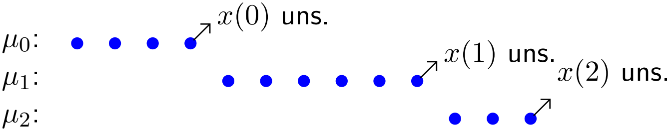
A class of smoothing DS methods



Initialization: Choose a function $r(\cdot)$ such that $\lim_{\mu \downarrow 0} r(\mu) = 0$.

Choose $\mu_0 > 0$, and $\sigma \in (0, 1)$

A class of smoothing DS methods



Initialization: Choose a function $r(\cdot)$ such that $\lim_{\mu \downarrow 0} r(\mu) = 0$.

Choose $\mu_0 > 0$, and $\sigma \in (0, 1)$

For $k = 0, 1, 2 \ldots$ (Until μ_k is suff. small)

- Apply DS to $\tilde{f}(\cdot, \mu_k)$ until step size $< r(\mu_k)$.
- Decrease the smoothing parameter: $\mu_{k+1} = \sigma \mu_k$.

Global convergence of smoothing DS (behavior of μ)

Assumption

Smoothing functions and their level sets are bounded for all k.

Global convergence of smoothing DS (behavior of μ)

Assumption

Smoothing functions and their level sets are bounded for all k.

If we let DS run forever for a given k, then $\alpha \longrightarrow 0$. Thus

Global convergence of smoothing DS (behavior of μ)

Assumption

Smoothing functions and their level sets are bounded for all k.

If we let DS run forever for a given k, then $\alpha \longrightarrow 0$. Thus

Theorem

The smoothing parameter goes to zero: $\lim_{k\to\infty} \mu_k = 0.$

Global convergence of smoothing DS (behavior of μ)

Assumption

Smoothing functions and their level sets are bounded for all k.

If we let DS run forever for a given k, then $\alpha \longrightarrow 0$. Thus

Theorem

The smoothing parameter goes to zero: $\lim_{k\to\infty} \mu_k = 0.$

Theorem

1. $\lim_{k\to+\infty} \alpha(k) = 0.$
2. $\exists x_*$ *and a subsequence* $K \subseteq \{(0),(1),\ldots\}$ *of unsucc. DS iterates such that* $x(k) \underset{K}{\longrightarrow} x_*$.

Now, $\|\nabla \tilde{f}(x(k), \mu_k)\| \;\leq\; C(\tilde{\nu}(\mu_k))\,\alpha(k)$

Global convergence of smoothing DS

Now, $\|\nabla \tilde{f}(x(k), \mu_k)\| \;\leq\; C(\tilde{\nu}(\mu_k))\, \alpha(k) \;\leq\; C(\tilde{\nu}(\mu_k))\, r(\mu_k)$.

Global convergence of smoothing DS

Now, $\|\nabla \tilde{f}(x(k), \mu_k)\| \;\leq\; C(\tilde{\nu}(\mu_k))\,\alpha(k) \;\leq\; C(\tilde{\nu}(\mu_k))\,r(\mu_k)$.

Thus, choosing $r(\cdot)$ appropriately, i.e., $r(\mu) = \mu^2$ when $\tilde{\nu}(\mu) = \mathcal{O}\left(\frac{1}{\mu}\right)$:

Global convergence of smoothing DS

Now, $\|\nabla \tilde{f}(x(k), \mu_k)\| \;\leq\; C(\tilde{\nu}(\mu_k))\, \alpha(k) \;\leq\; C(\tilde{\nu}(\mu_k))\, r(\mu_k)$.

Thus, choosing $r(\cdot)$ appropriately, i.e., $r(\mu) = \mu^2$ when $\tilde{\nu}(\mu) = \mathcal{O}\left(\frac{1}{\mu}\right)$:

Theorem

$$\lim_{k \in K} \|\nabla \tilde{f}(x(k), \mu_k)\| \;=\; 0$$

and x_* is stationary point associated with the smoothing function $\tilde{f}$.

Global convergence of smoothing DS

Now, $\|\nabla \tilde{f}(x(k), \mu_k)\| \;\leq\; C(\tilde{\nu}(\mu_k))\, \alpha(k) \;\leq\; C(\tilde{\nu}(\mu_k))\, r(\mu_k)$.

Thus, choosing $r(\cdot)$ appropriately, i.e., $r(\mu) = \mu^2$ when $\tilde{\nu}(\mu) = \mathcal{O}\left(\frac{1}{\mu}\right)$:

Theorem

$$\lim_{k \in K} \|\nabla \tilde{f}(x(k), \mu_k)\| \;=\; 0$$

and x_* is stationary point associated with the smoothing function $\tilde{f}$.

Definition

We say that x_* is a stationary point associated with the smoothing function $\tilde{f}$ if $0 \in G_{\tilde{f}}(x_*)$, where

$$G_{\tilde{f}}(x_*) \;=\; \{\textit{all limits of}\, \nabla \tilde{f}(x, \mu) \textit{ when } x \to x_* \textit{ and } \mu \to 0\}.$$

Clarke generalized derivative and subdifferential

Does $0 \in G_{\tilde{f}}(x_*)$ mean any form of true stationarity?

Clarke generalized derivative and subdifferential

Does $0 \in G_{\tilde{f}}(x_*)$ mean any form of true stationarity?

Definition

Let f be Lipschitz cont. near x.

Clarke generalized derivative and subdifferential

Does $0 \in G_{\tilde{f}}(x_*)$ mean any form of true stationarity?

Definition

Let f be Lipschitz cont. near x. The Clarke generalized directional derivative is defined by

$$f^\circ(x;v) \;=\; \limsup_{\bar{x}\to x \;\; t\downarrow 0} \frac{f(\bar{x}+tv)-f(\bar{x})}{t}.$$

Clarke generalized derivative and subdifferential

Does $0 \in G_{\tilde{f}}(x_*)$ mean any form of true stationarity?

Definition

Let f be Lipschitz cont. near x. The Clarke generalized directional derivative is defined by

$$f^\circ(x;v) \;=\; \limsup_{\bar{x}\to x \;\; t\downarrow 0} \frac{f(\bar{x}+tv)-f(\bar{x})}{t}.$$

The Clarke subdifferential is given by:

$$\partial f(x) \;=\; \{s \in \mathbb{R}^n : f^\circ(x;v) \geq \langle v, s\rangle,\; \forall v \in \mathbb{R}^n\}.$$

Clarke generalized derivative and subdifferential

Does $0 \in G_{\tilde{f}}(x_*)$ mean any form of true stationarity?

Definition

Let f be Lipschitz cont. near x. The Clarke generalized directional derivative is defined by

$$f^\circ(x; v) \;=\; \limsup_{\bar{x} \to x \;\; t \downarrow 0} \frac{f(\bar{x} + tv) - f(\bar{x})}{t}.$$

The Clarke subdifferential is given by:

$$\partial f(x) \;=\; \{s \in \mathbb{R}^n : f^\circ(x; v) \geq \langle v, s \rangle, \; \forall v \in \mathbb{R}^n\}.$$

Clarke stationarity

If x_* is a local minimizer, $0 \in \partial f(x_*)$.

Clarke generalized derivative and subdifferential

Does $0 \in G_{\tilde{f}}(x_*)$ mean any form of true stationarity?

Definition

Let f be Lipschitz cont. near x. The Clarke generalized directional derivative is defined by

$$f^\circ(x;v) \ = \ \limsup_{\bar{x}\to x \ t\downarrow 0} \frac{f(\bar{x}+tv)-f(\bar{x})}{t}.$$

The Clarke subdifferential is given by:

$$\partial f(x) \ = \ \{s \in \mathbb{R}^n : f^\circ(x;v) \geq \langle v, s\rangle, \ \forall v \in \mathbb{R}^n\}.$$

Clarke stationarity

If x_* is a local minimizer,

Clarke generalized derivative and subdifferential

Does $0 \in G_{\tilde{f}}(x_*)$ mean any form of true stationarity?

Definition

Let f be Lipschitz cont. near x. The Clarke generalized directional derivative is defined by

$$f^\circ(x; v) \;=\; \limsup_{\bar{x} \to x \;\; t \downarrow 0} \frac{f(\bar{x} + tv) - f(\bar{x})}{t}.$$

The Clarke subdifferential is given by:

$$\partial f(x) \;=\; \{s \in \mathbb{R}^n : f^\circ(x; v) \geq \langle v, s \rangle,\ \forall v \in \mathbb{R}^n\}.$$

Clarke stationarity

If x_* is a local minimizer, $f^\circ(x; v) \geq 0,\ \forall v \in \mathbb{R}^n$.

How to construct smoothing functions

There are forms of building smoothing functions $\tilde{f}$ such that

How to construct smoothing functions

There are forms of building smoothing functions $\tilde{f}$ such that

- $\tilde{f}$ satisfies the gradient consistency property

$$\partial f(x_*) \; = \; \text{co}\, G_{\tilde{f}}(x_*).$$

How to construct smoothing functions

There are forms of building smoothing functions $\tilde{f}$ such that

- $\tilde{f}$ satisfies the gradient consistency property

$$\partial f(x_*) = \text{co}\, G_{\tilde{f}}(x_*).$$

Thus, if $0 \in G_{\tilde{f}}(x_*) \subset \text{co}\, G_{\tilde{f}}(x_*)$, then $0 \in \partial f(x_*)$.

How to construct smoothing functions

There are forms of building smoothing functions $\tilde{f}$ such that

- $\tilde{f}$ satisfies the gradient consistency property

$$\partial f(x_*) \;=\; \text{co}\; G_{\tilde{f}}(x_*).$$

  Thus, if $0 \in G_{\tilde{f}}(x_*) \subset \text{co}\, G_{\tilde{f}}(x_*)$, then $0 \in \partial f(x_*)$.

- $\tilde{\nu}(\mu) \;=\; \mathcal{O}\left(\frac{1}{\mu}\right)$.

How to construct smoothing functions

There are forms of building smoothing functions $\tilde{f}$ such that

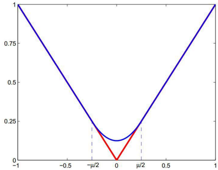
- $\tilde{f}$ satisfies the gradient consistency property

$$\partial f(x_*) \; = \; \text{co}\; G_{\tilde{f}}(x_*).$$

  Thus, if $0 \in G_{\tilde{f}}(x_*) \subset \text{co}\, G_{\tilde{f}}(x_*)$, then $0 \in \partial f(x_*)$.

- $\tilde{\nu}(\mu) \; = \; \mathcal{O}\left(\frac{1}{\mu}\right)$.

Chen and Zhou introduced such a smoothing function $\tilde{s}(t, \mu)$ of $|t|$:



How to construct smoothing functions

There are forms of building smoothing functions $\tilde{f}$ such that

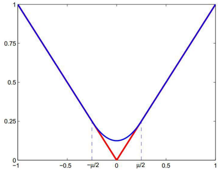
- $\tilde{f}$ satisfies the gradient consistency property

$$\partial f(x_*) \; = \; \text{co}\; G_{\tilde{f}}(x_*).$$

  Thus, if $0 \in G_{\tilde{f}}(x_*) \subset \text{co}\, G_{\tilde{f}}(x_*)$, then $0 \in \partial f(x_*)$.

- $\tilde{\nu}(\mu) \; = \; \mathcal{O}\left(\frac{1}{\mu}\right)$.

Chen and Zhou introduced such a smoothing function $\tilde{s}(t,\mu)$ of $|t|$:



Then we obtain $\tilde{F}(x,\mu) = \sum_{i=1}^{m} \tilde{s}(F_i(x),\mu)$ for $\|F\|_1 = \sum_{i=1}^{m} |F_i|$.

WCC of smoothing DS (to reduce μ)

Theorem

Let $\rho(\alpha) = \alpha^p$ *and* $r(\alpha) = \alpha^q$, *with* $p, q > 1$.

WCC of smoothing DS (to reduce μ)

Theorem

Let $\rho(\alpha) = \alpha^p$ and $r(\alpha) = \alpha^q$, with $p, q > 1$.

Any smoothing DS (based on sufficient decrease) takes at most

$$\mathcal{O}\left((-\log(\xi))\xi^{-pq}\right)$$

DS inner iterations to reduce μ below $\xi \in (0,1)$.

WCC of smoothing DS (to reduce smoothing gradient)

Corollary

Assume $\tilde{\nu}(\mu) = \mathcal{O}(1/\mu)$.

WCC of smoothing DS (to reduce smoothing gradient)

Corollary

Assume $\tilde{\nu}(\mu) = \mathcal{O}(1/\mu)$.

When μ *becomes lower than* ξ, $\nabla \tilde{f}$ *becomes*

$$\mathcal{O}\left(n^{\frac{1}{2}}(\xi^{q-1} + \xi^{(p-1)q})\right).$$

WCC of smoothing DS (to reduce smoothing gradient)

Corollary

Assume $\tilde{\nu}(\mu) = \mathcal{O}(1/\mu)$.

When μ *becomes lower than* ξ, $\nabla \tilde{f}$ *becomes*

$$\mathcal{O}\left(n^{\frac{1}{2}}(\xi^{q-1} + \xi^{(p-1)q})\right).$$

So, for having $\xi^{q-1} + \xi^{(p-1)q} = \mathcal{O}(\xi)$,

WCC of smoothing DS (to reduce smoothing gradient)

Corollary

Assume $\tilde{\nu}(\mu) = \mathcal{O}(1/\mu)$.

When μ *becomes lower than* ξ, $\nabla \tilde{f}$ *becomes*

$$\mathcal{O}\left(n^{\frac{1}{2}}(\xi^{q-1} + \xi^{(p-1)q})\right).$$

So, for having $\xi^{q-1} + \xi^{(p-1)q} = \mathcal{O}(\xi)$, one selects

$$p = \frac{3}{2} \quad \text{and} \quad q = 2$$

WCC of smoothing DS (to reduce smoothing gradient)

Corollary

Assume $\tilde{\nu}(\mu) = \mathcal{O}(1/\mu)$.

When μ *becomes lower than* ξ, $\nabla \tilde{f}$ *becomes*

$$\mathcal{O}\left(n^{\frac{1}{2}}(\xi^{q-1} + \xi^{(p-1)q})\right).$$

So, for having $\xi^{q-1} + \xi^{(p-1)q} = \mathcal{O}(\xi)$, one selects

$$p = \frac{3}{2} \quad \text{and} \quad q = 2$$

leading to

$$\mathcal{O}(n^{\frac{1}{2}}\xi).$$

WCC of smoothing DS (function evaluations)

Therefore, the number of iterations needed to reach $\|\nabla \tilde{f}\| \leq \epsilon$ and $\mu \leq \xi = \mathcal{O}(n^{-\frac{1}{2}}\epsilon)$ is

$$\mathcal{O}\left((-\log(\xi))\xi^{-pq}\right) \overset{p=\frac{3}{2},q=2}{=} \mathcal{O}\left(n^{\frac{3}{2}}[-\log(\epsilon)+\log(n)]\epsilon^{-3}\right).$$

Therefore, the number of iterations needed to reach $\|\nabla \tilde{f}\| \leq \epsilon$ and $\mu \leq \xi = \mathcal{O}(n^{-\frac{1}{2}}\epsilon)$ is

$$\mathcal{O}\left((-\log(\xi))\xi^{-pq}\right) \overset{p=\frac{3}{2},q=2}{=} \mathcal{O}\left(n^{\frac{3}{2}}[-\log(\epsilon)+\log(n)]\epsilon^{-3}\right).$$

In terms of function evaluations:

$$\mathcal{O}\left(n^{\frac{5}{2}}[-\log(\epsilon)+\log(n)]\epsilon^{-3}\right).$$

WCC of smoothing DS (function evaluations)

Therefore, the number of iterations needed to reach $\|\nabla \tilde{f}\| \leq \epsilon$ and $\mu \leq \xi = \mathcal{O}(n^{-\frac{1}{2}}\epsilon)$ is

$$\mathcal{O}\left((-\log(\xi))\xi^{-pq}\right) \overset{p=\frac{3}{2},q=2}{=} \mathcal{O}\left(n^{\frac{3}{2}}[-\log(\epsilon)+\log(n)]\epsilon^{-3}\right).$$

In terms of function evaluations:

$$\mathcal{O}\left(n^{\frac{5}{2}}[-\log(\epsilon)+\log(n)]\epsilon^{-3}\right).$$

This compares to $\mathcal{O}\left(n^3\epsilon^{-3}\right)$ using Gaussian densities (Nesterov, 2011).

WCC of smoothing DS (function evaluations)

Therefore, the number of iterations needed to reach $\|\nabla \tilde{f}\| \leq \epsilon$ and $\mu \leq \xi = \mathcal{O}(n^{-\frac{1}{2}}\epsilon)$ is

$$\mathcal{O}\left((-\log(\xi))\xi^{-pq}\right) \overset{p=\frac{3}{2},q=2}{=} \mathcal{O}\left(n^{\frac{3}{2}}[-\log(\epsilon)+\log(n)]\epsilon^{-3}\right).$$

In terms of function evaluations:

$$\mathcal{O}\left(n^{\frac{5}{2}}[-\log(\epsilon)+\log(n)]\epsilon^{-3}\right).$$

This compares to $\mathcal{O}\left(n^3\epsilon^{-3}\right)$ using Gaussian densities (Nesterov, 2011).

Reference:

- R. Garmanjani and L. N. Vicente, Smoothing and worst-case complexity for direct-search methods in nonsmooth optimization, to appear in IMA Journal of Numerical Analysis .

Summary of DS global rates

Imposing sufficient decrease to accept new iterates, as in derivative-based optimization:

Summary of DS global rates

Imposing sufficient decrease to accept new iterates, as in derivative-based optimization:

- $\mathcal{O}(\epsilon^{-1})$ smooth, convex (under sort of strong convexity...).

Summary of DS global rates

Imposing sufficient decrease to accept new iterates, as in derivative-based optimization:

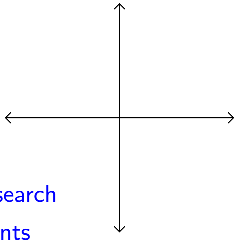
- $\mathcal{O}(\epsilon^{-1})$ smooth, convex (under sort of strong convexity...).
- $\mathcal{O}(\epsilon^{-2})$ smooth, non-convex.

Summary of DS global rates

Imposing sufficient decrease to accept new iterates, as in derivative-based optimization:

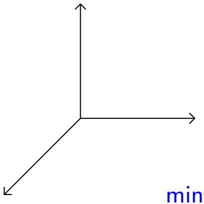
- $\mathcal{O}(\epsilon^{-1})$ smooth, convex (under sort of strong convexity...).
- $\mathcal{O}(\epsilon^{-2})$ smooth, non-convex.
- $\mathcal{O}(\epsilon^{-3})$ non-smooth, non-convex (using smoothing techniques...).

Positive spanning sets



coordinate search

$2n$ elements



minimal case

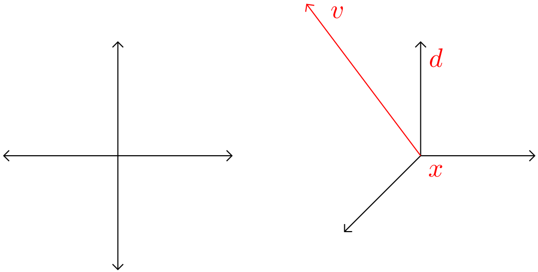
$n+1$ elements

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

$\mathcal{O}\left(\frac{1}{\mathrm{cm}(D_\oplus)^2}\epsilon^{-2}\right) = \mathcal{O}\left(n\epsilon^{-2}\right)$ iterations

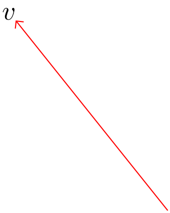
$$\mathrm{cm}(D) \; = \; \min_{0 \neq v \in \mathbb{R}^n} \max_{d \in D} \frac{v^\top d}{\|v\|\|d\|}$$

Positive spanning sets

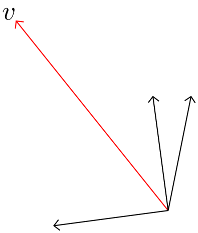


If $v = -\nabla f(x)$ then d is a descent direction.

Randomly generating 'positive spanning sets' ...



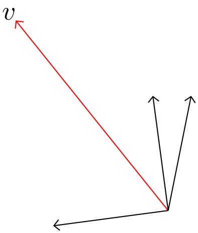
Randomly generating ‘positive spanning sets’ ...

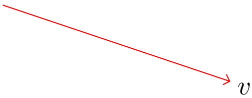


$n+1$ random polling directions

in this case not a PSS

Randomly generating 'positive spanning sets' ...

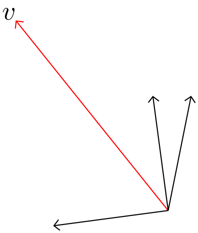




$n + 1$ random polling directions

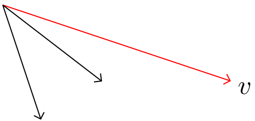
in this case not a PSS

Randomly generating 'positive spanning sets' ...



$n+1$ random polling directions

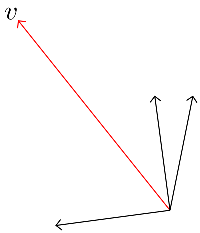
in this case not a PSS



$< n$ random polling directions

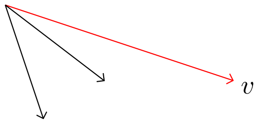
certainly not a PSS...

Randomly generating 'positive spanning sets' ...



$n+1$ random polling directions

in this case not a PSS

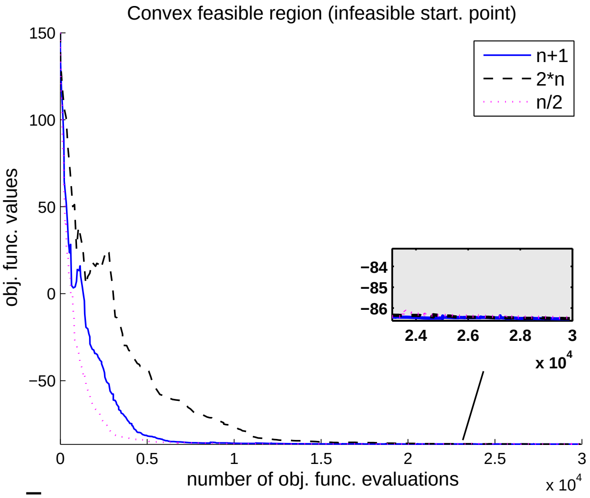


$< n$ random polling directions

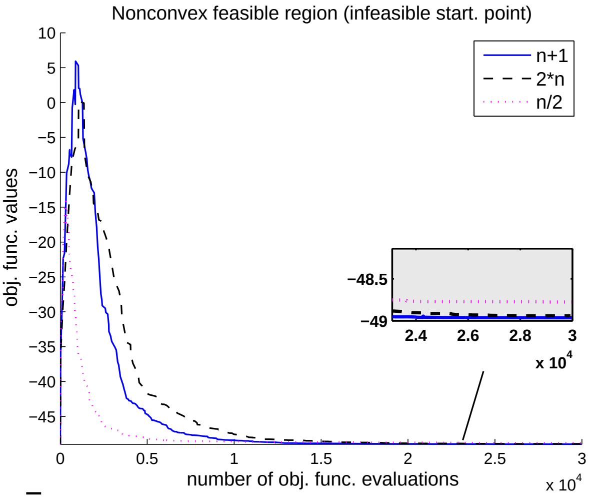
certainly not a PSS...

All we need is $\mathrm{cm}(D, v) \;=\; \max_{d \in D} \dfrac{v^\top d}{\|v\|\|d\|} \;\geq\; \kappa \in (0,1)$

Using $n/2$ random polling directions...



Using $n/2$ random polling directions...



problem	$[Q_k \ -Q_k]$	$2n$	$n+1$	$n/2$
arglina	3958	2954	1681	943
arglinb	266	94	62	44
arwhead	3903	2874	1735	945
bdqrtic	1198	1088	682	369
broydn3d	4196	3491	2005	1202
dqrtic	2485	1533	873	493
engval1	1642	888	566	308
freuroth	4	4	5	6
integreq	3796	3100	1789	956
nondia	882	1162	884	764
nondquar	3105	2719	1694	1052
penalty1	1422	1439	832	462
penalty2	2425	1391	744	458
tquartic	- (100)	28059	20087	14848
vardim	6	17	19	16

fevals to reach an opt. accuracy of 10^{-3}.

Here $n = 20$ and averages where taken for 30 runs.

DS based on probabilistic descent

Assumption

We say that a sequence of polling directions $\{D_k\}$ is (p)-probabilistically κ-descent for corresponding sequences $\{X_k\}$, $\{\mathrm{Alpha}_k\}$ if the events

DS based on probabilistic descent

Assumption

We say that a sequence of polling directions $\{D_k\}$ *is* (p)*-probabilistically* κ*-descent for corresponding sequences* $\{X_k\}$, $\{\mathrm{Alpha}_k\}$ *if the events*

$$S_k = \{\, \mathrm{cm}(D_k, -\nabla f(X_k)) \geq \kappa \,\}$$

DS based on probabilistic descent

Assumption

We say that a sequence of polling directions $\{D_k\}$ *is* (p)*-probabilistically* κ*-descent for corresponding sequences* $\{X_k\}$, $\{\mathrm{Alpha}_k\}$ *if the events*

$$S_k \;=\; \{\, \mathrm{cm}(D_k, -\nabla f(X_k)) \;\geq\; \kappa \,\}$$

satisfy the condition

$$P(S_k | \sigma(D_0, \ldots, D_{k-1})) \;\geq\; p.$$

DS based on probabilistic descent

Assumption

We say that a sequence of polling directions $\{D_k\}$ is (p)-probabilistically κ-descent for corresponding sequences $\{X_k\}$, $\{\mathrm{Alpha}_k\}$ if the events

$$S_k \;=\; \{\, \mathrm{cm}(D_k, -\nabla f(X_k)) \;\geq\; \kappa \,\}$$

satisfy the condition

$$P(S_k | \sigma(D_0, \ldots, D_{k-1})) \;\geq\; p.$$

Furthermore, if $p \geq \frac{1}{2}$, then we say that the polling directions are probabilistically κ-descent.

Global convergence of DS based on prob. descent

Lemma

For every realization of the algorithm, $\lim_{k\to\infty} \alpha_k = 0$.

Global convergence of DS based on prob. descent

Lemma

For every realization of the algorithm, $\lim_{k\to\infty} \alpha_k = 0$.

Theorem

Suppose that the polling directions $\{D_k\}$ *are probabilistically* κ*-descent for some* $\kappa \in (0,1)$.

Global convergence of DS based on prob. descent

Lemma

For every realization of the algorithm, $\lim_{k\to\infty} \alpha_k = 0$.

Theorem

Suppose that the polling directions $\{D_k\}$ *are probabilistically* κ*-descent for some* $\kappa \in (0,1)$.

Let $\{X_k\}$ *be a sequence of random iterates generated by the algorithm.*

Global convergence of DS based on prob. descent

Lemma

For every realization of the algorithm, $\lim_{k\to\infty} \alpha_k = 0$.

Theorem

Suppose that the polling directions $\{D_k\}$ *are probabilistically* κ*-descent for some* $\kappa \in (0,1)$.

Let $\{X_k\}$ *be a sequence of random iterates generated by the algorithm.*

Then,

$$P\left[\lim_{k\to\infty} \|\nabla f(X_k)\| = 0\right] = 1.$$

Global convergence of DS based on prob. descent

Lemma

For every realization of the algorithm, $\lim_{k\to\infty} \alpha_k = 0$.

Theorem

Suppose that the polling directions $\{D_k\}$ *are probabilistically* κ*-descent for some* $\kappa \in (0,1)$.

Let $\{X_k\}$ *be a sequence of random iterates generated by the algorithm.*

Then,

$$P\left[\lim_{k\to\infty} \|\nabla f(X_k)\| = 0\right] = 1.$$

The proof is based on the trust-region corresponding one:

- A. S. Bandeira, K. Scheinberg, and L. N. Vicente, Convergence of trust-region methods based on probabilistic models, submitted.

Worst case complexity of DS based on prob. descent

$p \nearrow$ when $|D| \nearrow$, but what is the effect of this on the performance/analysis of the algorithm?

Worst case complexity of DS based on prob. descent

$p \nearrow$ when $|D| \nearrow$, but what is the effect of this on the performance/analysis of the algorithm?

Theorem

Let $\epsilon \in (0,1)$*. In addition, suppose*

$$P\left[\exists K \;:\; \|\nabla f(X_j)\| \geq \epsilon,\ j = 0,\ldots,K-1,\ \|\nabla f(X_K)\| < \epsilon\right] \;=\; 1.$$

Worst case complexity of DS based on prob. descent

$p \nearrow$ when $|D| \nearrow$, but what is the effect of this on the performance/analysis of the algorithm?

Theorem

Let $\epsilon \in (0,1)$. *In addition, suppose*

$$P\left[\exists K \;:\; \|\nabla f(X_j)\| \geq \epsilon,\; j = 0,\ldots,K-1,\; \|\nabla f(X_K)\| < \epsilon\right] \;=\; 1.$$

Then,

$$P\left[\;\#\text{ function evals until } K \;\leq\; |D| C(\nu)\frac{1}{\epsilon^2}\;\Big|\cdots\right] \;\geq\; 2p(|D|)-1.$$

Worst case complexity of DS based on prob. descent

$p \nearrow$ when $|D| \nearrow$, but what is the effect of this on the performance/analysis of the algorithm?

Theorem

Let $\epsilon \in (0,1)$*. In addition, suppose*

$$P\left[\exists K \ : \ \|\nabla f(X_j)\| \geq \epsilon, \ j = 0,\ldots,K-1, \ \|\nabla f(X_K)\| < \epsilon\right] \ = \ 1.$$

Then,

$$P\left[\ \#\text{ function evals until } K \ \leq \ |D| C(\nu) \frac{1}{\epsilon^2} \ \Big| \cdots \right] \ \geq \ 2p(|D|) - 1.$$

$\cdots$ is (roughly) the σ-algebra until the index corresponding to the smallest step size up to K.

- One can relax the lower bound on p to

$$p \;\geq\; \frac{\ln(C_{dec})}{\ln(C_{dec}/C_{inc})}.$$

DS based on probabilistic descent

- One can relax the lower bound on p to

$$p \;\geq\; \frac{\ln(C_{dec})}{\ln(C_{dec}/C_{inc})}.$$

- When one imposes $p \geq \frac{1}{2}$, one must have $|D| \geq 2$.

problem	$[Q_k \ -Q_k]$	$2n$	$n+1$	$n/2$	$n/4$	2	1
arglina	8.60	6.42	3.65	2.05	1.23	1	- (100)
arglinb	11.08	3.92	2.58	1.83	1.46	1	4.17 (13)
arwhead	8.18	6.03	3.64	1.98	1.19	1	- (100)
bdqrtic	6.62	6.01	3.77	2.04	1.25	1	4.80 (80)
broydn3d	4.71	3.92	2.25	1.35	0.99	1	- (100)
dqrtic	8.28	5.11	2.91	1.64	1.06	1	4.67 (87)
engval1	11.09	6.00	3.82	2.08	1.36	1	4.60 (73)
freuroth	0.67	0.67	0.83	1	1	1	1
integreq	8.38	6.84	3.95	2.11	1.27	1	4.26 (93)
nondia	0.84	1.11	0.84	0.73	0.83	1	0.05 (13)
nondquar	4.27	3.73	2.33	1.45	1.02	1	- (100)
penalty1	5.51	5.58	3.22	1.79	1.17	1	3.82 (70)
penalty2	11.28	6.47	3.46	2.13	1.37	1	5.54 (90)
tquartic	- (100)	1.62	1.16	0.86	0.75	1	- (100)
vardim	0.46	1.31	1.46	1.23	1.08	1	5.54 (3.3)

Now, we display increase in # fevals relatively to using 2 directions.

References and support

References:

- S. Gratton, C. Royer, L. N. Vicente, and Z. Zhang, Direct search based on probabilistic descent, in preparation.
- S. Gratton and L. N. Vicente, A merit function approach for direct search, submitted.

Research supported by Réseau thématique de recherche avancée (RTRA), Fondation de Coopération Sciences et Technologies pour l'Aéronautique et l'Espace (STAE), under the grant ADTAO.