Regional Complexity Analysis of Algorithms for Nonconvex Smooth Optimization

Frank E. Curtis, Lehigh University

joint work with

Daniel P. Robinson, Johns Hopkins University

presented at

DIMACS/TRIPODS/MOPTA
Bethlehem, PA, USA

15 August 2018

Outline

Introduction

Contemporary Analyses

Regional Complexity Analysis: Step 1

Regional Complexity Analysis: Step 2

Summary & Perspectives

Outline

Introduction

Contemporary Analyses

Regional Complexity Analysis: Step 1

Regional Complexity Analysis: Step 2

Summary & Perspectives

Main concern / focus of this talk

Contemporary worst-case analyses for nonconvex optimization are

- overly conservative,
- not representative of actual performance, and
- too simplistic(?)

We should characterize complexity in a different way.

- Purpose of this talk is to convince you.
- (Otherwise, e.g., we may turn people off from second-order methods.)

Problem statement

Let's talk about the problem to minimize $f : \mathbb{R}^n \to \mathbb{R}$:

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

We'll focus on iterative algorithms of the form

$$x_{k+1} \leftarrow x_k + s_k \;\text{ for all }\; k \in \mathbb{N},$$

where $\{x_k\}$ is the iterate sequence and $\{s_k\}$ is the step sequence.

Problem statement

Let's talk about the problem to minimize $f : \mathbb{R}^n \to \mathbb{R}$:

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

We'll focus on iterative algorithms of the form

$$x_{k+1} \leftarrow x_k + s_k \;\; \text{for all} \;\; k \in \mathbb{N},$$

where $\{x_k\}$ is the iterate sequence and $\{s_k\}$ is the step sequence.

For the purposes of this talk. . .

- local search (not global optimization)
- deterministic methods (could extend to stochastic)

Problem statement

Let's talk about the problem to minimize $f : \mathbb{R}^n \to \mathbb{R}$:

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

We'll focus on iterative algorithms of the form

$$x_{k+1} \leftarrow x_k + s_k \;\text{ for all }\; k \in \mathbb{N},$$

where $\{x_k\}$ is the iterate sequence and $\{s_k\}$ is the step sequence.

For the purposes of this talk...

- local search (not global optimization)
- deterministic methods (could extend to stochastic)

Let's use $f_k := f(x_k)$, $g_k := \nabla f(x_k)$, and $H_k := \nabla^2 f(x_k)$.

Worst-case complexity: Contemporary approach

Worst-case complexity: Upper limit on the resources an algorithm will require to (approximately) solve a given problem

Worst-case complexity: Contemporary approach

Worst-case complexity: Upper limit on the resources an algorithm will require to (approximately) solve a given problem

. . . convex optimization: Bound on the number of iterations (or function or derivative evaluations) until

$$\begin{aligned} \|x_k - x_*\| &\leq \epsilon_x \\ \text{or} \quad f_k - f_* &\leq \epsilon_f, \end{aligned}$$

where x_* (f_*) is some global minimizer (minimum).

Worst-case complexity: Contemporary approach

Worst-case complexity: Upper limit on the resources an algorithm will require to (approximately) solve a given problem

. . . convex optimization: Bound on the number of iterations (or function or derivative evaluations) until

$$\|x_k - x_*\| \leq \epsilon_x$$
$$\text{or}\quad f_k - f_* \leq \epsilon_f,$$

where x_* (f_*) is some global minimizer (minimum).

. . . nonconvex optimization: Bound on the number of iterations (or function or derivative evaluations) until

$$\|g_k\| \leq \epsilon_g$$
$$\text{and maybe}\quad H_k \succeq -\epsilon_H I.$$

Worst-case complexity for nonconvex optimization

For example, **it is said** that for first-order stationarity we have the bounds. . .

Algorithm	$\|g_k\| \leq \epsilon_g$
Gradient descent	$\mathcal{O}(\epsilon_g^{-2})$
Second-order trust region (TR)	$\mathcal{O}(\epsilon_g^{-2})$
Cubic regularization (e.g., ARC)	$\mathcal{O}(\epsilon_g^{-3/2})$

(For “short-step versions”, second-order TR is $\mathcal{O}(\epsilon_g^{-3/2})$, but anyway. . .)

This should be surprising to anyone who has used these methods!

TR vs. ARC

TR		ARC
1: Solve to compute s_k: $\min_{s\in\mathbb{R}^n} \; q_k(s) := f_k + g_k^T s + \frac{1}{2} s^T H_k s$ s.t. $\|s\| \leq \delta_k$ (dual: λ_k)	$\xrightarrow{\sigma_k = \frac{\lambda_k}{\delta_k}}$ $\xleftarrow[\delta_k = \|s_k\|]{}$	1: Solve to compute s_k: $\min_{s\in\mathbb{R}^n} \; c_k(s) := f_k + g_k^T s + \frac{1}{2} s^T H_k s + \frac{1}{3}\sigma_k \|s\|^3$
2: Compute ratio: $\rho_k^q \leftarrow \frac{f_k - f(x_k+s_k)}{f_k - q_k(s_k)}$		2: Compute ratio: $\rho_k^c \leftarrow \frac{f_k - f(x_k+s_k)}{f_k - c_k(s_k)}$
3: Update radius: $\rho_k^q \geq \eta$: accept and $\delta_k \nearrow$ $\rho_k^q < \eta$: reject and $\delta_k \searrow$		3: Update regularization: $\rho_k^c \geq \eta$: accept and $\sigma_k \searrow$ $\rho_k^c < \eta$: reject and $\sigma_k \nearrow$

TR vs. ARC

TR

1: Solve to compute s_k:

$$\min_{s\in\mathbb{R}^n} \; q_k(s) := f_k + g_k^T s + \tfrac{1}{2} s^T H_k s \quad \text{s.t. } \|s\| \leq \delta_k \quad (\text{dual: } \lambda_k)$$

2: Compute ratio:

$$\rho_k^q \leftarrow \frac{f_k - f(x_k+s_k)}{f_k - q_k(s_k)}$$

3: Update radius:

$\rho_k^q \geq \eta$: accept and $\delta_k \nearrow$

$\rho_k^q < \eta$: reject and $\delta_k \searrow$

$$\sigma_k = \frac{\lambda_k}{\delta_k} \longrightarrow$$

$$\longleftarrow \delta_k = \|s_k\|$$

ARC

1: Solve to compute s_k:

$$\min_{s\in\mathbb{R}^n} \; c_k(s) := f_k + g_k^T s + \tfrac{1}{2} s^T H_k s + \tfrac{1}{3}\sigma_k \|s\|^3$$

2: Compute ratio:

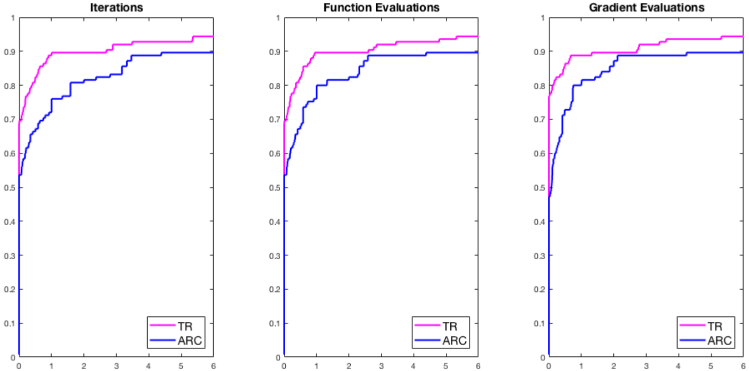
$$\rho_k^c \leftarrow \frac{f_k - f(x_k+s_k)}{f_k - c_k(s_k)}$$

3: Update regularization:

$\rho_k^c \geq \eta$: accept and $\sigma_k \searrow$

$\rho_k^c < \eta$: reject and $\sigma_k \nearrow$

Experiments with CUTEr



Complexity: Take-home message #1

Contemporary complexity theory for nonconvex optimization. . .

- might not be showing a deficiency of certain methods (e.g., 2nd-order TR);
- might be showing a **deficiency of the characterization strategy**.

Complexity: Take-home message #1

Contemporary complexity theory for nonconvex optimization. . .

- might not be showing a deficiency of certain methods (e.g., 2nd-order TR);
- might be showing a **deficiency of the characterization strategy**.

Our goal: A complementary approach to characterize algorithms.

- global convergence
- worst-case complexity, contemporary type + **our new approach**
- local convergence rate

Complexity: Take-home message #1

Contemporary complexity theory for nonconvex optimization. . .

- might not be showing a deficiency of certain methods (e.g., 2nd-order TR);
- might be showing a **deficiency of the characterization strategy**.

Our goal: A complementary approach to characterize algorithms.

- global convergence
- worst-case complexity, contemporary type + **our new approach**
- local convergence rate

We're admitting: Our approach **does not always** give the complete picture.

But the **contemporary approach can give a misleading picture**.

Outline

Introduction

Contemporary Analyses

Regional Complexity Analysis: Step 1

Regional Complexity Analysis: Step 2

Summary & Perspectives

Conservatism of contemporary analyses

Suppose $g := \nabla f$ is Lipschitz continuous with constant $L > 0$. Then,

$$f_{k+1} \leq f_k + g_k^T s_k + \tfrac{1}{2} L \|s_k\|^2.$$

Let $f_{\inf} := \min_{x \in \mathbb{R}^n} f(x)$. Suppose also that $\|g_k\|^2 \geq 2c(f_k - f_{\inf})$.

Conservatism of contemporary analyses

Suppose $g := \nabla f$ is Lipschitz continuous with constant $L > 0$. Then,

$$f_{k+1} \leq f_k + g_k^T s_k + \tfrac{1}{2} L \|s_k\|^2.$$

Let $f_{\inf} := \min_{x \in \mathbb{R}^n} f(x)$. Suppose also that $\|g_k\|^2 \geq 2c(f_k - f_{\inf})$.

$$f_k - f_{k+1} \geq \frac{1}{2L}\|g_k\|^2$$

$$f_0 - f_{\inf} \geq \frac{1}{2L}|\mathcal{K}_g|\epsilon_g^2$$

$$|\mathcal{K}_g| \leq \mathcal{O}\left(\frac{f_0 - f_{\inf}}{\epsilon_g^2}\right)$$

$$\begin{aligned} f_k - f_{k+1} &\geq \frac{1}{2L}\|g_k\|^2 \\ &\geq \frac{c}{L}(f_k - f_{\inf}) \end{aligned}$$

$$f_0 - f_{\inf} \geq \left(1 - \frac{c}{L}\right)^{-k}(f_k - f_{\inf})$$

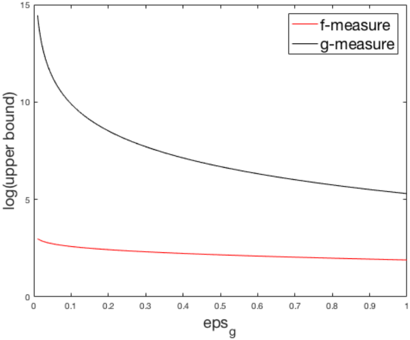
$$|\mathcal{K}_f| \leq \mathcal{O}\left(\log\left(\frac{f_0 - f_{\inf}}{\epsilon_f}\right)\right)$$

where

$$\mathcal{K}_g := \{k \in \mathbb{N} : \|g_k\| \geq \epsilon_g\} \text{ and } \mathcal{K}_f := \{k \in \mathbb{N} : f_k - f_{\inf} \geq \epsilon_f\}.$$

Upper bounds on $|\mathcal{K}_f|$ versus $|\mathcal{K}_g|$

Setting with $\{x \in \mathbb{R}^n : f_k - f_{\inf} \leq \epsilon_f\} = \{x \in \mathbb{R}^n : \|g_k\| \leq \epsilon_g\}$.



Worst-case examples

Worst-case performance bounds are tight; Cartis, Gould, Toint (2010).

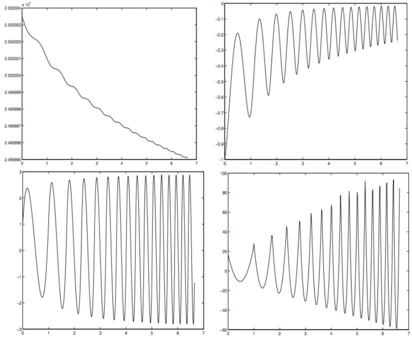


FIG. 2.1. *The function $f^{(1)}$ (top left) and its derivatives of order one (top right), two (bottom left), and three (bottom right) on the first* 16 *intervals.*

However, these examples for nonconvex optimization are. . . strange.

- Compared to convex optimization, for nonconvex. . .
- **there is a much wider gap between theory and practice.**

Motivation

We want a characterization strategy that

- attempts to capture behavior in actual practice
- i.e., is not “bogged down” by pedogogical examples
- **can be applied consistently across different classes of functions**

Motivation

We want a characterization strategy that

- attempts to capture behavior in actual practice
- i.e., is not "bogged down" by pedogogical examples
- **can be applied consistently across different classes of functions**

Our idea is to

- analyze how an algorithm behaves over different regions
- characterize an algorithm's behavior **by region**
- combine results to give complete perspectives on function classes

For some functions, there will be holes, but for many of interest there are none!

Motivation

We want a characterization strategy that

- attempts to capture behavior in actual practice
- i.e., is not “bogged down” by pedogogical examples
- **can be applied consistently across different classes of functions**

Our idea is to

- analyze how an algorithm behaves over different regions
- characterize an algorithm’s behavior **by region**
- combine results to give complete perspectives on function classes

For some functions, there will be holes, but for many of interest there are none!

We call this **regional complexity** analysis (RC analysis, for short).

Main idea

Algorithm

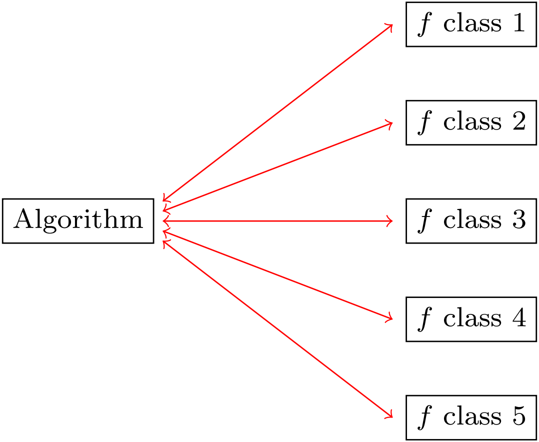
f class

Main idea

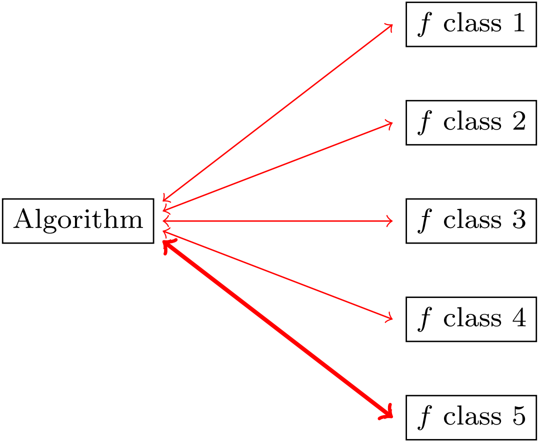
Algorithm

f class

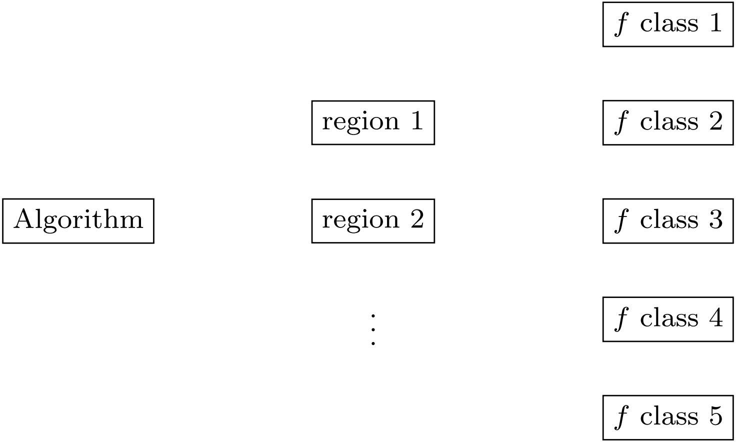
Main idea



Main idea



Main idea



Algorithm

region 1

region 2

⋮

f class 1

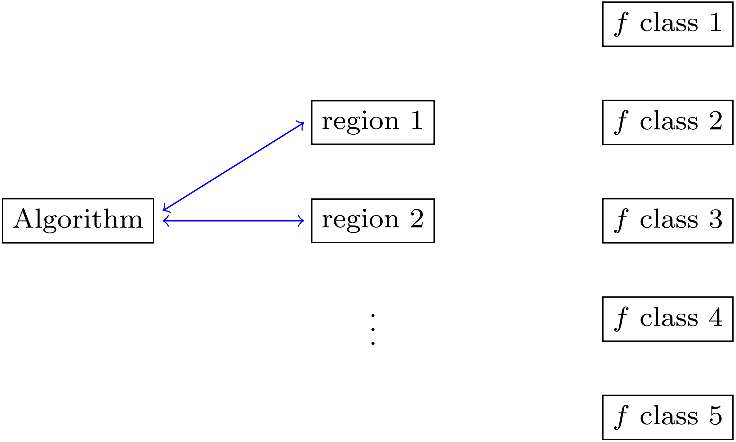
f class 2

f class 3

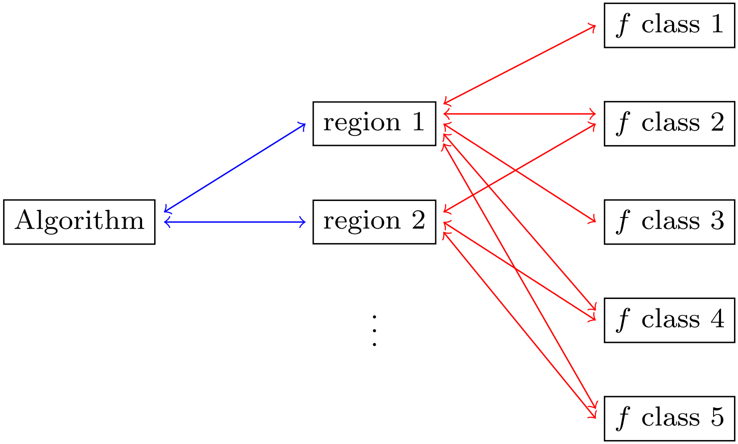
f class 4

f class 5

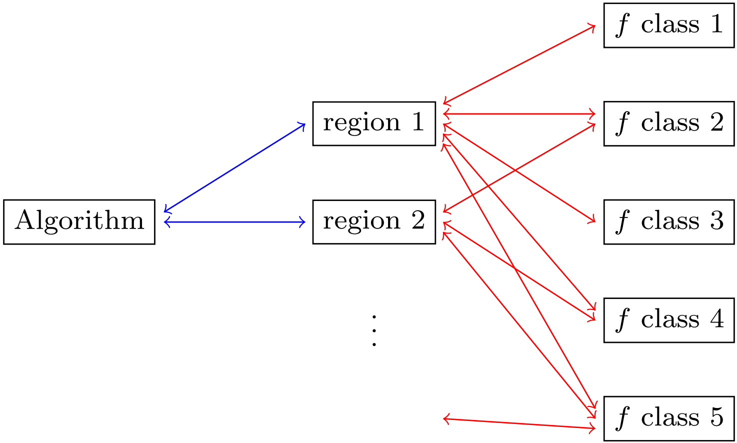
Main idea


f class 1
region 1
f class 2
Algorithm
region 2
f class 3
⋮
f class 4
f class 5

Main idea


Algorithm
region 1
region 2
⋮
f class 1
f class 2
f class 3
f class 4
f class 5

Main idea


Algorithm
region 1
region 2
⋮
f class 1
f class 2
f class 3
f class 4
f class 5

Outline

Introduction

Contemporary Analyses

Regional Complexity Analysis: Step 1

Regional Complexity Analysis: Step 2

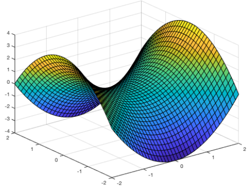
Summary & Perspectives

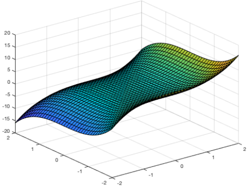
Intuition

Think about an arbitrary point in the search space, i.e.,

$$\mathcal{L} := \{x \in \mathbb{R}^n : f(x) \leq f_0\}.$$

- If $\|g_k\| \gg 0$, then “a lot” of progress can be made.
- If $\|g_k\| \approx 0$, but $\lambda(H_k) \ll 0$, then again “a lot” of progress can be made.





"Region 1"

Given $\kappa \in (0, \infty)$ and $f_{\rm ref} \in [f_{\rm inf}, f_0]$...

Let "Region 1" be points where the gradient is large relative to optimality error:

$$\mathcal{R}_1 := \{x \in \mathcal{L} : \|g(x)\|^\tau \geq \kappa(f(x) - f_{\rm ref}) \geq 0 \text{ for some } \tau \in [1,2]\}. \qquad (\star_1)$$

These are points with **gradient domination**.

- Let $\mathcal{R}_1^2$ be points where $(\star_1)$ holds with $\tau = 2$.
- Let $\mathcal{R}_1^1$ be all other points in $\mathcal{R}_1$.

“Region 1”

Given $\kappa \in (0, \infty)$ and $f_{\text{ref}} \in [f_{\inf}, f_0]\ldots$

Let “Region 1” be points where the gradient is large relative to optimality error:

$$\mathcal{R}_1 := \{x \in \mathcal{L} : \|g(x)\|^\tau \geq \kappa(f(x) - f_{\text{ref}}) \geq 0 \text{ for some } \tau \in [1, 2]\}. \qquad (\star_1)$$

These are points with **gradient domination**.

- Let $\mathcal{R}_1^2$ be points where $(\star_1)$ holds with $\tau = 2$.
- Let $\mathcal{R}_1^1$ be all other points in $\mathcal{R}_1$.

Theorem

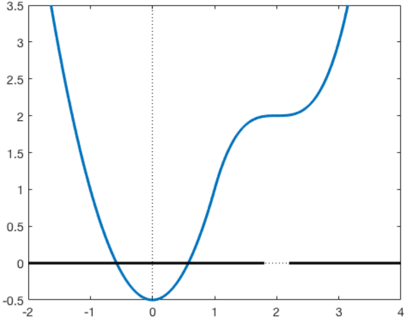
If f satisfies the Polyak-Łojasiewicz condition, then $\mathcal{R}_1^2 = \mathcal{R}_1 = \mathcal{L}$.

Theorem

If f is convex, then (over a ball containing its minimizers) $\mathcal{R}_1 = \mathcal{L}$.

Nesterov & Polyak (2006); Karimi, Nutini, and Schmidt (2016)

Region 1: Illustration



The set $\mathcal{R}_1$ (black line) covers almost the entire domain.

Example "Step 1–$\mathcal{R}_1$" result

Suppose ∇f is Lipschitz continuous with constant L_1.

Theorem

For a given algorithm, if $x_k \in \mathcal{R}_1$ implies that

$$f_k - f_{k+1} \geq \frac{1}{\zeta}\|g_k\|^2 \ \ \textit{for some} \ \ \zeta \in [L_1, \infty),$$

then the following hold.

(a) *If $x_k \in \mathcal{R}_1^2$, then (as in a **linear** rate)*

$$f_{k+1} - f_{\rm ref} \leq \left(1 - \frac{\kappa}{\zeta}\right)(f_k - f_{\rm ref}) \ \ \textit{where} \ \ \frac{\kappa}{\zeta} \in (0, 1].$$

(b) *If $x_k \in \mathcal{R}_1^1$, then (as in a **sublinear** rate)*

$$f_{k+1} - f_{\rm ref} \leq \left(1 - \frac{\kappa^2}{\zeta}(f_k - f_{\rm ref})\right)(f_k - f_{\rm ref}).$$

Characterizes gradient descent methods, second-order trust region methods, etc.

Example "Step 1–$\mathcal{R}_1$" result

Theorem

For a given algorithm, if $x_{k+1} \in \mathcal{R}_1$ implies that

$$f_k - f_{k+1} \geq \frac{1}{\zeta}\|g_{k+1}\|^{3/2} \quad \textit{for some} \quad \zeta \in (0,\infty),$$

then the following hold.

(a) *If $x_{k+1} \in \mathcal{R}_1^2$ and $f_k - f_{\text{ref}} \geq \kappa^3/\zeta^4$, then (as in a **linear** rate)*

$$f_{k+1} - f_{\text{ref}} \leq \left(\frac{(f_0 - f_{\text{ref}})^{1/4}}{\frac{\kappa^{3/4}}{\zeta} + (f_0 - f_{\text{ref}})^{1/4}}\right)(f_k - f_{\text{ref}})$$

*whereas, if $x_{k+1} \in \mathcal{R}_1^2$ and $f_k - f_{\text{ref}} < \kappa^3/\zeta^4$, then (as in a **superlinear** rate)*

$$f_{k+1} - f_{\text{ref}} \leq \left(\frac{\zeta^4(f_k - f_{\text{ref}})}{\kappa^3}\right)^{1/3}(f_k - f_{\text{ref}}).$$

Characterizes regularized Newton methods, etc.

Example "Step 1–$\mathcal{R}_1$" result

Theorem

(b) *If $x_{k+1} \in \mathcal{R}_1^1$ and $f_k - f_{\text{ref}} \geq \zeta^2/\kappa^3$, then (as in a **superlinear** rate)*

$$f_{k+1} - f_{\text{ref}} \leq \left(\frac{\zeta^2}{\kappa^3(f_k - f_{\text{ref}})}\right)^{1/3}(f_k - f_{\text{ref}}),$$

*whereas, if $x_{k+1} \in \mathcal{R}_1^1$ and $f_k - f_{\text{ref}} < \zeta^2/\kappa^3$, then (as in a **sublinear** rate)*

$$f_{k+1} - f_{\text{ref}} \leq \left(\frac{1}{1 + \frac{\kappa^{3/2}}{\zeta}\left(\frac{\sqrt{2}-1}{\sqrt{2}}\right)\sqrt{f_k - f_{\text{ref}}}}\right)^2(f_k - f_{\text{ref}}).$$

“Region 2”

Let “Region 2” be points not in $\mathcal{R}_1$ where negative curvature is large:

$$\mathcal{R}_2 := \{x \in \mathcal{L} : (\lambda(H(x)))_-^\tau \geq \kappa(f(x) - f_{\text{ref}}) \geq 0 \ \text{ for some } \tau \in [1,3]\} \setminus \mathcal{R}_1. \quad (\star_2)$$

These are points with **negative curvature domination**.

- Let $\mathcal{R}_2^3$ be points where $(\star_2)$ holds with $\tau = 3$.
- Let $\mathcal{R}_2^2$ be points where $(\star_2)$ holds with $\tau = 2$ (but not $\tau = 3$).
- Let $\mathcal{R}_2^1$ be all other points in $(\star_2)$.

"Region 2"

Let "Region 2" be points not in $\mathcal{R}_1$ where negative curvature is large:

$$\mathcal{R}_2 := \{x \in \mathcal{L} : (\lambda(H(x)))_-^\tau \geq \kappa(f(x) - f_{\rm ref}) \geq 0 \text{ for some } \tau \in [1,3]\} \setminus \mathcal{R}_1. \quad (\star_2)$$

These are points with **negative curvature domination**.

- Let $\mathcal{R}_2^3$ be points where $(\star_2)$ holds with $\tau = 3$.
- Let $\mathcal{R}_2^2$ be points where $(\star_2)$ holds with $\tau = 2$ (but not $\tau = 3$).
- Let $\mathcal{R}_2^1$ be all other points in $(\star_2)$.

Theorem

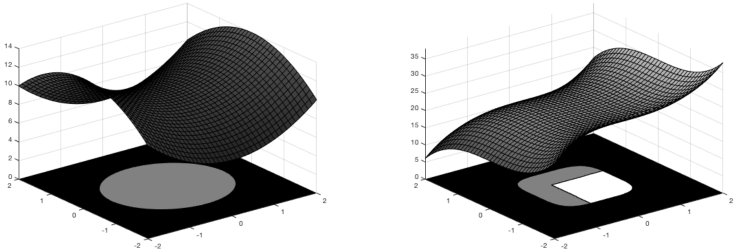
If f has the property that, for some $\kappa \in (0,\infty)$, one has

$$\max\{\|\nabla f(x)\|^2, -\lambda(\nabla^2 f(x))^3\} \geq \kappa(f(x) - f_{\inf}),$$

then $\mathcal{R}_1 \cup \mathcal{R}_2 = \mathcal{R}_1^2 \cup \mathcal{R}_2^3 = \mathcal{L}$.

This is a strict-saddle property.

Illustration



$(\bar{p} = 2)$ $\mathcal{R}_1$: black $\mathcal{R}_2$: gray $\overline{\mathcal{R}}$: white

Example "Step 1–$\mathcal{R}_2$" result

Suppose ∇f and $\nabla^2 f$ are Lipschitz continuous with constants L_1 and L_2.

Theorem

For a given algorithm, if $x_k \in \mathcal{R}_2$ implies that

$$f_k - f_{k+1} \geq \frac{1}{\zeta}(\lambda(H_k))_-^3 \quad \textit{for some} \quad \zeta \in [L_2, \infty),$$

then the following hold.

(a) *If $x_k \in \mathcal{R}_2^3$, then (as in a* ***linear*** *rate)*

$$f_{k+1} - f_{\rm ref} \leq \left(1 - \frac{\kappa}{\zeta}\right)(f_k - f_{\rm ref}) \quad \textit{where} \quad \frac{\kappa}{\zeta} \in (0, 1].$$

(b) *If $x_k \in \mathcal{R}_2^2$, then (as in a* ***sublinear*** *rate)*

$$f_{k+1} - f_{\rm ref} \leq \left(1 - \left(\frac{\kappa^{3/2}}{\zeta}\right)\sqrt{f_k - f_{\rm ref}}\right)(f_k - f_{\rm ref}).$$

(c) *If $x_k \in \mathcal{R}_2^1$, then (as in a* ***sublinear*** *rate)*

$$f_{k+1} - f_{\rm ref} \leq \left(1 - \frac{\kappa^3}{\zeta}(f_k - f_{\rm ref})^2\right)(f_k - f_{\rm ref}).$$

Characterizes (some!) second-order trust region methods, regularized Newton, etc.

Higher-order regions

This can be extended in a natural way to higher-order regions/algorithms.

If f is $\bar{p}$-times continuously differentiable, then we have the regions

$$\mathcal{R}_1 := \{x \in \mathcal{L} : \Delta_1(x)^\tau \geq \kappa(f(x) - f_{\inf}) \geq 0 \text{ for some } \tau \in [1,2]\},$$

$$\mathcal{R}_p := \{x \in \mathcal{L} : \Delta_2(x)^\tau \geq \kappa(f(x) - f_{\inf}) \geq 0 \text{ for some } \tau \in [1,p+1]\} \setminus \left(\bigcup_{j=1}^{p-1} \mathcal{R}_j\right)$$

$$\text{for all } \ p \in \{2,\dots,\bar{p}\},$$

$$\text{and } \ \overline{\mathcal{R}} := \mathcal{L} \setminus \left(\bigcup_{j=1}^{\bar{p}} \mathcal{R}_j\right).$$

Regions could be defined in other ways as well; key idea is to partition!

Outline

Introduction

Contemporary Analyses

Regional Complexity Analysis: Step 1

Regional Complexity Analysis: Step 2

Summary & Perspectives

Function classes

Definition ((g, H)-dominated function of degree (τ_1, τ_2))

A twice continuously differentiable function f is (g, H)-dominated of degree $(\tau_1, \tau_2) \in [1, 2] \times [1, 3]$ over $\mathcal{L}$ if for some constant $\kappa \in (0, \min\{L_1, L_2\}]$

$$\max\{\|g(x)\|^{\tau_1}, (\lambda(H(x)))_-^{\tau_2}\} \geq \kappa(f(x) - f_{\inf}) \quad \textit{for all} \quad x \in \mathcal{L}.$$

Definition (gradient-dominated function of degree τ)

A continuously differentiable function f is gradient-dominated of degree $\tau \in [1, 2]$ over $\mathcal{L}$ if for some constant $\kappa \in (0, L_1]$ it holds that

$$\|g(x)\|^{\tau} \geq \kappa(f(x) - f_{\inf}) \quad \textit{for all} \quad x \in \mathcal{L}.$$

Note: gradient-dominated $\implies$ (g, H)-dominated, but not vice versa

“Step 2” result

The following applies if f is (g, H)-dominated of degree $(2, 3)$.

Theorem

If $x_k \in \mathcal{R}_1^2 \cup \mathcal{R}_2^3$ *for all* $k \in \mathbb{N}$, *then*

$$\text{2nd-order TR}: \quad \text{linear} \implies \text{quadratic(?)}$$

$$|\{k : f_k - f_{\inf} > \epsilon\}| = \mathcal{O}\left(\log\left(\frac{f_0 - f_{\inf}}{\epsilon}\right)\right)$$

“Step 2” result

The following applies if f is (g, H)-dominated of degree $(2, 3)$.

Theorem

If $x_k \in \mathcal{R}_1^2 \cup \mathcal{R}_2^3$ *for all* $k \in \mathbb{N}$*, then*

$$\text{2nd-order TR} : \quad \text{linear} \implies \text{quadratic(?)}$$

$$|\{k : f_k - f_{\text{inf}} > \epsilon\}| = \mathcal{O}\left(\log\left(\frac{f_0 - f_{\text{inf}}}{\epsilon}\right)\right)$$

whereas

$$\text{regularized Newton} : \quad \text{linear} \implies \text{superlinear} \implies \text{quadratic(?)}.$$

$$|\{k : f_k - f_{\text{inf}} > \epsilon\}| = \mathcal{O}\left(\log\left(\frac{f_0 - f_{\text{inf}}}{\kappa^3/\zeta^4}\right)\right) + \mathcal{O}\left(\log\left(\log\left(\frac{\kappa^3/\zeta^4}{\epsilon}\right)\right)\right)$$

“Step 2” result

The following applies if f is (g, H)-dominated of degree $(2, 3)$.

Theorem

If $x_k \in \mathcal{R}_1^2 \cup \mathcal{R}_2^3$ *for all* $k \in \mathbb{N}$, *then*

$$\text{2nd-order TR}: \quad \text{linear} \implies \text{quadratic(?)}$$

$$|\{k : f_k - f_{\inf} > \epsilon\}| = \mathcal{O}\left(\log\left(\frac{f_0 - f_{\inf}}{\epsilon}\right)\right)$$

whereas

$$\text{regularized Newton}: \quad \text{linear} \implies \text{superlinear} \implies \text{quadratic(?).}$$

$$|\{k : f_k - f_{\inf} > \epsilon\}| = \mathcal{O}\left(\log\left(\frac{f_0 - f_{\inf}}{\kappa^3/\zeta^4}\right)\right) + \mathcal{O}\left(\log\left(\log\left(\frac{\kappa^3/\zeta^4}{\epsilon}\right)\right)\right)$$

(Compare with $|\{k : \|g_k\| > \epsilon_g\}| = \mathcal{O}(\epsilon_g^{-2})$ and $\mathcal{O}(\epsilon_g^{-3/2})$, respectively.)

"Step 2" result

The following holds if f is (g, H)-dominated of degree $(1, 1)$.

Theorem

If $x_k \in \mathcal{R}_1 \cup \mathcal{R}_2$ for all $k \in \mathbb{N}$, then the following hold.

- *$x_k \in \mathcal{R}_1^2 \cup \mathcal{R}_2^3$ for all large $k \in \mathbb{N}$, same as previous result.*
- *$x_k \in \mathcal{R}_1^2 \cup (\mathcal{R}_2^2 \cup \mathcal{R}_2^3)$ for all large $k \in \mathbb{N}$, then*

  2nd-order TR and regularized Newton: linear $\Longrightarrow$ superlinear

  $$\mathcal{O}\left(\log\left(\frac{f_0 - f_{\inf}}{1/\kappa}\right)\right) + \mathcal{O}\left(\frac{1/\kappa}{\sqrt{\epsilon}}\right)$$

- *$x_k \in (\mathcal{R}_1^1 \cup \mathcal{R}_1^2) \cup (\mathcal{R}_2^2 \cup \mathcal{R}_2^3)$ for all large $k \in \mathbb{N}$, then*

  2nd-order TR and regularized Newton: linear $\Longrightarrow$ superlinear

  $$\mathcal{O}\left(\log\left(\frac{f_0 - f_{\inf}}{1/\kappa}\right)\right) \quad \underbrace{+\mathcal{O}\left(\frac{1/\kappa}{\epsilon}\right)}_{\text{2nd-order TR}} \quad vs. \quad \underbrace{+\mathcal{O}\left(\frac{1/\kappa}{\sqrt{\epsilon}}\right)}_{\text{regularized Newton}}$$

- *$x_k \in \mathcal{R}_2^1$ for infinite number of $k \in \mathbb{N}$, then*

  2nd-order TR and regularized Newton: linear $\Longrightarrow$ superlinear

  $$\mathcal{O}\left(\log\left(\frac{f_0 - f_{\inf}}{1/\kappa}\right)\right) + \mathcal{O}\left(\frac{1/\kappa}{\epsilon^2}\right)$$

Outline

Introduction

Contemporary Analyses

Regional Complexity Analysis: Step 1

Regional Complexity Analysis: Step 2

Summary & Perspectives

Summary & Perspectives

Our goal: A **complementary** approach to characterize algorithms.

- global convergence
- worst-case complexity, contemporary type + **RC analysis**
- local convergence rate

Our idea is to

- analyze how an algorithm behaves over different regions
- characterize an algorithm's behavior **by region**
- combine results to give complete perspectives on function classes

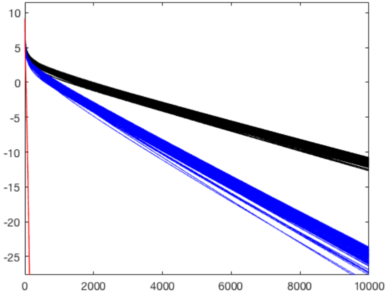
For some functions, there are holes, but for others the characterization is complete.

F. E. Curtis and D. P. Robinson, "How to Characterize the Worst-Case Performance of Algorithms for Nonconvex Optimization," Lehigh ISE/COR@L Technical Report 18T-003, February 3, 2018. *New version coming soon.*

Back to take-home message #2

Strongly convex quadratic

- gradient descent with a fixed stepsize (black)
- gradient descent with adaptive stepsizes / line searches (blue)
- conjugate gradient with adaptive stepsizes (red)



Focus on worst-case performance...

- is a **self-fulfilling prophecy**!
- Let's emphasize worst-case performance less when actual behavior is better!