Characterizing Worst-Case Complexity of Algorithms for Nonconvex Optimization

Frank E. Curtis, Lehigh University

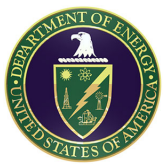
joint work with

Daniel P. Robinson, Johns Hopkins University

presented at

International Symposium on Mathematical Programming (ISMP)
Bordeaux, France

3 July 2018



Outline

Outline

Raising awareness

Issues that I believe nonlinear optimizers need to address:

Raising awareness

Issues that I believe nonlinear optimizers need to address:

State-of-the-art nonlinear optimization codes fail too often.

- Reasons are "high" nonlinearity, degeneracy, and infeasibility.
- People have disputed this, but I have results!
- (We'll never know the number of users that we've lost.)

Raising awareness

Issues that I believe nonlinear optimizers need to address:

State-of-the-art nonlinear optimization codes fail too often.

- Reasons are "high" nonlinearity, degeneracy, and infeasibility.
- People have disputed this, but I have results!
- (We'll never know the number of users that we've lost.)

Our worst-case analysis for nonconvex optimization is faulty.

- 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

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

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

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

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

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

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

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

where x_* is some global minimizer of f.

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

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

where x_* is some global minimizer of f.

. . . 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!

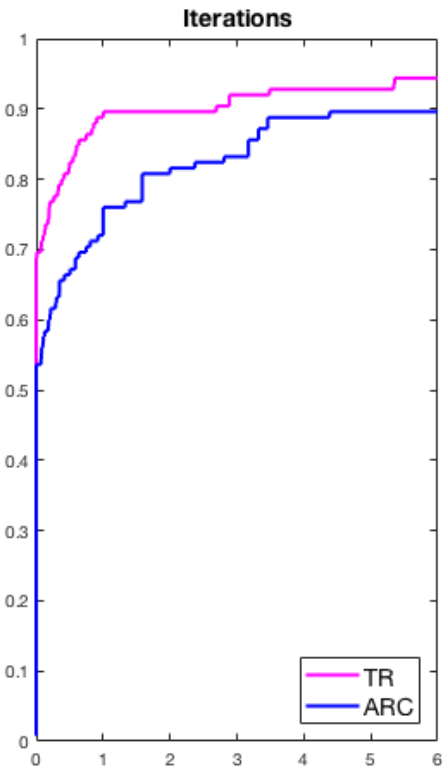
TR vs. ARC

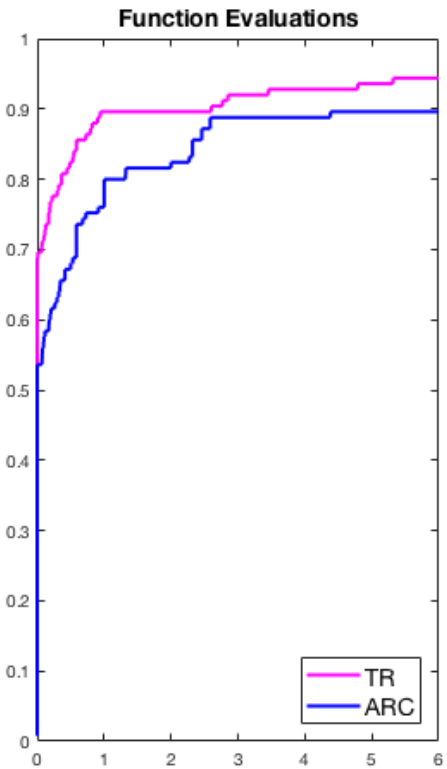
TR		ARC
1: Solve to compute s_k: $\min_{s\in\mathbb{R}^n} q_k(s) := f_k + g_k^T s + \frac{1}{2}s^T H_k s$ s.t. $\|s\| \le \delta_k$ (dual: λ_k)	$\sigma_k = \frac{\lambda_k}{\delta_k}$ →	1: Solve to compute s_k: $\min_{s\in\mathbb{R}^n} c_k(s) := f_k + g_k^T s + \frac{1}{2}s^T H_k s + \frac{1}{3}\sigma_k\|s\|^3$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

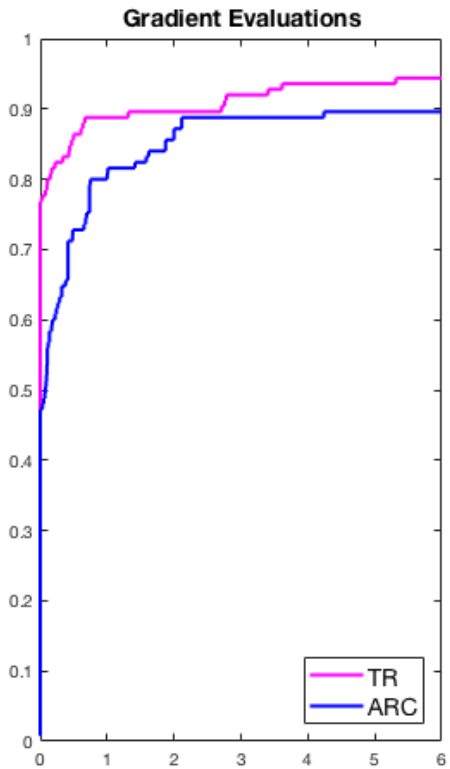
TR vs. ARC

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. $\|s\| \le \delta_k$ (dual: λ_k)	$\sigma_k = \frac{\lambda_k}{\delta_k}$ →	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: $\rho_k^q \leftarrow \frac{f_k - f(x_k+s_k)}{f_k - q_k(s_k)}$	← $\delta_k = \|s_k\|$	2: Compute ratio: $\rho_k^c \leftarrow \frac{f_k - f(x_k+s_k)}{f_k - c_k(s_k)}$
3: Update radius: $\rho_k^q \ge \eta$: accept and $\delta_k \nearrow$ $\rho_k^q < \eta$: reject and $\delta_k \searrow$		3: Update regularization: $\rho_k^c \ge \eta$: 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.

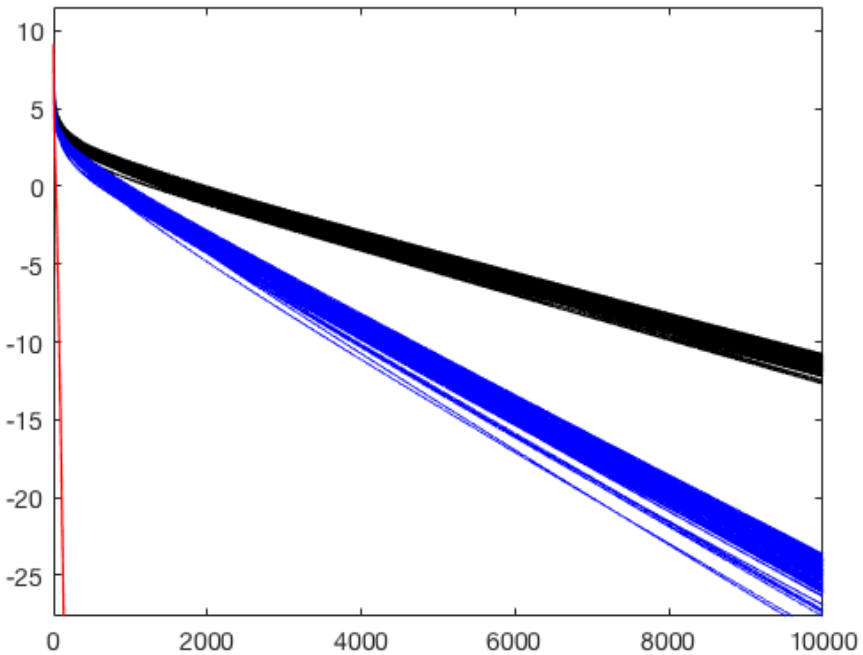
- global convergence
- worst-case complexity, contemporary type + **our new approach**
- local convergence rate

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

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

Complexity: Take-home message #2

Ideally, we would weigh worst-case analyses differently depending on the category of method. Some methods actually behave like their worst-case; others don't.



Let's return to this at the end of the talk. In short:

- **focus on worst-case analysis can be a self-fulfilling prophecy**

Outline

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

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

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

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

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

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

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

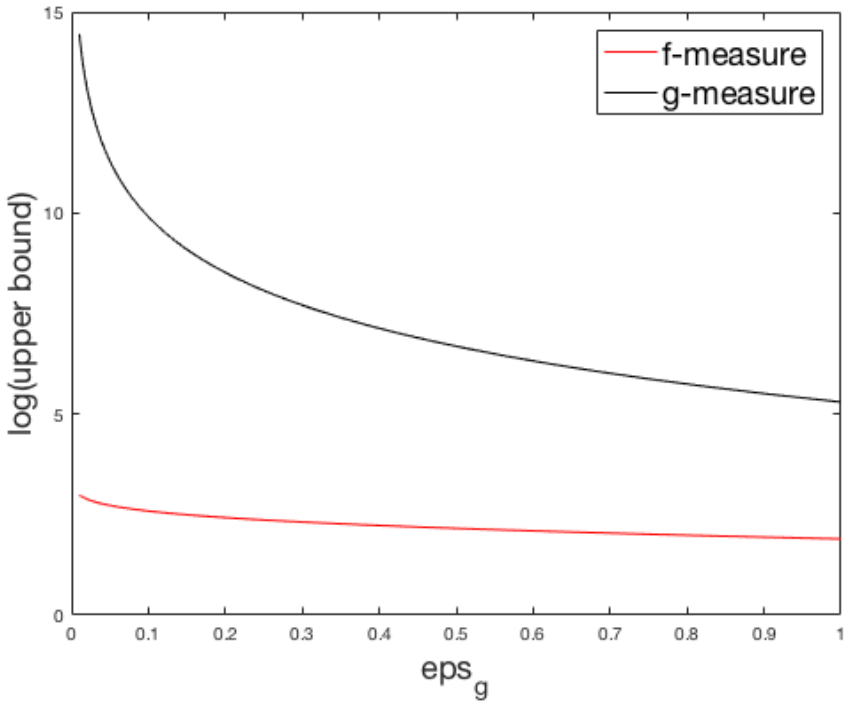
$$|\mathcal{K}_f| \leq \mathcal{O}\left(\log\left(\frac{f_0 - f_{\text{inf}}}{\epsilon_f} \right) \right)$$

where

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

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

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



Worst-case examples

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

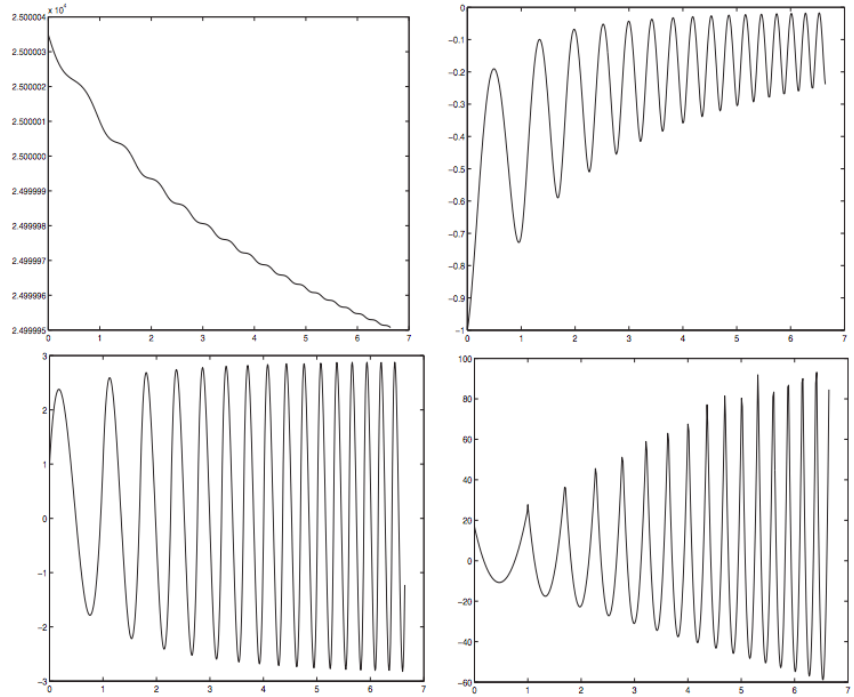


FIG. 2.1. *The function $f^{(1)}$ (top left) and its derivatives of order one (top right), two (bottom left), and three (bottom right) on the first* 16 *intervals.*

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

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

Outline

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

- partition the search space (dependent on f and x_0)
- analyze how an algorithm behaves over different regions
- characterize an algorithm's behavior **by region**

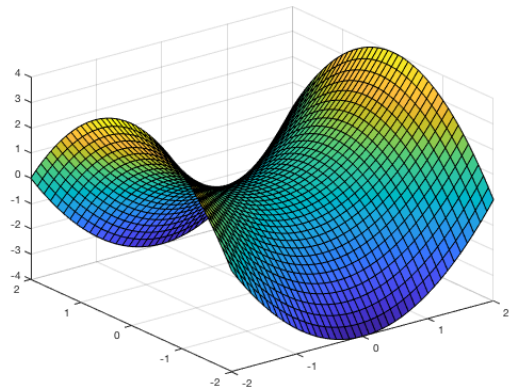
For some functions, there will be holes, but for many of interest there are none!

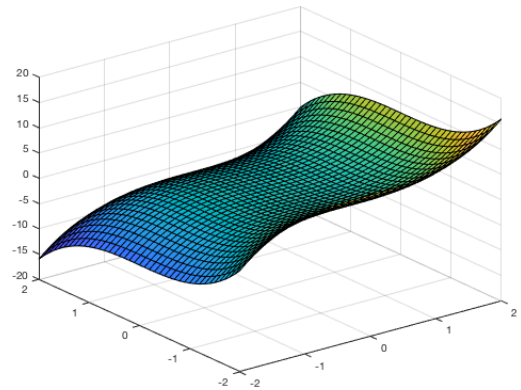
Intuition

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"

Definition (Reduction in a first-order model)

At a given point $x \in \mathcal{L}$, consider the model

$$m_1(x,s) = \nabla f(x)^T s + \tfrac{1}{2} r_1 \|s\|^2$$

Letting $s_{m_1}(x) := \arg\min_{s \in \mathbb{R}^n} m_1(x,s)$, the reduction in this model from x is

$$\Delta m_1(x) = m_1(x,0) - m_1(x, s_{m_1}(x)) = \tfrac{1}{2r_1} \|\nabla f(x)\|^2. \qquad (\star)$$

"Region 1"

Definition (Reduction in a first-order model)

At a given point $x \in \mathcal{L}$, consider the model

$$m_1(x,s) = \nabla f(x)^T s + \tfrac{1}{2} r_1 \|s\|^2$$

Letting $s_{m_1}(x) := \arg\min_{s \in \mathbb{R}^n} m_1(x,s)$, the reduction in this model from x is

$$\Delta m_1(x) = m_1(x,0) - m_1(x, s_{m_1}(x)) = \tfrac{1}{2r_1} \|\nabla f(x)\|^2. \qquad (\star)$$

Let "Region 1" be those points where this reduction is sufficiently large:

$$\mathcal{R}_1 := \{x \in \mathcal{L} : \Delta m_1(x) \geq \kappa(f(x) - f_{\text{inf}})\}.$$

Noting $(\star)$, these are "**big gradient**" points.

"Region 1"

Definition (Reduction in a first-order model)

At a given point $x \in \mathcal{L}$, consider the model

$$m_1(x,s) = \nabla f(x)^T s + \tfrac{1}{2} r_1 \|s\|^2$$

Letting $s_{m_1}(x) := \arg\min_{s \in \mathbb{R}^n} m_1(x,s)$, the reduction in this model from x is

$$\Delta m_1(x) = m_1(x,0) - m_1(x, s_{m_1}(x)) = \tfrac{1}{2r_1} \|\nabla f(x)\|^2. \qquad (\star)$$

Let "Region 1" be those points where this reduction is sufficiently large:

$$\mathcal{R}_1 := \{x \in \mathcal{L} : \Delta m_1(x) \geq \kappa(f(x) - f_{\text{inf}})\}.$$

Noting $(\star)$, these are "**big gradient**" points.

Theorem

A continuously differentiable f with a Lipschitz continuous gradient satisfies the Polyak-Łojasiewicz condition if and only if $\mathcal{R}_1 = \mathcal{L}$.

"Region 2"

Definition (Reduction in a second-order model)

At a given point $x \in \mathbb{R}^n$, consider the (non-Taylor-like) model

$$m_2(x,s) = \tfrac{1}{2} s^T \nabla^2 f(x) s + \tfrac{1}{3} r_2 \|s\|^3$$

Letting $s_{m_2}(x) := \arg\min_{s \in \mathbb{R}^n} m_2(x,s)$, the reduction in this model from x is

$$\Delta m_2(x) = m_2(x,0) - m_2(x, s_{m_2}(x)) = \tfrac{1}{6{r_2}^2} \max\{-\lambda(\nabla^2 f(x)), 0\}^3. \qquad (\star\star)$$

"Region 2"

Definition (Reduction in a second-order model)

At a given point $x \in \mathbb{R}^n$, consider the (non-Taylor-like) model

$$m_2(x,s) = \tfrac{1}{2} s^T \nabla^2 f(x) s + \tfrac{1}{3} r_2 \|s\|^3$$

Letting $s_{m_2}(x) := \arg\min_{s \in \mathbb{R}^n} m_2(x,s)$, the reduction in this model from x is

$$\Delta m_2(x) = m_2(x,0) - m_2(x, s_{m_2}(x)) = \tfrac{1}{6 {r_2}^2} \max\{-\lambda(\nabla^2 f(x)), 0\}^3. \qquad (\star\star)$$

Let "Region 2" be those points **not in** $\mathcal{R}_1$ where this reduction is sufficiently large:

$$\mathcal{R}_2 := \{x \in \mathbb{R}^n : \Delta m_2(x) \geq \kappa(f(x) - f_{\inf})\} \setminus \mathcal{R}_1.$$

Noting $(\star\star)$, these are "**very negative curvature**" points.

"Region 2"

Definition (Reduction in a second-order model)

At a given point $x \in \mathbb{R}^n$, consider the (non-Taylor-like) model

$$m_2(x,s) = \tfrac{1}{2} s^T \nabla^2 f(x) s + \tfrac{1}{3} r_2 \|s\|^3$$

Letting $s_{m_2}(x) := \arg\min_{s \in \mathbb{R}^n} m_2(x,s)$, the reduction in this model from x is

$$\Delta m_2(x) = m_2(x,0) - m_2(x, s_{m_2}(x)) = \tfrac{1}{6 r_2{}^2} \max\{-\lambda(\nabla^2 f(x)), 0\}^3. \qquad (\star\star)$$

Let "Region 2" be those points **not in** $\mathcal{R}_1$ where this reduction is sufficiently large:

$$\mathcal{R}_2 := \{x \in \mathbb{R}^n : \Delta m_2(x) \geq \kappa(f(x) - f_{\inf})\} \setminus \mathcal{R}_1.$$

Noting $(\star\star)$, these are "**very negative curvature**" points.

Theorem

If f is twice-continuously differentiable with Lipschitz continuous gradient and Hessian functions such that, at all $x \in \mathcal{L}$ and for some $\zeta \in (0, \infty)$, one has

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

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

Regions

This can be extended in a natural way to higher-order models.

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

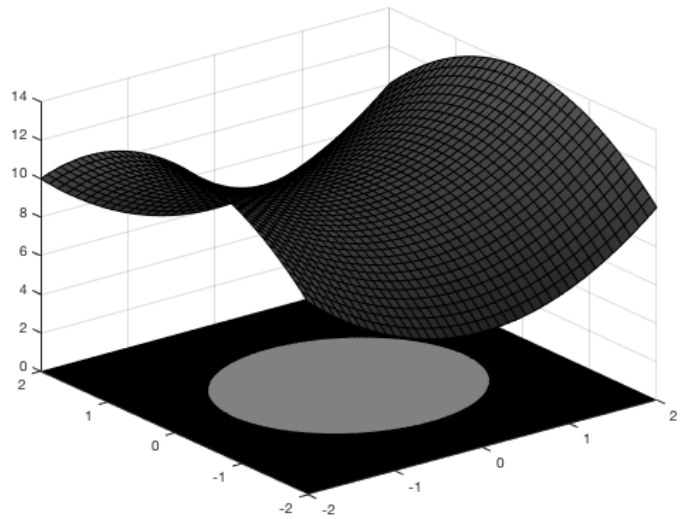
$$\mathcal{R}_1 := \{x \in \mathbb{R}^n : \Delta m_1(x) \geq \kappa(f(x) - f_{\text{inf}})\},$$

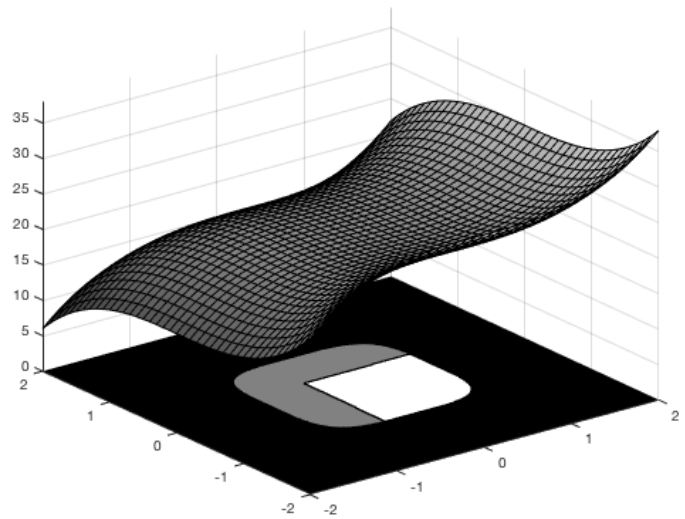
$$\mathcal{R}_p := \{x \in \mathbb{R}^n : \Delta m_p(x) \geq \kappa(f(x) - f_{\text{inf}})\} \setminus \left(\bigcup_{j=1}^{p-1} \mathcal{R}_j \right) \quad \text{for all} \quad p \in \{2, \ldots, \bar{p}\},$$

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!

Illustration





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

Outline

Regularized gradient and Newton methods

- Regularized gradient (RG) method: Computes s_k by solving

$$\min_{s\in\mathbb{R}^n} \; f_k + g_k^T s + \frac{l_1}{2}\|s\|^2 \quad \Longrightarrow \quad s_k = -\frac{1}{l_1} g_k$$

- Regularized Newton (RN) method: Computes s_k by solving

$$\min_{s\in\mathbb{R}^n} \; f_k + g_k^T s + \frac{1}{2} s^T H_k s + \frac{l_2}{3}\|s\|^3,$$

  also known as cubic regularization

Characterization: Contemporary

Theorem (Contemporary complexity result)

With $\bar{p} \geq 2$, let

$$\mathcal{K}_g := \{k \in \mathbb{N} : \|g_k\| \geq \epsilon_g\}$$
$$\text{and} \quad \mathcal{K}_H := \{k \in \mathbb{N} : H_k \succeq -\epsilon_H I\}.$$

Then, the cardinalities of $\mathcal{K}_g$ and $\mathcal{K}_H$ are of the order...

Algorithm	$\lvert\mathcal{K}_g\rvert$	$\lvert\mathcal{K}_H\rvert$
RG	$\mathcal{O}\left(\frac{l_1(f_0 - f_{inf})}{\epsilon_g^2}\right)$	$\infty^\star$
RN	$\mathcal{O}\left(\frac{l_2^{1/2}(f_0 - f_{inf})}{\epsilon_g^{3/2}}\right)$	$\mathcal{O}\left(\frac{l_2^2(f_0 - f_{inf})}{\epsilon_H^3}\right)$

*Could be better with a probabilistic analysis.

Characterization: Our approach

Theorem (New complexity result)

The numbers of iterations in $\mathcal{R}_1$ and $\mathcal{R}_2$ are of the order...

Algorithm	$\mathcal{R}_1$	$\mathcal{R}_2$
RG	$\mathcal{O}\left(\log\left(\frac{f_0 - f_{inf}}{\epsilon_f}\right)\right)$	$\infty^\star$
RN	$\mathcal{O}\left(\frac{l_2^2(f_0 - f_{inf})}{r_1^3}\right) + \mathcal{O}\left(\log\left(\frac{f_0 - f_{inf}}{\epsilon_f}\right)\right)$	$\mathcal{O}\left(\log\left(\frac{f_0 - f_{inf}}{\epsilon_f}\right)\right)$

*Could be better with a probabilistic analysis.

There is an initial phase, as seen in Nesterov & Polyak (2006)

Most interesting cases: Higher order method in lower-order region.

Trust region methods

$$\min_{s\in\mathbb{R}^n} \; f_k + g_k^T s + \tfrac{1}{2} s^T H_k s \quad \text{s.t.} \quad \|s\| \le \delta_k,$$

where

$$\text{(Case 1)} \qquad \delta_k \leftarrow \nu_k \|g_k\|;$$

$$\text{(Case 2)} \qquad \delta_k \leftarrow \nu_k \begin{cases} \|g_k\| & \text{if } \|g_k\|^2 \ge |\lambda(H_k)|^3 \\ |\lambda(H_k)| & \text{otherwise} \end{cases}$$

Trust region methods

$$\min_{s \in \mathbb{R}^n} \; f_k + g_k^T s + \tfrac{1}{2} s^T H_k s \quad \text{s.t.} \quad \|s\| \le \delta_k,$$

where

(Case 1) $\delta_k \leftarrow \nu_k \|g_k\|;$

(Case 2) $\delta_k \leftarrow \nu_k \begin{cases} \|g_k\| & \text{if } \|g_k\|^2 \ge |\lambda(H_k)|^3 \\ |\lambda(H_k)| & \text{otherwise} \end{cases}$

Trust region methods

$$\min_{s\in\mathbb{R}^n} \; f_k + g_k^T s + \tfrac{1}{2} s^T H_k s \quad \text{s.t.} \quad \|s\| \le \delta_k,$$

where

(Case 1) $\delta_k \leftarrow \nu_k \|g_k\|;$

(Case 2) $\delta_k \leftarrow \nu_k \begin{cases} \|g_k\| & \text{if } \|g_k\|^2 \ge |\lambda(H_k)|^3 \\ |\lambda(H_k)| & \text{otherwise} \end{cases}$

Theorem (Case 1)

# of iterations in $\mathcal{R}_1$ is at most $\mathcal{O}\left(\chi \log\left(\frac{f_0 - f_{inf}}{\epsilon_f}\right)\right)$. For $\mathcal{R}_2$, no guarantee.

Theorem (Case 2)

# of iterations in $\mathcal{R}_1$ is at most $\mathcal{O}\left(\chi \log\left(\frac{f_0 - f_{inf}}{\epsilon_f}\right)\right)$.

# of iterations in $\mathcal{R}_2$ is at most $\mathcal{O}\left(\bar{\chi} \log\left(\frac{f_0 - f_{inf}}{\epsilon_f}\right)\right)$.

pth-order method: Behavior over $\mathcal{R}_p$

Let $s_{w_p}(x)$ be a minimum norm global minimizer of the regularized Taylor model

$$w_p(x,s) = \text{“}p\text{th-order Taylor model”} + \frac{l_p}{p+1}\|s\|^{p+1}$$

Theorem

If $\{x_k\}$ *is generated by the iteration*

$$x_{k+1} \leftarrow x_k + s_{w_p}(x),$$

then, with $\epsilon_f \in (0, f_0 - f_{inf})$*, the number of iterations in*

$$\mathcal{R}_p \cap \{x \in \mathbb{R}^n : f(x) - f_{inf} \geq \epsilon_f\}$$

is bounded above by

$$\left\lceil \log\left(\frac{f_0 - f_{inf}}{\epsilon_f}\right)\left(\log\left(\frac{1}{1-\kappa}\right)\right)^{-1}\right\rceil = \mathcal{O}\left(\log\left(\frac{f_0 - f_{inf}}{\epsilon_f}\right)\right)$$

Outline

Summary & Perspectives

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

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

Our idea is to

- partition the search space (dependent on f and x_0)
- analyze how an algorithm behaves over different regions
- characterize an algorithm's behavior **by region**

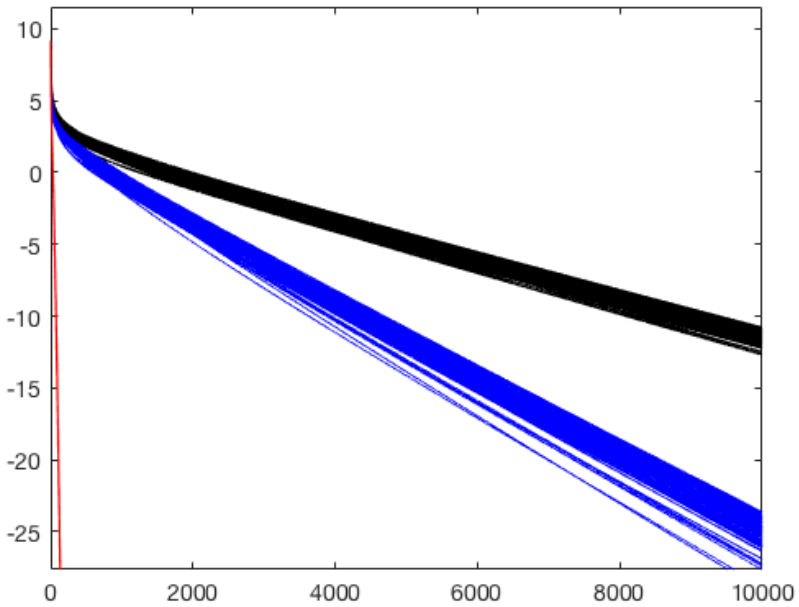
For some functions, there are holes, but for others the characterization is complete.

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.

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!