How to Characterize the Worst-Case Performance of Algorithms for Nonconvex Optimization

Frank E. Curtis, Lehigh University

joint work with

Daniel P. Robinson, Johns Hopkins University

U.S.-Mexico Workshop on Optimization and its Applications

8 January 2018

Thanks, Don!

Outline

Motivation

Contemporary Analyses

Partitioning the Search Space

Behavior of Regularization Methods

Summary & Perspectives

Outline

Motivation

Contemporary Analyses

Partitioning the Search Space

Behavior of Regularization Methods

Summary & Perspectives

History

Nonlinear optimization has had parallel developments

convexity	smoothness
Rockafellar Fenchel Nemirovski Nesterov	Powell Fletcher Goldfarb Nocedal
subgradient inequality	sufficient decrease
convergence, complexity guarantees	convergence, fast local convergence

Worlds are (finally) colliding!

Worst-case complexity for nonconvex optimization

Here is how we do it now:

Assuming Lipschitz continuity of derivatives...

...upper bound on # of iterations until $\|\nabla f(x_k)\|_2 \leq \epsilon$?

Gradient descent	Newton / trust region	Cubic regularization
$\mathcal{O}(\epsilon^{-2})$	$\mathcal{O}(\epsilon^{-2})$	$\mathcal{O}(\epsilon^{-3/2})$

Self-examination

But. . .

- Is this the best way to ***characterize*** our algorithms?
- Is this the best way to ***represent*** our algorithms?

Self-examination

But…

- Is this the best way to ***characterize*** our algorithms?
- Is this the best way to ***represent*** our algorithms?

People listen! Cubic regularization…

- Griewank (1981)
- Nesterov & Polyak (2006)
- Weiser, Deuflhard, Erdmann (2007)
- Cartis, Gould, Toint (2011), the `ARC` method

…is a framework to which researchers have been attracted…

- Agarwal, Allen-Zhu, Bullins, Hazan, Ma (2017)
- Carmon, Duchi (2017)
- Kohler, Lucchi (2017)
- Peng, Roosta-Khorasan, Mahoney (2017)

However, there remains a large gap between theory and practice!

Purpose of this talk

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

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

Purpose of this talk

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

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

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

But we believe it ***is*** useful!

Purpose of this talk

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

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

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

But we believe it ***is*** useful!

Nonconvexity is difficult in every sense!

- Can we accept a characterization strategy with some (literal) holes?
- Or should we be purists, even if we throw out the baby with the bathwater...

Outline

Motivation

Contemporary Analyses

Partitioning the Search Space

Behavior of Regularization Methods

Summary & Perspectives

Simple setting

Consider the iteration

$$x_{k+1} \leftarrow x_k - \frac{1}{L} g_k \quad \text{for all} \quad k \in \mathbb{N}.$$

A contemporary complexity analysis considers the set

$$\mathcal{G}(\epsilon_g) := \{x \in \mathbb{R}^n : \|g(x)\|_2 \leq \epsilon_g\}$$

and aims to find an upper bound on the cardinality of

$$\mathcal{K}_g(\epsilon_g) := \{k \in \mathbb{N} : x_k \notin \mathcal{G}(\epsilon_g)\}.$$

$g_k := \nabla f(x_k)$, $g := \nabla f$

Upper bound on $|\mathcal{K}_g(\epsilon_g)|$

Using $s_k = -\frac{1}{L} g_k$ and the upper bound

$$f_{k+1} \leq f_k + g_k^T s_k + \tfrac{1}{2} L \|s_k\|_2^2,$$

one finds with $f_{\inf} := \inf_{x \in \mathbb{R}^n} f(x)$ that

$$\begin{aligned} & & f_k - f_{k+1} &\geq \tfrac{1}{2L} \|g_k\|_2^2 \\ &\Longrightarrow & (f_0 - f_{\inf}) &\geq \tfrac{1}{2L} |\mathcal{K}_g(\epsilon_g)| \epsilon_g^2 \\ &\Longrightarrow & |\mathcal{K}_g(\epsilon_g)| &\leq 2L(f_0 - f_{\inf}) \epsilon_g^{-2}. \end{aligned}$$

"Nice" f

But what if f is "nice"?

...e.g., satisfying the Polyak-Łojasiewicz condition for $c \in (0, \infty)$, i.e.,

$$f(x) - f_{\inf} \leq \tfrac{1}{2c} \|g(x)\|_2^2 \ \text{ for all } \ x \in \mathbb{R}^n.$$

Now consider the set

$$\mathcal{F}(\epsilon_f) := \{x \in \mathbb{R}^n : f(x) - f_{\inf} \leq \epsilon_f\}$$

and consider an upper bound on the cardinality of

$$\mathcal{K}_f(\epsilon_f) := \{k \in \mathbb{N} : x_k \notin \mathcal{F}(\epsilon_f)\}.$$

Upper bound on $|\mathcal{K}_f(\epsilon_f)|$

Using $s_k = -\frac{1}{L} g_k$ and the upper bound

$$f_{k+1} \leq f_k + g_k^T s_k + \tfrac{1}{2} L \|s_k\|_2^2,$$

one finds that

$$\begin{aligned}
& & f_k - f_{k+1} &\geq \tfrac{1}{2L} \|g_k\|_2^2 \\
& & &\geq \tfrac{c}{L}(f_k - f_{\inf}) \\
\Longrightarrow & & (1 - \tfrac{c}{L})(f_k - f_{\inf}) &\geq f_{k+1} - f_{\inf} \\
\Longrightarrow & & (1 - \tfrac{c}{L})^k (f_0 - f_{\inf}) &\geq f_k - f_{\inf} \\
\Longrightarrow & & |\mathcal{K}_f(\epsilon_f)| &\leq \log\left(\frac{f_0 - f_{\inf}}{\epsilon_f}\right)\left(\log\left(\frac{L}{L-c}\right)\right)^{-1}.
\end{aligned}$$

For the first step. . .

In the "general nonconvex" analysis. . .

. . . the expected decrease for the first step is much more pessimistic:

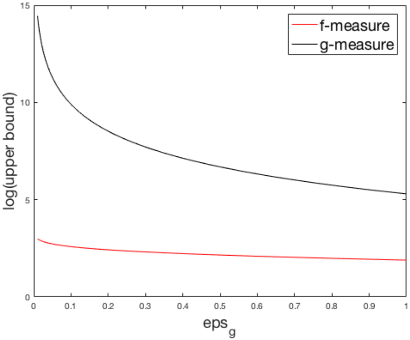
$$\begin{aligned} \text{general nonconvex:} \quad & f_0 - f_1 \geq \tfrac{1}{2L}\epsilon_g^2 \\ \text{PL condition:} \quad & (1 - \tfrac{c}{L})(f_0 - f_{\inf}) \geq f_1 - f_{\inf} \end{aligned}$$

. . . and it remains more pessimistic throughout!

Upper bounds on $|\mathcal{K}_f(\epsilon_f)|$ versus $|\mathcal{K}_g(\epsilon_g)|$

Let $f(x) = \frac{1}{2}x^2$, meaning that $g(x) = x$.

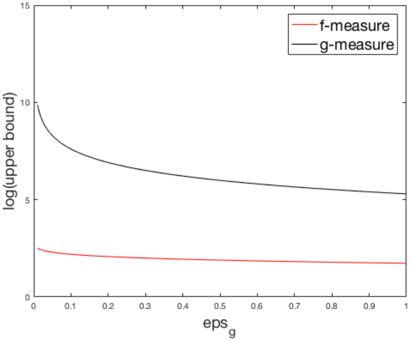
- Let $\epsilon_f = \frac{1}{2}\epsilon_g^2$, meaning that $\mathcal{F}(\epsilon_f) = \mathcal{G}(\epsilon_g)$.
- Let $x_0 = 10$, $c = 1$, and $L = 2$. (Similar pictures for any $L > 1$.)



Upper bounds on $|\mathcal{K}_f(\epsilon_f)|$ versus $|\{k \in \mathbb{N} : \frac{1}{2}\|g_k\|_2^2 > \epsilon_g\}|$

Let $f(x) = \frac{1}{2}x^2$, meaning that $\frac{1}{2}g(x)^2 = \frac{1}{2}x^2$.

- Let $\epsilon_f = \epsilon_g$, meaning that $\mathcal{F}(\epsilon_f) = \mathcal{G}(\epsilon_g)$.
- Let $x_0 = 10$, $c = 1$, and $L = 2$. (Similar pictures for any $L > 1$.)



Bad worst-case!

Worst-case complexity bounds in the general nonconvex case are very pessimistic.

- The analysis immediately admits a large gap when the function is nice.
- The "essentially tight" examples for the worst-case bounds are... weird.[1]

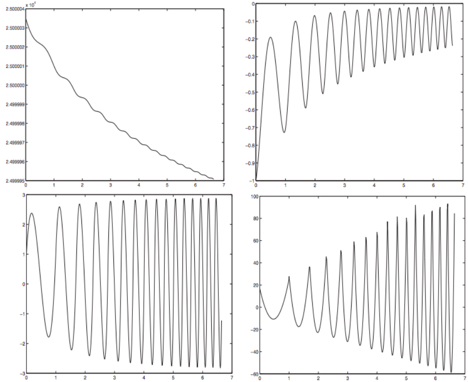


FIG. 2.1. *The function $f^{(1)}$ (top left) and its derivatives of order one (top right), two (bottom left), and three (bottom right) on the first 16 intervals.*

[1] Cartis, Gould, Toint (2010)

Plea

Let's not have these be the problems that dictate how we

- characterize our algorithms and
- represent our algorithms to the world!

Outline

Motivation

Contemporary Analyses

Partitioning the Search Space

Behavior of Regularization Methods

Summary & Perspectives

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
- shows more than just the worst of the worst case

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
- shows more than just the worst of the worst case

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 some of interest there are none!

Intuition

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

$$\mathcal{L} := \{x \in \mathbb{R}^n : f(x) \leq f(x_0)\}.$$

- If $\|g(x)\|_2 \gg 0$, then "a lot" of progress can be made.
- If $\min(\text{eig}(\nabla^2 f(x))) \ll 0$, then "a lot" of progress can also be made.

Assumption

Assumption 1

- *f is $\bar{p}$-times continuously differentiable*
- *f is bounded below by $f_{inf} := \inf_{x \in \mathbb{R}^n} f(x)$*
- *for all $p \in \{1, \ldots, \bar{p}\}$, there exists $L_p \in (0, \infty)$ such that*

$$f(x+s) \leq \underbrace{f(x) + \sum_{j=1}^{p} \frac{1}{j!} \nabla^j f(x)[s]^j}_{t_p(x,s)} + \frac{L_p}{p+1} \|s\|_2^{p+1}$$

pth-order term reduction

Definition 2

For each $p \in \{1, \dots, \bar{p}\}$, define the function

$$m_p(x, s) = \frac{1}{p!} \nabla^p f(x)[s]^p + \frac{r_p}{p+1} \|s\|_2^{p+1}.$$

Letting $s_{m_p}(x) := \arg\min_{s \in \mathbb{R}^n}$, the *reduction in the pth-order term from x* is

$$\Delta m_p(x) = m_p(x, 0) - m_p(x, s_{m_p}(x)) \geq 0.$$

*Exact definition of r_p is not complicated, but we'll skip it here

Regions

We propose to partition the search space, given $(\kappa, f_{\text{ref}}) \in (0,1) \times [f_{\inf}, f(x_0))$, into

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

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

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

*We don't need $f_{\text{ref}} = f_{\inf}$, but, for simplicity, think of it that way here

Functions satisfying Polyak-Łojasiewicz

Theorem 3

A continuously differentiable f with a Lipschitz continuous gradient satisfies the Polyak-Łojasiewicz condition if and only if $\mathcal{R}_1 = \mathcal{L}$ for any $x_0 \in \mathbb{R}^n$.

Hence, if we prove something about the behavior of an algorithm over $\mathcal{R}_1$, then

- we know how it behaves if f satisfies PL *and*
- we know how it behaves *at any point* satisfying the PL inequality.

Functions satisfying a strict-saddle-type property

Theorem 4

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

$$\max\{\|\nabla f(x)\|_2^2, -\lambda_{\min}(\nabla^2 f(x))^3\} \geq \zeta(f(x) - f_{inf}),$$

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

Outline

Motivation

Contemporary Analyses

Partitioning the Search Space

Behavior of Regularization Methods

Summary & Perspectives

Linearly convergent behavior over $\mathcal{R}_p$

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

$$w_p(x, s) = t_p(x, s) + \frac{l_p}{p+1}\|s\|_2^{p+1}$$

Theorem 5

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

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

then, with $\epsilon_f \in (0, f(x_0) - f_{ref})$, the number of iterations in

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

is bounded above by

$$\left\lceil \log\left(\frac{f(x_0) - f_{ref}}{\epsilon_f}\right)\left(\log\left(\frac{1}{1-\kappa}\right)\right)^{-1}\right\rceil = \mathcal{O}\left(\log\left(\frac{f(x_0) - f_{ref}}{\epsilon_f}\right)\right)$$

Characterization: Contemporary

Let RG and RN represent regularized gradient and Newton, respectively.

Theorem 6

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

$$\mathcal{K}_1(\epsilon_g) := \{k \in \mathbb{N} : \|\nabla f(x_k)\|_2 > \epsilon_g\}$$
$$\textit{and} \;\; \mathcal{K}_2(\epsilon_H) := \{k \in \mathbb{N} : \lambda_{\min}(\nabla^2 f(x_k)) < -\epsilon_H\}.$$

Then, the cardinalities of $\mathcal{K}_1(\epsilon_g)$ and $\mathcal{K}_2(\epsilon_H)$ are of the order...

Algorithm	$\lvert\mathcal{K}_1(\epsilon_g)\rvert$	$\lvert\mathcal{K}_2(\epsilon_H)\rvert$
RG	$\mathcal{O}\left(\frac{l_1(f(x_0)-f_{inf})}{\epsilon_g^2}\right)$	∞
RN	$\mathcal{O}\left(\frac{l_2^{1/2}(f(x_0)-f_{inf})}{\epsilon_g^{3/2}}\right)$	$\mathcal{O}\left(\frac{l_2^2(f(x_0)-f_{inf})}{\epsilon_H^3}\right)$

Characterization: Our approach

Theorem 7

The numbers of iterations in $\mathcal{R}_1$ and $\mathcal{R}_2$ with $f_{ref} = f_{inf}$ are of the order...

Algorithm	$\mathcal{R}_1$	$\mathcal{R}_2$
RG	$\mathcal{O}\left(\log\left(\frac{f(x_0)-f_{inf}}{\epsilon_f}\right)\right)$	∞
RN	$\mathcal{O}\left(\frac{l_2^2(f(x_0)-f_{inf})}{r_1^3}\right) + \mathcal{O}\left(\log\left(\frac{f(x_0)-f_{inf}}{\epsilon_f}\right)\right)$	$\mathcal{O}\left(\log\left(\frac{f(x_0)-f_{inf}}{\epsilon_f}\right)\right)$

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

Characterization: Our approach

Theorem 7

The numbers of iterations in $\mathcal{R}_1$ and $\mathcal{R}_2$ with $f_{ref} = f_{inf}$ are of the order...

Algorithm	$\mathcal{R}_1$	$\mathcal{R}_2$
RG	$\mathcal{O}\left(\log\left(\frac{f(x_0)-f_{inf}}{\epsilon_f}\right)\right)$	∞
RN	$\mathcal{O}\left(\frac{l_2^2(f(x_0)-f_{inf})}{r_1^3}\right) + \mathcal{O}\left(\log\left(\frac{f(x_0)-f_{inf}}{\epsilon_f}\right)\right)$	$\mathcal{O}\left(\log\left(\frac{f(x_0)-f_{inf}}{\epsilon_f}\right)\right)$

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

A ∞ can appear, but one could consider probabilistic bounds, too

Outline

Motivation

Contemporary Analyses

Partitioning the Search Space

Behavior of Regularization Methods

Summary & Perspectives

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.