Slope and geometry in variational mathematics

Dmitriy Drusvyatskiy
Combinatorics & Optimization, University of Waterloo

Joint work with
A. Daniilidis (Chile), A.D. Ioffe (Technion), A.S. Lewis (Cornell)

Oct 5, 2013
Fix a metric space \((\mathcal{X}, d)\) and a function \(f : \mathcal{X} \to \overline{\mathbb{R}}\).
Fix a metric space \((X, d)\) and a function \(f : X \rightarrow \mathbb{R}\).

Slope: “Fastest instantaneous rate of decrease”
Fix a metric space \((\mathcal{X}, d)\) and a function \(f : \mathcal{X} \to \overline{\mathbb{R}}\).

**Slope:** “Fastest instantaneous rate of decrease”

\[
|\nabla f| (\bar{x}) := \limsup_{x \to \bar{x}} \frac{f(\bar{x}) - f(x)}{d(\bar{x}, x)}
\]
Fix a metric space \((\mathcal{X}, d)\) and a function \(f : \mathcal{X} \to \bar{\mathbb{R}}\).

**Slope:** “Fastest instantaneous rate of decrease”

\[
|\nabla f|(\bar{x}) := \limsup_{x \to \bar{x}} \frac{f(\bar{x}) - f(x)}{d(\bar{x}, x)}
\]

**Limiting slope:**

\[
|\nabla f|(\bar{x}) := \liminf_{x \to \bar{x}} |\nabla f|(x)
\]
Fix a metric space \((X, d)\) and a function \(f : X \to \mathbb{R}\).

Slope: “Fastest instantaneous rate of decrease”

\[
|\nabla f|(\bar{x}) := \limsup_{x \to \bar{x}} \frac{f(\bar{x}) - f(x)}{d(\bar{x}, x)}
\]

Limiting slope:

\[
|\nabla f|(\bar{x}) := \liminf_{x \to \bar{x}} |\nabla f|(x)
\]

Critical points:

\(\bar{x}\) is critical for \(f\) \iff \(|\nabla f|(\bar{x}) = 0\)
Fix a metric space \((X, d)\) and a function \(f : X \to \mathbb{R}\).

Slope: “Fastest instantaneous rate of decrease”

\[
|\nabla f|(\bar{x}) := \limsup_{x \to \bar{x}} \frac{f(\bar{x}) - f(x)}{d(\bar{x}, x)}
\]

Limiting slope:

\[
|\nabla f|(\bar{x}) := \liminf_{x \to \bar{x}} |\nabla f|(x)
\]

Critical points:

\[
\bar{x} \text{ is critical for } f \iff |\nabla f|(\bar{x}) = 0
\]

Eg:
Method of alternating projections

Common problem:

Given sets $A, B \subset \mathbb{R}^n$, find some point $x \in A \cap B$. 

Diagram: 

- Set $A$ is shaded.
- Set $B$ is represented by the line.
- Point $x$ is the intersection of $A$ and $B$. 

Distance and projection:

- For set $B$: $d_B(x) = \min_{y \in B} |x - y|$ and $P_B(x) = \{\text{nearest points of } B \text{ to } x\}$. 

Finding points in $P_A$ and $P_B$ is often easy!

Eg 1 (simple example): Linear programming:

$\{x: x_{\geq 0}\} \cap \{x: Ax = b\}$.

Eg 2 (more interesting): Low-order control:

$\{X \succeq 0: \text{rank}(X) \leq r\} \cap \{X: A(X) = b\}$. 

Method of alternating projections

Common problem:

Given sets $A, B \subset \mathbb{R}^n$, find some point $x \in A \cap B$.

Distance and projection:

$$d_B(x) = \min_{y \in B} |x - y| \quad \text{and} \quad P_B(x) = \{\text{nearest points of } B \text{ to } x\}.$$
Method of alternating projections

Common problem:

Given sets $A, B \subset \mathbb{R}^n$, find some point $x \in A \cap B$.

Distance and projection:

$$d_B(x) = \min_{y \in B} |x - y| \quad \text{and} \quad P_B(x) = \{ \text{nearest points of } B \text{ to } x \}.$$  

Finding points in $P_A$ and $P_B$ is often easy!
Method of alternating projections

Common problem:

Given sets $A, B \subset \mathbb{R}^n$, find some point $x \in A \cap B$.

Distance and projection:

$$d_B(x) = \min_{y \in B} |x - y| \quad \text{and} \quad P_B(x) = \{\text{nearest points of } B \text{ to } x\}.$$ 

Finding points in $P_A$ and $P_B$ is often easy!

Eg 1 (simple example): Linear programming:

$$\{x : x \geq 0\} \cap \{x : Ax = b\}.$$
Method of alternating projections

Common problem:

Given sets \( A, B \subset \mathbb{R}^n \), find some point \( x \in A \cap B \).

Distance and projection:

\[
d_B(x) = \min_{y \in B} |x - y| \quad \text{and} \quad P_B(x) = \{ \text{nearest points of } B \text{ to } x \}.
\]

Finding points in \( P_A \) and \( P_B \) is often easy!

Eg 1 (simple example): Linear programming:

\[
\{ x : x \geq 0 \} \cap \{ x : Ax = b \}.
\]

Eg 2 (more interesting): Low-order control:

\[
\{ X \succeq 0 : \text{rank } X \leq r \} \cap \{ X : A(X) = b \}.
\]
Method of alternating projections (von Neumann ’33):

\[ x_{k+1} \in P_B(x_k) \]
\[ x_{k+2} \in P_A(x_{k+1}) \]
Method of alternating projections (von Neumann ’33):

\[
\begin{align*}
    x_{k+1} & \in P_B(x_k) \\
x_{k+2} & \in P_A(x_{k+1})
\end{align*}
\]

The “angle” between \( A \) and \( B \) drives the convergence!
Method of alternating projections (von Neumann ’33):

\[ x_{k+1} \in P_B(x_k) \]
\[ x_{k+2} \in P_A(x_{k+1}) \]

The “angle” between \( A \) and \( B \) drives the convergence!

Quantifying the angle:

\[ \psi(x, y) := \begin{cases} 
|x - y| & \text{if } x \in A, y \in B \\
+\infty & \text{otherwise}
\end{cases} \]

Comparison of \( |\nabla \psi_y|(x) \) and \( |\nabla \psi_x|(y) \) quantifies the angle!
• When is slope an adequate tool?
  — Semi-algebraic case
  — Slope & error bounds

• Applications:
  — Alternating projections & transversality
  — Steepest descent curves
  — Active sets & generic properties.
Subdifferentials (an interlude)

Subdifferential:

\[ v \in \partial f(\bar{x}) \iff \bar{x} \text{ is critical for } f - \langle v, \cdot \rangle \]
Subdifferentials (an interlude)

Subdifferential:

\[ v \in \partial f(\bar{x}) \iff \bar{x} \text{ is critical for } f - \langle v, \cdot \rangle \]

Surprising consequence (Ioffe ’00):

\[ |\nabla f|(x) = \text{dist}(0, \partial f(x)). \]
Subdifferentials (an interlude)

Subdifferential:
\[ v \in \partial f(\bar{x}) \iff \bar{x} \text{ is critical for } f - \langle v, \cdot \rangle \]

Surprising consequence (Ioffe ’00):
\[ |\nabla f|(x) = \text{dist } (0, \partial f(x)). \]

Example:
\[ \partial (-|\cdot|)(x) = \begin{cases} 1 & \text{if } x < 0 \\ \{-1, 1\} & \text{if } x = 0 \\ -1 & \text{if } x > 0 \end{cases} \]
Subdifferentials (an interlude)

Subdifferential:

\[ v \in \partial f(\bar{x}) \iff \bar{x} \text{ is critical for } f - \langle v, \cdot \rangle \]

Surprising consequence (Ioffe ’00):

\[ |\nabla f|(x) = \text{dist} (0, \partial f(x)). \]

Example:

\[ \partial (-|\cdot|)(x) = \begin{cases} 
1 & \text{if } x < 0 \\
\{-1, 1\} & \text{if } x = 0 \\
-1 & \text{if } x > 0
\end{cases} \]

Subdifferential graph:

\[ \text{gph } \partial f := \{(x, v) \in \mathbb{R}^n \times \mathbb{R}^n : v \in \partial f(x)\} \]
Are these notions adequate?
Subdifferentials

Are these notions adequate?

Pathology: \( \text{gph} \partial f \) can be very large!
Subdifferentials

Are these notions adequate?

**Pathology:** $\text{gph} \partial f$ can be very large!

What do we expect the size of $\text{gph} \partial f$ to be?

- If $f: \mathbb{R}^n \to \mathbb{R}$ is **smooth**, then $\text{gph} \partial f$ is $n$-dimensional smooth manifold.
- If $f: \mathbb{R}^n \to \mathbb{R}$ is **convex**, then $\text{gph} \partial f$ is $n$-dimensional Lipschitz manifold (Minty ’62).
Subdifferentials

Are these notions adequate?

Pathology: $\text{gph} \, \partial f$ can be very large!

What do we expect the size of $\text{gph} \, \partial f$ to be?

- If $f : \mathbb{R}^n \to \mathbb{R}$ is smooth, then $\text{gph} \, \partial f$ is $n$-dimensional smooth manifold.
- If $f : \mathbb{R}^n \to \mathbb{R}$ is convex, then $\text{gph} \, \partial f$ is $n$-dimensional Lipschitz manifold (Minty ’62).

Multiple authors (Rockafellar, Borwein, Wang, . . .):
There are functions $f : \mathbb{R}^n \to \mathbb{R}$ with “$2n$-dimensional” $\text{gph} \, \partial f$. 

$Q \subset \mathbb{R}^n$ is semi-algebraic if is a finite union of solution sets to finitely many polynomial inequalities.
Semi-algebraic geometry

\( Q \subset \mathbb{R}^n \) is semi-algebraic if is a finite union of solution sets to finitely many polynomial inequalities.

Eg: semi-definite representable sets (Nesterov-Nemirovskii)
Semi-algebraic geometry

\( Q \subset \mathbb{R}^n \) is semi-algebraic if is a finite union of solution sets to finitely many polynomial inequalities.

Eg: semi-definite representable sets (Nesterov-Nemirovskii)

Semi-algebraicity is robust (Tarski-Seidenberg theorem).
Eg:
\( f \) semi-algebraic \( \implies \) \( \text{gph} \, \partial f \) and \( |\nabla f| \) are semi-algebraic.
Semi-algebraic geometry

$Q \subset \mathbb{R}^n$ is \textit{semi-algebraic} if it is a finite union of solution sets to finitely many polynomial inequalities.

\textbf{Eg:} semi-definite representable sets (Nesterov-Nemirovskii)

Semi-algebraicity is robust (Tarski-Seidenberg theorem).

\textbf{Eg:}

\[
\text{if } f \text{ semi-algebraic } \implies \text{gph } \partial f \text{ and } |\nabla f| \text{ are semi-algebraic.}
\]

Semi-algebraic $Q$ “stratify” into finitely many manifolds $\{\mathcal{M}_i\}$.

Dimension:

\[
\dim Q := \max_{i=1,\ldots,k} \{\dim \mathcal{M}_i\}.
\]
Theorem (D-Ioffe-Lewis)

For semi-algebraic $f : \mathbb{R}^n \to \overline{\mathbb{R}}$, we have

$$\dim \text{gph} \partial f = n,$$

even locally around any point in $\text{gph} \partial f$.
Semi-algebraic geometry

Theorem (D-Ioffe-Lewis)

For semi-algebraic $f: \mathbb{R}^n \rightarrow \overline{\mathbb{R}}$, we have

$$\dim \text{gph } \partial f = n,$$

even locally around any point in $\text{gph } \partial f$.

Conclusion: Criticality is meaningful for concrete variational problems!
Theorem (D-Ioffe-Lewis)

For semi-algebraic $f : \mathbb{R}^n \to \mathbb{R}$, we have

$$\dim \text{gph} \partial f = n,$$

even locally around any point in $\text{gph} \partial f$.

Conclusion: Criticality is meaningful for concrete variational problems!

Semi-algebraic $f$ have only finitely many critical values (cf. Sard’s Theorem)
Semi-algebraic geometry

Theorem (D-Ioffe-Lewis)

For semi-algebraic $f : \mathbb{R}^n \to \mathbb{R}$, we have

$$\dim \text{gph} \partial f = n,$$

even locally around any point in \text{gph} \partial f.

Conclusion: Criticality is meaningful for concrete variational problems!

Semi-algebraic $f$ have only finitely many critical values (cf. Sard’s Theorem) \Rightarrow intervals $(a, b)$ of non-critical values.
Theorem (D-Ioffe-Lewis)

For semi-algebraic \( f : \mathbb{R}^n \rightarrow \overline{\mathbb{R}} \), we have

\[
\dim \text{gph } \partial f = n,
\]

even locally around any point in \( \text{gph } \partial f \).

Conclusion: Criticality is meaningful for concrete variational problems!

Semi-algebraic \( f \) have only finitely many critical values
(cf. Sard's Theorem) \( \Rightarrow \) intervals \((a, b)\) of non-critical values.

What can we learn from non-criticality?
Common problem: Estimate

\[ \text{dist} (x, [f \leq r]) \quad \text{(difficult)}. \]
Common problem: Estimate

\[ \text{dist} (x, [f \leq r]) \] (difficult).

“The residual”:

\[ f(x) - r \] (easy).

\[ \text{Desirable quality:} \exists \kappa \text{ with} \text{dist} (x, [f \leq r]) \leq \kappa (f(x) - r). \]
Slope & error bounds

Common problem: Estimate
\[ \text{dist} \left( x, [f \leq r] \right) \] (difficult).

“The residual”:
\[ f(x) - r \] (easy).

Desirable quality: Exists \( \kappa \) with
\[ \text{dist} \left( x, [f \leq r] \right) \leq \kappa (f(x) - r). \]
Common problem: Estimate

\[ \text{dist}(x, [f \leq r]) \]  (difficult).

“The residual”:  

\[ f(x) - r \]  (easy).

Desirable quality: Exists \( \kappa \) with

\[ \text{dist}(x, [f \leq r]) \leq \kappa (f(x) - r). \]

Restrict \( f : \mathbb{R}^n \rightarrow \overline{\mathbb{R}} \) to a “slice” \( f^{-1}(a, b) \).
Common problem: Estimate
\[ \text{dist} (x, [f \leq r]) \] (difficult).

“The residual”:
\[ f(x) - r \] (easy).

Desirable quality: Exists \( \kappa \) with
\[ \text{dist} (x, [f \leq r]) \leq \kappa (f(x) - r). \]

Restrict \( f : \mathbb{R}^n \to \mathbb{R} \) to a “slice” \( f^{-1}(a, b) \).

Lemma (Error bound)
The following are equivalent.
Non-criticality:
\[ |\nabla f| \geq \frac{1}{\kappa}. \]
Error-bound:
\[ \text{dist} (x, [f \leq r]) \leq \kappa (f(x) - r), \text{ when } r \in (a, f(x)). \]
Slope & error bounds

Common problem: Estimate
\[ \text{dist} (x, [f \leq r]) \text{ (difficult).} \]

“The residual”:
\[ f(x) - r \text{ (easy).} \]

Desirable quality: Exists \( \kappa \) with
\[ \text{dist} (x, [f \leq r]) \leq \kappa (f(x) - r). \]

Restrict \( f : \mathbb{R}^n \to \bar{\mathbb{R}} \) to a “slice” \( f^{-1}(a, b) \).

Lemma (Error bound)
The following are equivalent.
Non-criticality:
\[ |\nabla f| \geq \frac{1}{\kappa}. \]
Error-bound:
\[ \text{dist} (x, [f \leq r]) \leq \kappa (f(x) - r), \text{ when } r \in (a, f(x)). \]

- Observed by Azé-Corvellec ’04, Ioffe ’00.
Semi-algebraicity & error bounds

Figure: $f(x) = x^2$

Desingularization: (Bolte-Daniilidis-Lewis ’07)

For semi-algebraic $f$, there exists "nice" $\phi$ with $|\nabla (\phi \circ f)(x)| \geq 1$ for $x / \in \text{crit } f$.

Error bounds always applicable for semi-algebraic functions!
Semi-algebraicity & error bounds

Figure: $f(x) = x^2$

Figure: $\sqrt{f(x)} = |x|$
Semi-algebraicity & error bounds

Figure: \( f(x) = x^2 \)

Figure: \( \sqrt{f(x)} = |x| \)

**Desingularization:** (Bolte-Daniilidis-Lewis ’07)
For semi-algebraic \( f \), there exists “nice” \( \phi \) with

\[
|\nabla (\phi \circ f)\( (x) \) \geq 1 \quad \text{for } x \notin \text{crit } f.
\]
**Semi-algebraicity & error bounds**

**Figure:** \( f(x) = x^2 \)

**Figure:** \( \sqrt{f(x)} = |x| \)

**Desingularization:** (Bolte-Daniilidis-Lewis ’07)

For semi-algebraic \( f \), there exists “nice” \( \phi \) with

\[
|\nabla (\phi \circ f)|(x) \geq 1 \quad \text{for } x \notin \text{crit } f.
\]

**Error bounds always applicable for semi-algebraic functions!**
Slope & Error bounds

Second-order behaviour (error bounds of the slope):

\[ x \mapsto |\nabla f|(x) \]

Theorem (D-Ioffe)

For semi-algebraic \( f \) and a strict minimizer \( \bar{x} \), the following are equivalent:

Quadratic growth:

\[ f(x) \geq f(\bar{x}) + \alpha |x - \bar{x}|^2 \]

for \( x \) near \( \bar{x} \).

Error-bound:

\[ |x - \bar{x}| \leq \kappa \cdot d(0, \partial f(x)) \]

for \( x \) near \( \bar{x} \).

• Not true for general functions; e.g. \( f(x) = 2x^2 + \frac{1}{2}x^2 \sin \left( \frac{1}{x} \right) \)

Second-order growth/Regularity: Poliquin-Rockafellar '98, D-Nghia-Mordukhovich '13, Artacho-Geoffroy '08 '13
Slope & Error bounds

Second-order behaviour (error bounds of the slope):

\[ x \mapsto |\nabla f(x)| \]

Theorem (D-Ioffe)

For semi-algebraic \( f \) and a strict minimizer \( \bar{x} \), the following are equivalent:

Quadratic growth:

\[ f(x) \geq f(\bar{x}) + \alpha |x - \bar{x}|^2 \quad \text{for } x \text{ near } \bar{x}. \]

Error-bound:

\[ |x - \bar{x}| \leq \kappa \cdot d(0, \partial f(x)) \quad \text{for } x \text{ near } \bar{x}. \]
Second-order behaviour (error bounds of the slope):

\[ x \mapsto |\nabla f|(x) \]

**Theorem (D-Ioffe)**

For semi-algebraic \( f \) and a strict minimizer \( \bar{x} \), the following are equivalent:

**Quadratic growth:**

\[ f(x) \geq f(\bar{x}) + \alpha |x - \bar{x}|^2 \quad \text{for } x \text{ near } \bar{x}. \]

**Error-bound:**

\[ |x - \bar{x}| \leq \kappa \cdot d(0, \partial f(x)) \quad \text{for } x \text{ near } \bar{x}. \]

- Not true for general functions; eg. \( f(x) = 2x^2 + \frac{1}{2} x^2 \sin(\frac{1}{x}) \)
Slope & Error bounds

Second-order behaviour (error bounds of the slope):

\[ x \mapsto |\nabla f|(x) \]

Theorem (D-Ioffe)

For semi-algebraic \( f \) and a strict minimizer \( \bar{x} \), the following are equivalent:

Quadratic growth:

\[ f(x) \geq f(\bar{x}) + \alpha |x - \bar{x}|^2 \quad \text{for } x \text{ near } \bar{x}. \]

Error-bound:

\[ |x - \bar{x}| \leq \kappa \cdot d(0, \partial f(x)) \quad \text{for } x \text{ near } \bar{x}. \]

- Not true for general functions; eg. \( f(x) = 2x^2 + \frac{1}{2}x^2 \sin(\frac{1}{x}) \)

Alternating projections & transversality
Convergence of alternating projections

Transversality: \( N_A(\bar{x}) \cap -N_B(\bar{x}) = \{0\} \) \( (N_A = \partial \delta_A) \)

Figure: Not transverse

Figure: Transverse
Convergence of alternating projections

Transversality: \( N_A(\bar{x}) \cap -N_B(\bar{x}) = \{0\} \quad (N_A = \partial \delta_A) \)

Figure: Not transverse

Local convergence (D-Ioffe-Lewis ’13):

\( A \) and \( B \) transverse at \( \bar{x} \) \( \implies \) local \( \mathbb{R} \)-linear convergence.
Coupling function:

\[
\psi(x, y) = \delta_A(x) + |x - y| + \delta_B(y).
\]
Transversality & error bounds

Coupling function:

\[
\psi(x, y) = \delta_A(x) + |x - y| + \delta_B(y).
\]

Error bound:

\[
N_B(\bar{x}) \cap -N_A(\bar{x}) = \{0\}
\]

\[
\downarrow
\]

\[
\max \{|\nabla \psi_x|(y), |\nabla \psi_y|(x)| \geq \kappa
\]

for \(x \in A\) and \(y \in B\), not in \(A \cap B\).
Transversality & error bounds

Coupling function:

\[ \psi(x, y) = \delta_A(x) + |x - y| + \delta_B(y). \]

Error bound:

\[ N_B(\bar{x}) \cap -N_A(\bar{x}) = \{0\} \]
\[ \downarrow \]
\[ \max \{ |\nabla \psi_x|(y), |\nabla \psi_y|(x) \} \geq \kappa \]

for \( x \in A \) and \( y \in B \), not in \( A \cap B \).

\[ \downarrow \]

Local linear convergence
Convergence

- Transversality is necessary but not verifiable.
Convergence

- Transversality is necessary but not verifiable.

Eg:

\[
\begin{align*}
A & \quad B \\
\end{align*}
\]
Convergence

- Transversality is necessary but not verifiable.

Eg:

What about sublinear convergence?
Convergence

- Transversality is necessary but not verifiable.

**Eg:**

![Diagram](image)

What about sublinear convergence?

**Theorem (D-Ioffe-Lewis)**

\[ A \text{ and } B \text{ are semi-algebraic, } A \cap B \text{ is compact, } x_0 \text{ near } A \cap B \]

\[ \implies \text{ alternating projections converge.} \]
Convergence

- Transversality is necessary but not verifiable.

Eg:

What about sublinear convergence?

Theorem (D-Ioffe-Lewis)

A and B are semi-algebraic, A ∩ B is compact, x₀ near A ∩ B
⇒ alternating projections converge.

Generic transversality (D-Ioffe-Lewis):
If A and B are semi-algebraic, then
A + a and B + b are transverse for a.e. (a, b)
Open question

General paradigm:

\[ \text{no convexity} \implies \text{no global convergence}. \]
Open question

General paradigm:

\[
\text{no convexity} \implies \text{no global convergence.}
\]

Variants of alternating projections work globally!
Open question

General paradigm:

no convexity $\implies$ no global convergence.

Variants of alternating projections work globally!

- Integer programming:

$$\mathbb{Z}^n \cap \{x : Ax \leq b\}$$

(eg: sudoku, 3-SAT, 4 queens problem, etc ...)

Ongoing work Artacho, Borwein, Tam.
Steepest descent curves


**Steepest descent curves**

**Bounded speed:** A curve \( x: [0, T] \rightarrow \mathcal{X} \) is 1-Lipschitz if

\[
\text{dist} (x(t), x(s)) \leq |t - s|.
\]
Steepest descent curves

**Bounded speed:** A curve $x: [0, T] \to \mathcal{X}$ is 1-Lipschitz if

$$\text{dist}(x(t), x(s)) \leq |t - s|.$$
Steepest descent curves

Bounded speed: A curve $x: [0, T] \to \mathcal{X}$ is 1-Lipschitz if

$$\text{dist} \ (x(t), x(s)) \leq |t - s|.$$ 

Motivation: 1-Lipschitz curves $x: [0, T] \to \mathcal{X}$ satisfy

$$|\nabla (f \circ x)| \leq |\nabla f|(x).$$
Steepest descent curves

Bounded speed: A curve $x: [0, T] \rightarrow \mathcal{X}$ is 1-Lipschitz if

$$\text{dist}(x(t), x(s)) \leq |t - s|.$$ 

What are steepest descent curves?

Motivation: 1-Lipschitz curves $x: [0, T] \rightarrow \mathcal{X}$ satisfy

$$|\nabla(f \circ x)| \leq |\nabla f|(x).$$

Definition (Near-steepest descent curves)

Curve $x: [0, T] \rightarrow \mathcal{X}$ is a near-steepest descent curve if

- $x$ is 1-Lipschitz,
- $f \circ x$ is decreasing,
- $|\nabla(f \circ x)| \geq |\nabla f|(x)$, a.e. on $[0, T]$. 
Example

Figure: \( f(x, y) = \max\{x + y, |x - y|\} + x(x + 1) + y(y + 1) + 100 \)
Theorem (D-Ioffe-Lewis)

For reasonable $f : \mathbb{R}^n \to \overline{\mathbb{R}}$ and a curve $x : [0, T] \to \mathbb{R}^n$ the following are equivalent.

1. $x$ is a near-steepest descent curve,
2. $f \circ x$ is decreasing and (after reparametrizing)

$$\dot{x} \in -\partial f(x), \quad \text{a.e. on } [0, T].$$

Remark: When 2. holds, $\dot{x}$ is the shortest element of $-\partial f(x)$, a.e. on $[0, T]$. Reasonable conditions: $f$ is smooth, convex, or semi-algebraic.
Theorem (D-Ioffe-Lewis)

For reasonable $f: \mathbb{R}^n \rightarrow \overline{\mathbb{R}}$ and a curve $x: [0, T] \rightarrow \mathbb{R}^n$ the following are equivalent.

1. $x$ is a near-steepest descent curve,
2. $f \circ x$ is decreasing and (after reparametrizing)

$$\dot{x} \in -\partial f(x), \quad \text{a.e. on } [0, T].$$

Remark: When 2. holds,

$$\dot{x} \text{ is the shortest element of } -\partial f(x), \quad \text{a.e. on } [0, T].$$
Theorem (D-Ioffe-Lewis)

For reasonable $f: \mathbb{R}^n \to \overline{\mathbb{R}}$ and a curve $x: [0, T] \to \mathbb{R}^n$ the following are equivalent.

1. $x$ is a near-steepest descent curve,
2. $f \circ x$ is decreasing and (after reparametrizing)

$$\dot{x} \in -\partial f(x), \quad \text{a.e. on } [0, T].$$

**Remark:** When 2. holds,

$$\dot{x} \text{ is the shortest element of } -\partial f(x), \quad \text{a.e. on } [0, T].$$

**Reasonable conditions:** $f$ is smooth, convex, or semi-algebraic.
Existence

Theorem (Ambrosio et al. ’05, De Giorgi ’93)

For reasonable $f : \mathcal{X} \rightarrow \overline{\mathbb{R}}$, there exist near-steepest descent curves starting from any point.
Existence

Theorem (Ambrosio et al. '05, De Giorgi ’93)

For reasonable $f: \mathcal{X} \to \overline{\mathbb{R}}$, there exist near-steepest descent curves starting from any point.

Proof ingredients:

• Moreau-Yosida approximation:

$$x_{k+1} = \arg\min_{x \in \mathcal{X}} \left\{ f(x) + \frac{1}{2\tau} d^2(x, x_k) \right\}. $$

• Extraneous topologies $\Rightarrow$ existence of minimizers and convergence.
Existence

Theorem (Ambrosio et al. ’05, De Giorgi ’93)

For reasonable $f : \mathcal{X} \rightarrow \mathbb{R}$, there exist near-steepest descent curves starting from any point.

Proof ingredients:

• Moreau-Yosida approximation:

$$x_{k+1} = \arg\min_{x \in \mathcal{X}} \left\{ f(x) + \frac{1}{2\tau} d^2(x, x_k) \right\}.$$ 

• Extraneous topologies $\Rightarrow$ existence of minimizers and convergence.

Proof is opaque and uses heavy machinery!
New proof idea: Error bound lemma

Equipartition: For $\eta > 0$,

$$f(x_0) - \eta = \tau_0 < \ldots < \tau_k = f(x_0).$$
New proof idea: Error bound lemma

**Equipartition:** For $\eta > 0$,

$$f(x_0) - \eta = \tau_0 < \ldots < \tau_k = f(x_0).$$

**Initialize:** $j = 0$;

**while** $i \leq k$ **do**

\[ x_{j+1} \leftarrow P_{[f \leq \tau_{j+1}]}(x_j); \]

**end**
New proof idea: Error bound lemma

**Equipartition:** For $\eta > 0$,

$$f(x_0) - \eta = \tau_0 < \ldots < \tau_k = f(x_0).$$

**Initialize:** $j = 0$;  
while $i \leq k$ do  
\[ x_{j+1} \leftarrow P_{[f \leq \tau_{j+1}]}(x_j); \]
end

Consider resulting trajectories as $k \to \infty$. 

![Diagram showing trajectories](image-url)
Semi-algebraic descent

Theorem (D-Ioffe-Lewis ’13)
f semi-algebraic, \( \bar{x} \) not a local minimizer \( \implies \)

There exists a nontrivial descent curve starting from \( \bar{x} \)
Semi-algebraic descent

Theorem (D-Ioffe-Lewis ’13)

\( f \) semi-algebraic, \( \bar{x} \) not a local minimizer \( \implies \)

There exists a nontrivial descent curve starting from \( \bar{x} \)

- Not true for \( C^\infty \) functions; eg. \( f(x) = e^{-\frac{1}{|x|}} \sin\left(\frac{1}{x}\right) \)
Semi-algebraic descent

Theorem (D-Ioffe-Lewis ’13)

\[ f \text{ semi-algebraic, } \bar{x} \text{ not a local minimizer } \implies \]

There exists a nontrivial descent curve starting from \( \bar{x} \)

- Not true for \( C^\infty \) functions; eg. \( f(x) = e^{-\frac{1}{|x|}} \sin\left(\frac{1}{x}\right) \)

Theorem (Daniilidis-Bolte-D-Ioffe-Lewis, Lojasiewicz)

If \( f \) is semi-algebraic, then any bounded near-steepest descent curve (with maximal domain) has bounded length and converges to a critical point.
Semi-algebraic descent

Theorem (D-Ioffe-Lewis ’13)

- Not true for $C^\infty$ functions; eg. $f(x) = e^{-\frac{1}{|x|}} \sin\left(\frac{1}{x}\right)$

Theorem (Daniilidis-Bolte-D-Ioffe-Lewis, Lojasiewicz)

If $f$ is semi-algebraic, then any bounded near-steepest descent curve (with maximal domain) has bounded length and converges to a critical point.

- Not true for $C^\infty$ functions (Palis, de Melo)
Semi-algebraic descent

Theorem (D-Ioffe-Lewis ’13)

\( f \) semi-algebraic, \( \bar{x} \) not a local minimizer \( \implies \)

There exists a nontrivial descent curve starting from \( \bar{x} \)

• Not true for \( C^\infty \) functions; eg. \( f(x) = e^{-\frac{1}{|x|}} \sin(\frac{1}{x}) \)

Theorem (Daniilidis-Bolte-D-Ioffe-Lewis, Lojasiewicz)

If \( f \) is semi-algebraic, then any bounded near-steepest descent curve (with maximal domain) has bounded length and converges to a critical point.

• Not true for \( C^\infty \) functions (Palis, de Melo)

One motivation: Algorithm complexity (Eg: Attouch et al.)
Active sets in optimization.
Active sets in optimization

Figure: $Q$ is $4 \times 4$ Toeplitz spectrahedron
Active sets in optimization

Figure: \( Q \) is 4 \( \times \) 4 Toeplitz spectrahedron

Definition (Partial Smoothness)
A set \( Q \) is partly smooth relative to \( \mathcal{M} \subset Q \) if

1. (Smoothness) \( \mathcal{M} \) is a smooth manifold,
2. (Sharpness) \( N_{\mathcal{M}} = \text{span} \ N_Q \) on \( \mathcal{M} \),
3. (Continuity) \( N_Q \) varies continuously on \( \mathcal{M} \).
Active sets in optimization

Figure: $Q$ is $4 \times 4$ Toeplitz spectrahedron

Definition (Partial Smoothness)
A set $Q$ is partly smooth relative to $\mathcal{M} \subset Q$ if

1. (Smoothness) $\mathcal{M}$ is a smooth manifold,
2. (Sharpness) $N_{\mathcal{M}} = \text{span} \ N_Q$ on $\mathcal{M}$,
3. (Continuity) $N_Q$ varies continuously on $\mathcal{M}$.

(Originates in Lewis ’03)
Active sets in optimization

Partial smoothness has classical roots!

Eg: **Smooth constraints**

\[ Q := \{ x : g_i(x) \leq 0, \quad \text{for } i = 1, \ldots, m \} \]

where \( g_1, \ldots, g_m \) are smooth.
Active sets in optimization

Partial smoothness has classical roots!

**Eg: Smooth constraints**

\[ Q := \{ x : g_i(x) \leq 0, \text{ for } i = 1, \ldots, m \} \]

where \( g_1, \ldots, g_m \) are smooth.

**Active indices:**

\[ I(x) = \{ i : g_i(x) = 0 \} \]
Active sets in optimization

Partial smoothness has classical roots!

Eg: Smooth constraints

\[ Q := \{ x : g_i(x) \leq 0, \quad \text{for } i = 1, \ldots, m \} \]

where \( g_1, \ldots, g_m \) are smooth.

Active indices:

\[ I(x) = \{ i : g_i(x) = 0 \} \]

Reasonable conditions \( \implies \)

\[ \mathcal{M} := \{ x : I(x) = I(\bar{x}) \} \quad \text{is a partly smooth manifold near } \bar{x} \]
Active sets in optimization

Partial smoothness has classical roots!

Eg: Smooth constraints

\[ Q := \{ x : g_i(x) \leq 0, \quad \text{for } i = 1, \ldots, m \} \]

where \( g_1, \ldots, g_m \) are smooth.

Active indices:

\[ I(x) = \{ i : g_i(x) = 0 \} \]

Reasonable conditions \( \implies \)

\[ M := \{ x : I(x) = I(\bar{x}) \} \quad \text{is a partly smooth manifold near } \bar{x} \]

Figure: \( Q = \{(x, y, z) : z \geq x(1-x) + y^2, \quad z \geq -x(1+x) + y^2 \} \)
Active sets in optimization

Partial smoothness has classical roots!

Eg: **Sum of perturbed norms**

\[
\min_x f(x) := \sum_{i=1}^{m} \|F_i(x)\|
\]

where \(F_1, \ldots, F_m\) are smooth.
Active sets in optimization

Partial smoothness has classical roots!

Eg: **Sum of perturbed norms**

$$\min_x f(x) := \sum_{i=1}^m \|F_i(x)\|$$

where $F_1, \ldots, F_m$ are smooth.

Active indices:

$$I(x) := \{i : F_i(x) = 0\}$$
Active sets in optimization

Partial smoothness has classical roots!

Eg: **Sum of perturbed norms**

\[
\min_x f(x) := \sum_{i=1}^{m} \|F_i(x)\|
\]

where \( F_1, \ldots, F_m \) are smooth.

Active indices:

\[
I(x) := \{ i : F_i(x) = 0 \}
\]

Reasonable conditions \( \implies \) epi \( f \) is partly smooth relative to

\[
\mathcal{M} := \{(x, f(x)) : I(x) = I(\bar{x})\}
\]
Active sets in optimization

Partial smoothness has classical roots!

Eg: **Sum of perturbed norms**

\[
\min_x f(x) := \sum_{i=1}^{m} \|F_i(x)\|
\]

where \(F_1, \ldots, F_m\) are smooth.

Active indices:

\[
I(x) := \{i : F_i(x) = 0\}
\]

Reasonable conditions \(\implies\) \(\text{epi } f\) is partly smooth relative to

\[
\mathcal{M} := \{(x, f(x)) : I(x) = I(\bar{x})\}
\]

Figure: \(f(x, y) := |x^2 + y^2 - 1| + |x - y|\)
Active sets in optimization

Why do optimizers care?
Active sets in optimization

Why do optimizers care?

- Many optimization algorithms identify $\mathcal{M}$ in finite time!

Eg: Gradient projection, Newton-like, proximal point.
Active sets in optimization

Why do optimizers care?

- Many optimization algorithms **identify** $\mathcal{M}$ in finite time!
  
  **Eg:** Gradient projection, Newton-like, proximal point.
  
  $\rightarrow$ Acceleration strategies!
Active sets in optimization

Why do optimizers care?

• Many optimization algorithms identify $\mathcal{M}$ in finite time!

**Eg:** Gradient projection, Newton-like, proximal point.

→ Acceleration strategies!

**Eg:** Wright, Burke-Moré, Ferris, Flåm, ...
Active sets in optimization

Why do optimizers care?

• Many optimization algorithms identify $\mathcal{M}$ in finite time!

Eg: Gradient projection, Newton-like, proximal point.

$\implies$ Acceleration strategies!

Eg: Wright, Burke-Moré, Ferris, Flåm, . . .

Finite Identification: For $\bar{x} \in \mathcal{M}$ and $\bar{v} \in \text{ri} N_Q(\bar{x})$, have

\[
\begin{aligned}
  x_i &\rightarrow \bar{x}, v_i \rightarrow \bar{v} \\
  v_i &\in N_Q(x_i)
\end{aligned}
\]

$\implies x_i \in \mathcal{M}$ for all large $i$
Active sets in optimization

Why do optimizers care?

- Many optimization algorithms identify $M$ in finite time!

  Eg: Gradient projection, Newton-like, proximal point.

  $\implies$ Acceleration strategies!

  Eg: Wright, Burke-Moré, Ferris, Flåm, ...

Finite Identification: For $\bar{x} \in M$ and $\bar{v} \in \text{ri } N_Q(\bar{x})$, have

\[
\begin{align*}
  x_i &\rightarrow \bar{x}, \quad v_i \rightarrow \bar{v} \\
  v_i &\in N_Q(x_i)
\end{align*}
\]

$\implies x_i \in M$ for all large $i$

finite identification $\iff$ partial smoothness (D-Lewis ’13)
Active sets in optimization

Why do optimizers care?

• Many optimization algorithms identify $\mathcal{M}$ in finite time!
  Eg: Gradient projection, Newton-like, proximal point.
  $\implies$ Acceleration strategies!
  Eg: Wright, Burke-Moré, Ferris, Flåm, ...

Finite Identification: For $\bar{x} \in \mathcal{M}$ and $\bar{v} \in \text{ri } N_Q(\bar{x})$, have

\[
\begin{align*}
x_i &\to \bar{x}, \quad v_i \to \bar{v} \\
v_i &\in N_Q(x_i)
\end{align*}
\]

$\implies x_i \in \mathcal{M}$ for all large $i$

finite identification $\iff$ partial smoothness (D-Lewis ’13)

Generic existence (D-Lewis): For semi-algebraic $Q$ consider

\[
\max \{ \langle v, x \rangle : x \in Q \}\]
Active sets in optimization

Why do optimizers care?

• Many optimization algorithms identify $\mathcal{M}$ in finite time!

Eg: Gradient projection, Newton-like, proximal point.

$\implies$ Acceleration strategies!

Eg: Wright, Burke-Moré, Ferris, Flâm, ... 

Finite Identification: For $\bar{x} \in \mathcal{M}$ and $\bar{v} \in \text{ri } N_Q(\bar{x})$, have

\[
x_i \to \bar{x}, \quad v_i \to \bar{v}
\]

\[
v_i \in N_Q(x_i)
\]

$\implies x_i \in \mathcal{M}$ for all large $i$

finite identification $\iff$ partial smoothness (D-Lewis ’13)

Generic existence (D-Lewis): For semi-algebraic $Q$ consider

\[
\max \{ \langle v, x \rangle : x \in Q \}
\]

Then for “typical” $v$ at any local minimizer: unique partly smooth manifold, strict complementarity, quadratic growth.
Active sets in optimization

Why do optimizers care?

• Many optimization algorithms identify $\mathcal{M}$ in finite time!

Eg: Gradient projection, Newton-like, proximal point.

$\Rightarrow$ Acceleration strategies!

Eg: Wright, Burke-Moré, Ferris, Flâm, ...

Finite Identification: For $\bar{x} \in \mathcal{M}$ and $\bar{v} \in \text{ri } N_Q(\bar{x})$, have

$$x_i \to \bar{x}, v_i \to \bar{v}$$
$$v_i \in N_Q(x_i)$$

$\Rightarrow x_i \in \mathcal{M}$ for all large $i$

finite identification $\iff$ partial smoothness (D-Lewis ’13)

Generic existence (D-Lewis): For semi-algebraic $Q$ consider

$$\max \{ \langle v, x \rangle : x \in Q \}$$

Then for “typical” $v$ at any local minimizer: unique partly smooth manifold, strict complementarity, quadratic growth.

• cf. Spingarn-Rockafellar ’79, Pataki-Tunçel ’01, ...
Summary

- Slope is an elegant tool.
- Variational analysis is especially effective for semi-algebraic function.
- Consequences for alternating projections, steepest descent, active sets.
Thank you.
Spectral sets

How to see this structure in \textit{eigenvalue optimization}?
Spectral sets

How to see this structure in eigenvalue optimization?

Consider $S^n := \{n \times n$ symmetric matrices$\}$ and the eigenvalue map

$$A \mapsto (\lambda_1(A), \ldots, \lambda_n(A))$$

where

$$\lambda_1(A) \leq \ldots \leq \lambda_n(A).$$
How to see this structure in eigenvalue optimization?

Consider $\mathbb{S}^n := \{n \times n \text{ symmetric matrices}\}$ and the eigenvalue map

$$A \mapsto (\lambda_1(A), \ldots, \lambda_n(A))$$

where

$$\lambda_1(A) \leq \ldots \leq \lambda_n(A).$$

Spectral sets:

$$\lambda^{-1}(Q) = \{A \in \mathbb{S}^n : \lambda(A) \in Q\}$$
Spectral sets

How to see this structure in eigenvalue optimization?

Consider $S^n := \{n \times n$ symmetric matrices\} and the eigenvalue map

$$A \mapsto (\lambda_1(A), \ldots, \lambda_n(A))$$

where

$$\lambda_1(A) \leq \ldots \leq \lambda_n(A).$$

Spectral sets:

$$\lambda^{-1}(Q) = \{A \in S^n : \lambda(A) \in Q\}$$

where $Q$ is permutation-invariant.
Spectral sets

How to see this structure in eigenvalue optimization?

Consider $S^n := \{n \times n \text{ symmetric matrices}\}$ and the eigenvalue map

$$A \mapsto (\lambda_1(A), \ldots, \lambda_n(A))$$

where

$$\lambda_1(A) \leq \ldots \leq \lambda_n(A).$$

Spectral sets:

$$\lambda^{-1}(Q) = \{A \in S^n : \lambda(A) \in Q\}$$

where $Q$ is permutation-invariant.

(e.g. $\{A : \lambda_n(A) \leq 1\}$ or $\{A : \text{rank } A \leq r\}$)
Spectral sets

How to see this structure in eigenvalue optimization?

Consider $S^n := \{n \times n$ symmetric matrices\} and the eigenvalue map

$$A \mapsto (\lambda_1(A), \ldots, \lambda_n(A))$$

where

$$\lambda_1(A) \leq \ldots \leq \lambda_n(A).$$

Spectral sets:

$$\lambda^{-1}(Q) = \{A \in S^n : \lambda(A) \in Q\}$$

where $Q$ is permutation-invariant.

(e.g. $\{A : \lambda_n(A) \leq 1\}$ or $\{A : \text{rank } A \leq r\}$)

- Partial smoothness of $\lambda^{-1}(Q)$ / partial smoothness of $Q$.  

Recognizing partial smoothness (Daniilidis-D-Lewis):

\[ Q \text{ partly smooth at } \lambda(\bar{X}) \text{ relative to } M \]
\[ \implies \lambda^{-1}(Q) \text{ partly smooth at } \bar{X} \text{ relative to } \lambda^{-1}(M). \]
Recognizing partial smoothness (Daniilidis-D-Lewis):

\[ Q \text{ partly smooth at } \lambda(\bar{X}) \text{ relative to } M \]
\[ \implies \lambda^{-1}(Q) \text{ partly smooth at } \bar{X} \text{ relative to } \lambda^{-1}(M). \]

Eg:

\[ |x| + |y| \leq 1 \]
\[ |\lambda_1(X)| + |\lambda_2(X)| \leq 1 \]
Recognizing partial smoothness (Daniilidis-D-Lewis):

\[ Q \text{ partly smooth at } \lambda(\bar{X}) \text{ relative to } \mathcal{M} \]
\[ \implies \lambda^{-1}(Q) \text{ partly smooth at } \bar{X} \text{ relative to } \lambda^{-1}(\mathcal{M}). \]

Eg:

\[ |x| + |y| \leq 1 \]
\[ |\lambda_1(X)| + |\lambda_2(X)| \leq 1 \]

Related work: Daniilidis-Malick-Sendov ’11