How nonsmooth optimization usually is

Adrian Lewis

ORIE Cornell

Joint work with:

J. Burke and, D. Drusvyatskiy (Washington), J. Guo (Cornell), D. Henrion (Toulouse),

A. Ioffe (Technion), J. Liang (Cambridge), M. Overton (NYU), S. Wright (Wisconsin)

Pinhas Naor Lecture Be'er Sheva, May 2018

Outline

Can we minimize nonsmooth and (maybe) nonconvex functions?

- Algorithms
  - General-purpose quasi-Newton
  - ProxDescent for composite problems
  - Primal-dual for saddlepoints
- Examples
  - Eigenvalue optimization
  - Systems control
  - Transient dynamics
  - Sparse estimation
- Geometry
  - The typical picture — partial smoothness
  - Active set philosophy and acceleration
  - Constant rank.

Nonsmooth optimization in practice

Practitioners often value optimization algorithms that are:

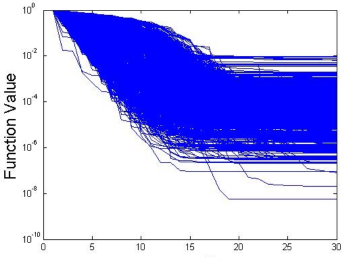
simple, reliable, intuitive, general-purpose (black-box).

Example: gradient descent for minimizing **smooth** f on $\mathbf{R}^n$.
At current iterate x, set $t = 1$:

$$\textbf{repeat} \quad x_{\text{new}} = x - t\nabla f(x); \quad t = \frac{t}{2}; \quad \textbf{until} \quad f(x_{\text{new}}) < f(x).$$

But f is often **nonsmooth**.

- Gradient descent **fails**.
  Eg: 1000 random runs on
  $f(u, v) = |u| + v^2 \quad \longrightarrow$
- Subgradient method slow.
- Bundle methods tricky.
- Fast methods structured.



Nonsmooth optimization via "smooth" BFGS

Current iterate x, and H approximating $\nabla^2 f(x)^{-1}$.

- x_{new} approximately minimizes f in quasi-Newton direction:

$$-\mathbf{R}_+ H \nabla f(x).$$

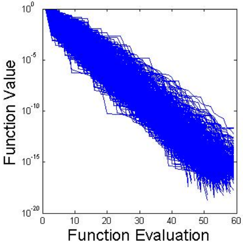
- H_{new} chosen as close to H as possible...

$$\text{measured by} \quad \text{trace}\, H^{-1} H_{\text{new}} - \log\det H_{\text{new}} \ldots$$

subject to curvature information:

$$H_{\text{new}}\big(\nabla f(x_{\text{new}}) - \nabla f(x)\big) = x_{\text{new}} - x.$$

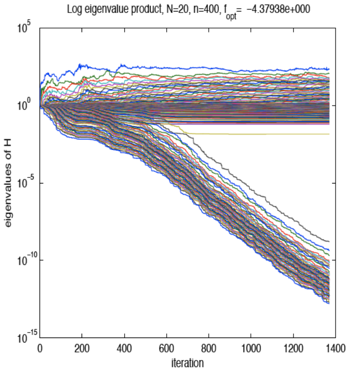
Effective for nonsmooth f too! (L-Overton '13)
Example (L-Zhang '18):
1000 random runs on
$f(u, v) = |u| + v^2 \quad \longrightarrow$
Invariably converges, at consistent linear rate. **Why?**



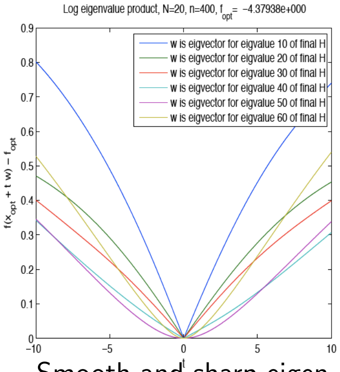
Typical "partly smooth" behavior

Example (Anstreicher-Lee '04):
Minimize product of 10 largest eigenvalues of symmetric matrix

$$(a_{ij} v^i \cdot v^j) \quad \text{for unit } v^i \in \mathbf{R}^{20} \quad (i = 1, \ldots, 20).$$



Eigenvalues of H



Smooth and sharp eigen-directions for product.

Theme: typical nonsmooth geometry

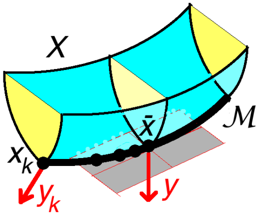
Practical optimization involves minimizing $\langle y, \cdot \rangle$ over closed $X \subset \mathbf{R}^n$ that may be

- nonsmooth
- nonconvex,

but typically

- nonpathological.

Optimization reveals **ridges**: the problem parameters y determine solutions varying over **smooth** manifolds $\mathcal{M} \subset X$, around which X is **sharp**.



Aim: illustrate this **partial smoothness**, define it, explain why it's typical, and capitalize on it.

Example: simultaneous control system stabilization

Problem (Blondel '94) Find **stable** real polynomials p, q so

$$(z^2 - 2\delta z + 1)p(z) + (z^2 - 1)q(z)$$

also stable (all roots lie in left half-plane).

- $\delta = 1$ clearly impossible;
- $\delta = 0.99999$ impossible (Blondel)
- $\delta = 0.9$? Prize: **1 kg Belgian chocolate**;
- Which δ are possible? Prize: **+1 kg**.

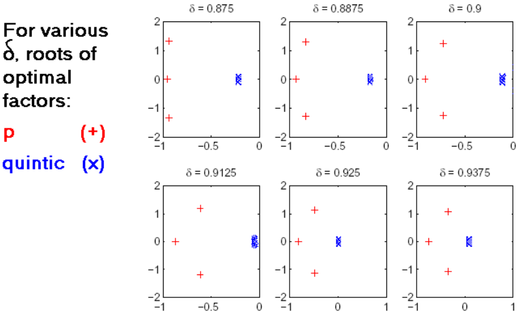
Computational approach (Burke-Henrion-L-Overton '05)
Restrict (eg) to cubic p and scalar q, minimize real t over

$$X \;=\; \big\{(p,q,t) : t \geq \operatorname{Re} z \text{ for all roots } z\big\}$$

and eat chocolate if optimal $t < 0$.

Optimal roots for chocolate problem

In this case $\mathcal{M} = \{(p, q, t) : \text{quintic has quintuple root at } t\}$.



- As parameter δ varies, solution varies **smoothly** on $\mathcal{M}$.
- Such solutions are easy to calculate algebraically.
- As $(p, q, t) \in X$ moves off $\mathcal{M}$, t increases **sharply**.

Numerical radius and control systems

Matrices Z with **field of values** satisfying

$$W(Z) \;=\; \{u^* Z u : \text{unit } u\} \;\subset\; \text{unit disk } \mathbf{D}$$

- form a compact convex set Ω, and
- have dynamics $x \leftarrow Zx$ with good transient stability.

After optimization (L-Overton '18),

- $W(Z)$ often **equals** $\mathbf{D}$, and
- such Z form a manifold $\mathcal{M}$.

Example: Any unit matrix (in Frobenius norm) with sparsity

$$\begin{bmatrix} 0 & \times & 0 & \cdots & 0 \\ 0 & 0 & \times & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & \times \\ 0 & 0 & 0 & \cdots & 0 \end{bmatrix}$$

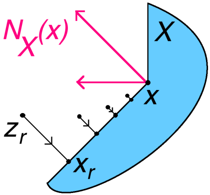
is the projection onto Ω of some $Y \notin \Omega$.
As Y varies, the projection varies over $\mathcal{M}$.

Mathematical foundations

The **normal cone** $N_X(x)$ at $x \in X$ consists of

$$n = \lim_r \lambda_r (z_r - x_r)$$

where $\lambda_r > 0$, $z_r \to x$, and x_r is a projection of z_r onto X.



The **tangent cone** $T_X(x)$ consists of $t = \lim_r \mu_r (y_r - x)$, where $\mu_r > 0$ and $y_r \to x$ in X.

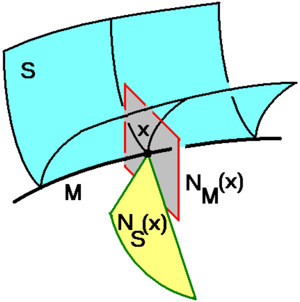
X is **(Clarke) regular** at x when these cones are polar: $\langle n, t \rangle \leq 0$.

Examples. Manifolds, convex sets, or **prox-regular** sets: points near x have unique projections onto X.

Partly smooth sets

$S \subset \mathbf{R}^n$ is **partly smooth** relative to a manifold $\mathcal{M} \subset S$ if

- S is regular throughout $\mathcal{M}$
- $\mathcal{M}$ is a **ridge** of S:
  $N_S(x)$ spans $N_{\mathcal{M}}(x)$
  for $x \in \mathcal{M}$.
- $N_S(\cdot)$ is **continuous** on $\mathcal{M}$.



Examples

- Polyhedra, relative to their **faces**
- $\{x : \text{smooth } g_i(x) \leq 0\}$, relative to $\{x : \textbf{active } g_i(x) = 0\}$
- Semidefinite cone, relative to **fixed rank** manifolds (Oustry).

Semi-algebraic sets

A good model for concrete feasible regions...

Polynomial level sets in $\mathbf{R}^n$:

$$\{x : p(x) < 0\} \quad \text{and} \quad \{x : p(x) \leq 0\}.$$

Basic sets are finite intersections of these.
Finite unions of basic sets are called **semi-algebraic**.

Semi-algebraicity is prevalent and easy to recognize,
since linear projection maps preserve it (Tarski-Seidenberg).

Typical variational problems

Theorem (Drusvyatskiy-Ioffe-L '13) For a problem $y \in \Phi(x)$, if

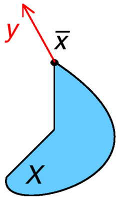
$$\text{semi-algebraic } \Phi\colon \mathbf{E} \rightrightarrows \mathbf{F} \quad \text{has} \quad \dim(\operatorname{graph}\Phi) \le \dim \mathbf{F},$$

then for almost all data y at every solution $\bar{x}$,

strong regularity: Φ^{-1} single-valued and Lipschitz near $(y, \bar{x})$.

Example Any maximizer $\bar{x}$ of $\langle y, \cdot \rangle$ over closed $X \subset \mathbf{E}$ is **critical**:

$$y \in N_X(\bar{x}).$$



Semi-algebraic X have $\dim(\operatorname{graph} N_X) \le \dim \mathbf{E}$, so, for almost all y, strong regularity holds for all $\bar{x}$. And more...

Identifiability and "active set" philosophy

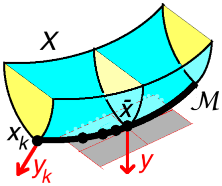
Many methods for $\max_X \langle y, \cdot \rangle$ (high-dimensional and nonsmooth) generate **asymptotically critical** $x_k \in X$:

$$\text{there exist } y_k \in N_X(x_k) \text{ such that } y_k \to y.$$

Example. Proximal point: $\rho(x_k - x_{k+1}) + y \in N_X(x_{k+1})$.

Suppose X is semi-algebraic and y is generic.
Any maximizer $\bar{x}$ lies on an **identifiable manifold** $\mathcal{M} \subset X$:
every asymptotically critical sequence eventually lies in $\mathcal{M}$.

Equivalently (almost),
X is partly smooth relative to $\mathcal{M}$,
and prox-regular at $\bar{x}$ for $y \in \operatorname{ri} N_X(\bar{x})$.
Hence low-dimensional smooth reduction
$\max_{\mathcal{M}} \langle y, \cdot \rangle$, and acceleration...



Example: composite optimization

Minimize **"simple"** nonsmooth $h\colon \mathbf{R}^m \to \mathbf{R}$ (here finite convex) composed with smooth $c\colon \mathbf{R}^n \to \mathbf{R}^m$. Around current x,

$$\tilde{c}(d) \;=\; c(x) + \nabla c(x) d \;\approx\; c(x+d).$$

Step d solves **easy** subproblem

$$\min_d \; h\big(\tilde{c}(d)\big) + \mu\|d\|^2.$$

Update step control μ: **if**

$$\text{actual decrease} \;=\; h\big(c(x)\big) - h\big(c(x+d)\big)$$

less than half

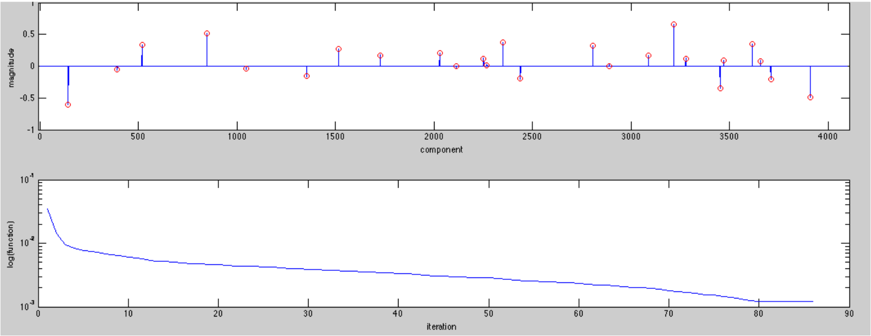
$$\text{predicted decrease} \;=\; h\big(c(x)\big) - h\big(\tilde{c}(d)\big),$$

reject: $\mu \leftarrow 2\mu$; otherwise,
accept: $x \leftarrow x + d, \;\; \mu \leftarrow \frac{\mu}{2}$.
Repeat. (L-Wright '15: ProxDescent)

Example: nonconvex regularizers for sparse estimation

$$\min_{\mathbf{x}} \|A\mathbf{x} - b\|^2 + \tau \sum_i \phi(\mathbf{x}_i) \qquad \text{(Zhao et al. '10).}$$

Random 256-by-4096 A, sparse $\hat{\mathbf{x}}$, and $b = A\hat{\mathbf{x}} + \text{noise}$.



Eventual slow linear convergence.

Acceleration

ProxDescent for $f = h\big(c(\cdot)\big)$ generates steps d_k.
Limit points $\bar{x}$ of the corresponding iterates x_k are stationary.

If h partly smooth at $c(\bar{x})$ relative to $\mathcal{N}$, and f grows quadratically, then $x_k \to \bar{x}$ (linearly).

Identifiability $\Rightarrow$ $c(x_k) + \nabla c(x_k) d_k \in \mathcal{N}$ eventually.

Classical algorithms

- use d_k to predict the active set.
- accelerate using a second-order model.

Generalize for simple h (L-Wright '15, Mifflin-Sagastizábal '05):

- "Track" $\mathcal{N}$.
- Build a second-order model from c and $h|_{\mathcal{N}}$.

Partly smooth operators

Partial smoothness of sets $X \subset \mathbf{R}^n$ illuminates optimality:

$$y \in N_X(x).$$

What about $y \in \Phi(x)$ for set-valued $\Phi\colon \mathbf{R}^n \rightrightarrows \mathbf{R}^m$ (eg monotone)?

Definition Φ is **partly smooth** at $\bar{x}$ for $\bar{y} \in \Phi(\bar{x})$ if:

- Its graph $\operatorname{gph}\Phi$ is a manifold around $(\bar{x}, \bar{y})$;
- $P\colon \operatorname{gph}\Phi \to \mathbf{R}^n$ defined by $P(x, y) = x$ is **constant rank**.

$\mathcal{M} = P(\operatorname{gph}\Phi)$ is then an identifiable manifold for $\bar{y} \in \Phi(x)$.

Partial smoothness and primal-dual methods

For convex f and g and a matrix A, **saddlpoints** for

$$\min_x \max_y \left\{ f(x) + y^T A x - g(y) \right\}$$

satisfy

$$\begin{bmatrix} 0 \\ 0 \end{bmatrix} \in \Phi \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} \partial f & -A^T \\ A & \partial g \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix}.$$

(Chambolle-Pock '11) seeks saddlepoints by updating (x, y) to

$$\begin{aligned} x_{\text{new}} \quad & \text{minimizing} \quad f(\cdot) + \frac{1}{2}\| \cdot - x + A^T y\|^2 \\ y_{\text{new}} \quad & \text{minimizing} \quad g(\cdot) + \frac{1}{2}\Big\| \cdot - y + A(x - 2x_{\text{new}})\Big\|^2. \end{aligned}$$

If f, g are partly smooth relative to $\mathcal{M}, \mathcal{N}$, then Φ is partly smooth relative to $\mathcal{M} \times \mathcal{N}$. Hence identification (Lewis-Zhang '18).

Summary

- Appealingly simple nonsmooth algorithms (like BFGS).
- Diverse examples: classical, spectral, control...
- Typical partly smooth geometry of "ridges":
  - Each ridge is a smooth manifold;
  - Around the ridge, the set is "sharp".
- Partial smoothness is typical (especially if semi-algebraic)...
- ...and active-set methods depend on it.