Nonsmooth geometry and active sets

Adrian Lewis

ORIE Cornell

Optimization, Games and Dynamics Paris, November 2011

Outline

- Partly smooth functions:
  - power
  - ubiquity
  - elegance
- BFGS and nonsmoothness
- A composite proximal algorithm
- Semi-algebraic sets and generic variational geometry
- The foundations of active-set methods

Example: minimizing a max-function

Suppose $\bar{x} \in \mathbf{R}^n$ minimizes a pointwise max of smooth functions

$$f(x) = \max_{i \in I} f_i(x),$$

with affine-independent $\nabla f_i(\bar{x})$ for i in the active set

$$\bar{I} \; = \; \big\{ i : f_i(\bar{x}) = f(\bar{x}) \big\} \; = \; I(\bar{x}).$$

Since f is smooth on the active manifold

$$\mathcal{M} \; = \; \{x : I(x) = \bar{I}\},$$

classical calculus shows Clarke stationarity: zero lies in

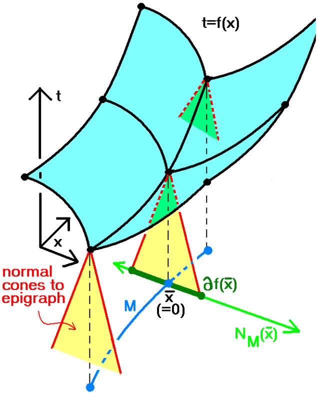
$$\Big\{ \sum_{i \in \bar{I}} \lambda_i \nabla f_i(\bar{x}) : \lambda \geq 0, \; \sum_{i \in \bar{I}} \lambda_i = 1 \Big\},$$

and this set is just the subdifferential

$$\partial f(\bar{x}) \; = \; \mathrm{conv}\{\lim \nabla f(x^r) : x^r \to \bar{x}\}.$$

Partial smoothness of f relative to $\mathcal{M}$

- Good behavior on the active manifold: as $x \in \mathcal{M}$ varies, $f(x)$ varies smoothly and $\partial f(x)$ varies continuously.
- Prox-regularity: points near $\big(\bar{x}, f(\bar{x})\big)$ have unique nearest points in the epigraph $\big\{(x,t) : t \geq f(x)\big\}$.
- Sharpness: $\partial f(\bar{x})$ spans the normal space $N_{\mathcal{M}}(\bar{x})$.



What the active manifold captures

Assume nondegeneracy: $0 \in \operatorname{ri} \partial f(\bar{x})$ ("strict complementarity").

Active set methods Approximately stationary points lie on $\mathcal{M}$:

$$x^r \to \bar{x},\ y^r \to 0,\ y^r \in \partial f(x^r) \quad \Rightarrow \quad x^r \in \mathcal{M} \text{ eventually.}$$

We call such $\mathcal{M}$ identifiable (Wright '93).

Partly smooth 2nd-order conditions Around $\bar{x}$,

$$f \text{ grows at least quadratically} \quad \Leftrightarrow \quad f|_{\mathcal{M}} \text{ grows quadratically.}$$

(verifiable simply via a Hessian.)

Sensitivity analysis In this case, $\mathcal{M}$ consists of *all* nearby approximately stationary points: for small $\delta > 0$,

$$\mathcal{M} \ = \ (\partial f)^{-1}(\delta B) \ \text{ locally around } \bar{x}.$$

These properties involve only f, **NOT** its algebraic presentation.

Example: minimizing eigenvalue products via BFGS

The active manifold emerges, even without explicit structure in f. Given

$$A \in \mathbf{S}^{20}_{+} \quad \text{(the 20-by-20 positive definite matrices)}$$

consider an eigenvalue-product problem (Anstreicher-Lee '04)

$$\min \Big\{ \prod_{i=1}^{14} \lambda_i(A \circ X) : X \in \mathbf{S}^{20}_{+},\ X_{ii} = 1 \ \forall i \Big\}.$$

Numerically, the optimal $\bar{X}$ has $\lambda_{14}(A \circ \bar{X})$ having multiplicity 9:

$$\lambda_5 > \lambda_6 = \cdots = \lambda_{14} > \lambda_{15}.$$

Matrix analysis predicts partial smoothness relative to a manifold $\mathcal{M}$ of dimension $\frac{9 \cdot 10}{2} - 1 = 44$. **We "see" $\mathcal{M}$ numerically!**

Minimization by BFGS

To minimize smooth $f\colon \mathbf{R}^n \to \mathbf{R}\ldots$

Current iterate $x \in \mathbf{R}^n$ and positive definite $H \approx \nabla^2 f(x)^{-1}$. Define

$$p = -H\nabla f(x), \quad x_{\text{new}} = x + \bar{\alpha} p,$$

where step $\bar{\alpha} > 0$ chosen by line search (eg doubling and bisection) on $\phi(\alpha) = f(x + \alpha p)$ to satisfy Wolfe conditions:

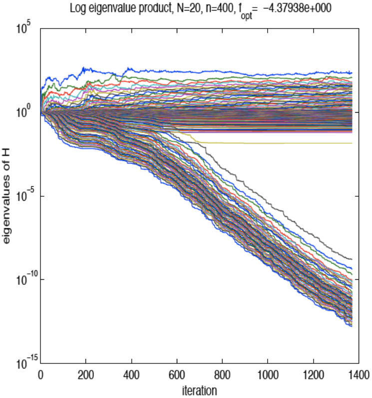
$$\phi(\bar{\alpha}) - \phi(0) < \frac{1}{3}\phi'(0)\bar{\alpha} \ \text{ and } \ \phi'(\bar{\alpha}) > \frac{2}{3}\phi'(0).$$

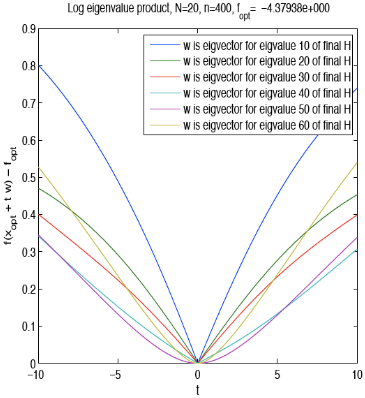
Update H and **repeat.**

- In practice, if feasible, BFGS is often most popular.
- In theory, BFGS converges for convex coercive f (Powell '76), but may fail for $\mathbf{C}^\infty$ nonconvex f (Dai '02).
- BFGS often works well for nonsmooth f (Lemaréchal '82)!

Revealing the active manifold numerically

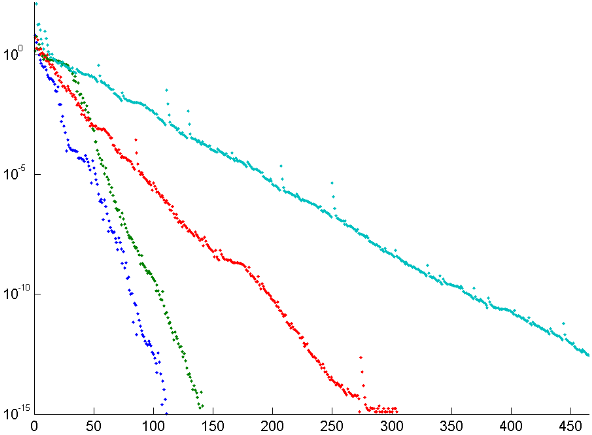
For Anstreicher-Lee, **44** ($= \dim \mathcal{M}$) H-eigenvalues $\to 0\ldots$





…and the corresponding eigenspace is tangent to $\mathcal{M}$:
the objective is smooth along $\mathcal{M}$ and “sharp” orthogonally.

BFGS for nonsmooth optimization (L-Overton '10)



Function values for BFGS applied to
$f(x, y) = w|y - x^2| + (1 - y)^2$, with $w = 1, 2, 4, 8$.

A conjecture

Apply BFGS to any "concrete" Lipschitz $f : \mathbf{R}^n \to \mathbf{R}$, with random initial point and H. Then almost surely:

- function values converge linearly;
- limit points of iterates are Clarke stationary;
- assuming convergence to a partly smooth point, the eigenstructure of H reveals the active manifold.

"Concrete" might mean semi-algebraic — graph of f a finite union of sets, each defined by finitely-many polynomial inequalities.

What if we assume more structure?
How, then, are active manifolds useful?

Composite optimization: the framework

Solve

$$\min_{x \in \mathbf{R}^n} h\big(c(x)\big)$$

for given functions

$$\begin{array}{lll} \text{nonsmooth} & h\colon \mathbf{R}^m \to \mathbf{R} & \text{finite and convex} \\ \mathbf{C}^2\text{-smooth} & c\colon \mathbf{R}^n \to \mathbf{R}^m. & \end{array}$$

Key computational assumption

"Structure" in h lets us easily solve proximal linearizations

$$\min_{d \in \mathbf{R}^n} \; h\big(\tilde{c}(d)\big) + \mu \|d\|^2,$$

for linear approximations $\tilde{c}$.

(Extensions allow prox-regular and extended-valued h.)

A proximal algorithm (L-Wright '09)

Current iterate x, prox parameter $\mu > 0$.
Linear approximation

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

Find the unique proximal step $d(x,\mu)$ minimizing

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

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: $x \leftarrow x + d, \quad \mu \leftarrow \frac{\mu}{2}$.
Repeat.

Example: exact penalties

Replace constrained optimization

$$\min_x \left\{ f(x) : g_i(x) \leq 0 \right\}$$

by unconstrained minimization of

$$f(x) + \nu \sum_i g_i^+(x) \;=\; h\big(c(x)\big)$$

(for some $\nu > 0$), where

$$c = (f, g_1, \ldots, g_k), \quad h(f, g_1, \ldots, g_k) = f + \nu \sum_i g_i^+.$$

Easy proximal linearizations

$$\min_d \;\; a_0^T d + \sum_i (a_i^T d + b_i)^+ + \mu \|d\|^2$$

(via specialized quadratic programming).

Related ideas: Yuan '85, Burke '85, Fletcher-Sainz de la Maza '89, Wright '90, KNITRO (Byrd et al. '05), Friedlander et al. 07.

Examples: Compressive sensing...

(Candès, Donoho, Tao et al. '06...)

We seek sparse solutions to linear systems $Ex = g$ via

$$\min_x \ \|Ex - g\|^2 + \tau\|x\|_1.$$

In statistics, LASSO and LARS (Tibshirani et al. '96, '04) similar.

Proximal linearizations are separable:

$$\min_{d \in \mathbf{R}^n} \ a^T d + \tau\|x + d\|_1 + \mu\|d\|^2.$$

Need just $O(n)$ operations: implemented as SpaRSA

(Wright-Nowak-Figueiredo '09)

Analogously, for low-rank X satisfying a linear system $E(X) = g$, Candès et al. '08 suggest

$$\min_X \|E(X) - g\|^2 + \tau\|X\|_*,$$

where $\|\cdot\|_*$ is the nuclear norm (sum of singular values).

Speed

The proximal algorithm is

- simple
- versatile
- applicable to huge problems

but slow. For example:

- $h = \mathrm{id}$ gives steepest descent with trust region radius $\frac{1}{2\mu}$.
- $c = \mathrm{id}$ gives the classical proximal point method (Rockafellar '76).

Both methods typically converge linearly but slowly.

Previous special cases use the initial step d to predict active constraints, and hence accelerate using a 2nd-order model.

Accelerating the proximal algorithm

Minimizing $h \circ c$ generates iterates x_r and proximal steps d_r.

Theorem (L-Wright '09)

Any limit point $\bar{x}$ of (x_r) is stationary.

Assume the partly smooth 2nd-order conditions (so $x_r \to \bar{x}$). In particular, h is partly smooth at $c(\bar{x})$ relative to a manifold $\mathcal{M}$.

Theorem (Hare-L '05)

Eventually $c_r = c(x_r) + \nabla c(x_r) d_r \in \mathcal{M}$.

Proof.

Use the identifiability property of $\mathcal{M}$. □

If h is simple, $\partial h(c_r)$ is computable, and orthogonal to M at c_r.

So we

- "track" M
- use 2nd-order properties of c and $h|_M$.

(Cf. earlier references and Mifflin-Sagastizábal '05.)

Structure versus intrinsic geometry

Explicit structure in the presentation of h may help us

- implement acceleration ideas
- check 2nd-order conditions for sensitivity analysis.

But our key idea, partial smoothness, is geometric: intrinsic to h.

How typically do the partly smooth 2nd-order conditions hold?

Generic strict complementarity and primal-dual nondegeneracy holds in various structured settings:

- nonlinear programs (Spingarn-Rockafellar '79)
- complementarity problems (Saigal-Simon '73)
- semidefinite programs (Alizadeh et al. '97, Shapiro '97)
- conic convex programs (Pataki-Tunçel '01)
- sublinear-smooth composites (Bonnans-Shapiro '00).

Classical results

For simplicity, fix $c = \mathrm{id}$. Given data $v \in \mathbf{R}^n$, consider conjugation:

$$\min_x \left\{ h(x) - v^T x \right\} \quad \big(= \ - h^*(v) \big).$$

Theorem (Mazur '33)

For convex coercive h and generic v, the optimal solution is unique (and also, for almost all v, nondegenerate (Drusvyatskiy-L '10).)

Theorem (Sard '42, Spingarn-Rockafellar '79)

For $\mathbf{C}^2$ h and almost all v, quadratic growth holds at all local mins.

An intrinsic approach: semi-algebraic sets

Earlier work on generic optimality relies on

> the structural presentation of h.

By contrast, we assume only that

> the graph of h is semi-algebraic.

That is, it can be presented as

> a finite union of sets, each defined by finitely-many polynomial inequalities.

But our approach is intrinsic, independent of this presentation.

We can recognize semi-algebraic sets via "quantifier elimination": linear maps preserve semi-algebraicity (Tarski-Seidenberg '31).

Furthermore, semi-algebraic sets have dimension, so, for a semi-algebraic subset of a convex set generic $\Leftrightarrow$ dense.

Prevalence of partial smoothness

Theorem (Bolte-Daniilidis-L '09)

Given semi-algebraic convex $h \colon \mathbf{R}^n \to \bar{\mathbf{R}} = \mathbf{R} \cup \{+\infty\}$, *consider*

$$\min_x \Big\{ h(x) - v^T x \Big\}.$$

For generic $v \in \operatorname{dom} h^*$ *(ensuring finite value), the unique optimal solution satisfies the partly smooth 2nd-order conditions.*

For nonconvex h, these properties generically hold around all the (finitely-many) stationary points (Drusvyatskiy-L '11).

Semi-algebraic geometry gives an excellent testbed for "concrete" variational analysis...

A semi-algebraic aside: thin subdifferential graphs

If $f : \mathbf{R}^n \to \mathbf{R}$ is smooth, ∇f has everywhere n-dimensional graph.

Theorem (Minty '62)

If $f : \mathbf{R}^n \to \bar{\mathbf{R}}$ is convex, ∂f has everywhere n-dimensional graph.

(...with computational implications for equations on the graph.)

We say $y \in \partial^P f(x)$ (the proximal subdifferential) if
some quadratic $q \leq f$ (locally) satisfies $q(x) = f(x)$, $\nabla q(x) = y$.

$\partial^P f$ usually has large graph: $2n$-dimensional (Borwein-Wang '00).

But...

Theorem (Drusvyatskiy-L-Ioffe '10)

If $f : \mathbf{R}^n \to \bar{\mathbf{R}}$ is semi-algebraic, $\partial^P f$ has everywhere n-dimensional graph.

Identifying active sets: mathematical foundations

The partly smooth 2nd-order conditions are

- powerful ✓
- ubiquitous ✓
- mathematically elegant??

Focus on the identifiability property of $\mathcal{M}$ at stationary $\bar{x}$:

$$\boxed{x^r \to \bar{x},\ y^r \to 0,\ y^r \in \partial f(x^r) \quad \Rightarrow \quad x^r \in \mathcal{M} \text{ eventually.}}$$

Call an identifiable set locally minimal if any other identifiable set contains it, locally around $\bar{x}$. When do such sets exist?

- Not always, even for finite convex f: for example $\sqrt{x_1^2 + x_2^4}$.
- Always for polyhedral (or "fully amenable") f.

Identifiable manifolds

Suppose $0 \in \partial f(\bar{x})$. We've seen:

partial smoothness $+$ nondegeneracy $\Rightarrow$ $\exists$ identifiable manifold.

Partial smoothness (and prox-regularity) at $\bar{x}$ for 0 is enough.

Theorem (Drusvyatskiy-L-Zhang '11)

The converse is also true. Manifold $\mathcal{M}$ is then locally minimal, and

$$\partial f = \partial(f + \delta_{\mathcal{M}}) \text{ locally around } (\bar{x}, 0).$$

So, in essence, partial smoothness is simple and natural.

(Note: the Mordukhovich generalized Hessian is then easy:

$$\partial^2 f(\bar{x}|0)w = \begin{cases} \nabla^2_{\mathcal{M}} f(\bar{x})w + N_{\mathcal{M}}(\bar{x}) & \text{if } w \perp N_{\mathcal{M}}(\bar{x}) \\ \emptyset & \text{otherwise,} \end{cases}$$

where $\nabla^2_{\mathcal{M}}$ is the Riemannian Hessian.)

Summary

- Partial smoothness as a conceptual tool for sensitivity and acceleration
- Nonsmooth optimization via BFGS.
- A simple and versatile proximal algorithm for composite optimization
- Generic properties in semi-algebraic variational analysis
- The foundations of active-set methods