Partial smoothness and active sets: a fresh approach

Adrian Lewis

ORIE Cornell

Joint work with: J. Liang (Cambridge)

ISMP Bordeaux 2018

Outline

Three examples identifying activity in variational problems.

- **Active set methods** for SDP.
- **Primal-dual splitting** for saddlepoints.
- **ProxDescent** for composite optimization.

Three ideas of partial smoothness:

- **Differential-geometric:** constant-rank
- **Algorithmic:** identification
- **Variational-analytic:** nonsmooth geometry…

…and their equivalence and ubiquity.

Example 1: active sets in semidefinite optimization

For $C^{(2)}$-smooth, strongly convex f, the optimal solution $\bar{X}$ of

$$\min\{f(X) : X \in \mathbf{S}^n_+\}$$

is just the zero of "gradient + normal cone" operator:

$$\Phi(X) \;=\; \nabla f(X) + \underbrace{N_{\mathbf{S}^n_+}(X)}_{X \perp Y \in -\mathbf{S}^n_+}.$$

Projected gradient iteration $X \leftarrow \big(X - \alpha \nabla f(X)\big)_+$ converges to $\bar{X}$ with $\min \|\Phi(X)\| \to 0$. If $0 \in \mathbf{ri}\,\Phi(\bar{X})$ (strict complementarity), iterates **identify** an **active manifold**: eventually,

$$X \in \mathcal{M} \;=\; \{X : \operatorname{rank} X = \operatorname{rank} \bar{X}\}.$$

Linear convergence, and faster via projected Newton steps in $\mathcal{M}$.

Example 2: primal-dual splitting

For convex f, g, p, q with p, q smooth, and a matrix A, **saddlpoints** for

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

are zeros of the monotone operator

$$\Phi \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} \partial f + \nabla p & -A^T \\ A & \partial g + \nabla q \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix}.$$

Generalized proximal point seeks saddlepoints by updating (x, y):

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

Identification for saddlepoint problems

Primal-dual splitting for

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

includes many special cases.

- $g = \delta_{\{0\}}$ (forcing $y = 0$): proximal gradient method.
- $p = 0$ and $q = 0$: (Chambolle-Pock '11, ...).

(Liang-Fadili-Peyré '18) give conditions for convergence, with

$$\min \left\| \Phi \begin{bmatrix} x \\ y \end{bmatrix} \right\| \to 0,$$

identification of active manifolds,

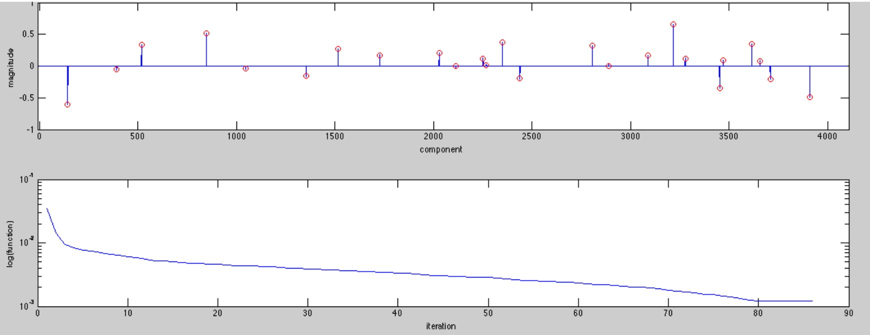
$$\begin{bmatrix} x \\ y \end{bmatrix} \in \mathcal{M} \times \mathcal{N} \quad \text{eventually},$$

and linear convergence.

Example 3: 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 support identification and linear convergence.

Composite minimization via ProxDescent

Minimize nonsmooth (but prox-friendly) $h\colon \mathbf{R}^m \to \bar{\mathbf{R}}$ composed with smooth $c\colon \mathbf{R}^n \to \mathbf{R}^m$. Around current x,

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

Proximal step d minimizes

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

Update step control μ: **if**

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

less than half

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

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

The pattern of partial smoothness

Each example involves an **active manifold** of solutions to a variational problem, **identified** by diverse algorithms.

Three, often equivalent perspectives on partial smoothness:

- Differential-geometric: constant-rank;
- Algorithmic: identification;
- Variational-analytic: nonsmooth geometry.

Partly smooth operators

Definition 1 Set-valued $\Phi\colon \mathbf{R}^n \rightrightarrows \mathbf{R}^m$ is **partly smooth** at $\bar{x}$ for $\bar{y} \in \Phi(\bar{x})$ if:

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

(Equivalently, the range and tangent spaces

$$\{0\} \times \mathbf{R}^m \quad \text{and} \quad T_{\operatorname{gph}\Phi}(x, y)$$

intersect with constant dimension as (x, y) varies.)

Definition 2 Manifold $\mathcal{M}$ **identifiable** for $\bar{y} \in \Phi(\bar{x})$ means $y_k \in \Phi(x_k)$ and $(x_k, y_k) \to (\bar{x}, \bar{y})$ implies $x_k \in \mathcal{M}$ eventually.

(Definition 1 implies Definition 2: $\mathcal{M}$ is the active manifold.)

Identification and the "active set" philosophy

Consider a high-dimensional nonsmooth **generalized equation**

$$(*) \qquad y \in \Phi(x)$$

described by set-valued $\Phi \colon \mathbf{R}^n \rightrightarrows \mathbf{R}^m$.

- **Variable** x.
- **Data** y.

If Φ is partly smooth at $\bar{x}$ for $\bar{y}$, with identifiable manifold $\mathcal{M}$, then $(*)$ reduces locally to a lower-dimensional smooth problem

$$(x, \bar{y}) \in \operatorname{gph} \Phi \quad \text{and} \quad x \in \mathcal{M},$$

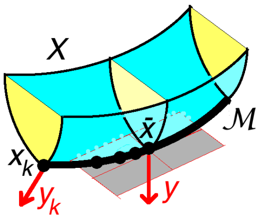
open to Newton-type acceleration.

The geometry of partial smoothness

Special case: Minimize $\langle y, \cdot \rangle$ over closed $X \subset \mathbf{R}^n$.
Critical points are zeros of

$$\Phi(x) = y + N_X(x).$$

For concrete sets X, optimization typically reveals **ridges**: varying the problem parameters y determines solutions varying over **smooth** manifolds $\mathcal{M} \subset X$, around which X is **sharp**.



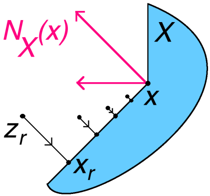
Explanation: concrete sets X are **partly smooth**.
More precisely...

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.

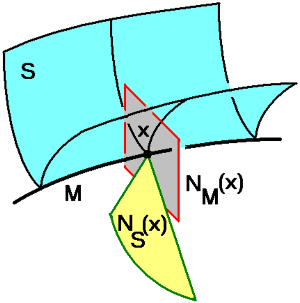
X is **(Clarke) regular** at x when these cones are polar: $\langle n, t \rangle \leq 0$.
(Eg: X **prox-regular**: points near x have unique nearest points...
...and then $\lim_r$ not needed for normals.)

Examples. Manifolds and convex sets are prox-regular, with classical normal and tangent cones/spaces.

Partly smooth sets

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

- S **prox-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.

(L '02)

Equivalent partly smooth ideas

Consider a point

$$\bar{x} \in \mathcal{M} \subset S \subset \mathbf{R}^n,$$

where the set S is prox-regular throughout the manifold $\mathcal{M}$, with normal vector $\bar{y} \in N_S(\bar{x})$. The following notions of partial smoothness are all equivalent.

- Differential-geometric: The operator N_S is partly smooth at $\bar{x}$ for $\bar{y}$, with active manifold $\mathcal{M}$.
- Algorithmic: $\mathcal{M}$ is identifiable at $\bar{x}$ for $\bar{y}$, for the operator N_S.
- Variational-analytic: The set S is partly smooth relative to $\mathcal{M}$...
  - "locally", at $\bar{x}$ for $\bar{y}$...
  - and $\bar{y} \in \operatorname{ri} N_S(\bar{x})$.

Analogous result for partly smooth **functions** f (and ∂f).

(Drusvyatskiy-L '14, L-Liang '18)

Sard-type behavior: partial smoothness is common

Consider a **semi-algebraic** generalized equation

$$y \in \Phi(x)$$

described by set-valued $\Phi \colon \mathbf{R}^n \rightrightarrows \mathbf{R}^m$.

- Variable x unknown.
- Data y **generic**.

Suppose Φ has **small graph**:

$$\dim(\operatorname{gph} \Phi) \leq m.$$

Then:

- Solution set $\Phi^{-1}(y)$ is finite (possibly empty);
- Φ is partly smooth at every solution for y;
- Near each solution, Φ^{-1} is single-valued and Lipschitz.

Example (Drusvyatskiy-L-Ioffe '16) Normal cones, subdifferentials.

Summary (from various "nebulous" perspectives)

Many algorithms (or formulations, or post-optimality analyses. . .)
for optimization (and broader variational problems)
identify (or target, or reveal)
activity (or structure)
in solutions.

The reason: a blend of smooth and nonsmooth geometry —
partial smoothness.

A simple unifying explanation:

$\boxed{\textbf{constant-rank properties of first-order conditions}}$.