Nonsmooth optimization: conditioning, convergence, and semi-algebraic models

Adrian Lewis

ORIE Cornell

International Congress of Mathematicians Seoul, August 2014

Outline

- Optimization and inverse problems via variational analysis
- A fundamental web of ideas:
  - error bounds and sensitivity to data
  - robustness to perturbation
  - angle of transversality
  - linearly convergent algorithms.
- Semi-algebraic geometry and generic regularity
- Some algorithms:
  - alternating projections
  - nonsmooth quasi-Newton
  - a prox-linear method.
- Foundations of active-set methods.

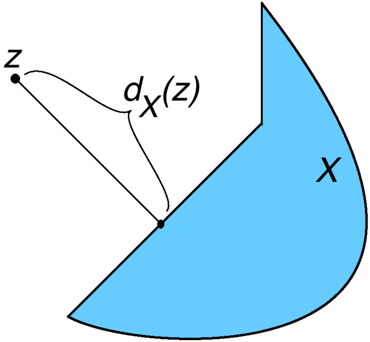
Theme

variational analysis $\longleftrightarrow$ **computational inversion**

$\downarrow$

optimization, equilibrium,
control, etc. . . via
nonsmooth geometry
of closed sets X
(maybe **nonconvex**)
in Euclidean space $\mathbf{E}$.

Key tool: **distance** d_X.



Computational inversion of $y \in \Phi(x)$

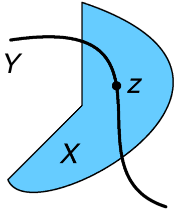
Problem

Given **set-valued mapping** Φ (between Euclidean spaces), find a **solution** x with **data** $y \in \underbrace{\Phi(x)}_{\text{easy to compute}}$. Equivalently, $x \in \underbrace{\Phi^{-1}(y)}_{\text{hard}}$.

Examples

- Linear programming: $Ax \leq y$. Define $\Phi(x) = Ax + \mathbf{R}^m_+$.
- (Banach, 1922) If $\mathrm{Id} - \Phi$ is a single-valued contraction, the iteration $x_{k+1} = y + x_k - \Phi(x_k)$ converges to the solution.
- Set intersection:
  Given sets X and Y,
  find $z \in X \cap Y$.

  Define $\Phi(z) = (X - z) \times (Y - z)$.
  Then solve $(0, 0) \in \Phi(z)$.



A circle of ideas

Consider the problem

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

locally around $(\bar{x}, \bar{y}) \in \operatorname{graph} \Phi$.

- **Regularity** — a linear **error bound**:

$$\frac{d_{\Phi^{-1}(y)}(x)}{d_{\Phi(x)}(y)} = \frac{\text{distance to a true solution}}{\text{measured error}}$$

  is bounded above. (The sup near $(\bar{x}, \bar{y})$ is the **modulus**.)
- **Sensitivity** of solutions x to data y. (Condition number)
- **Robustness** of regularity to changes in Φ.
  (Distance to ill-posedness: Demmel '87, Renegar '94.)
- Local **linear convergence** of algorithms: $d_{\Phi(x_k)}(y) \leq \alpha^k$.
- What happens for **generic data** y?

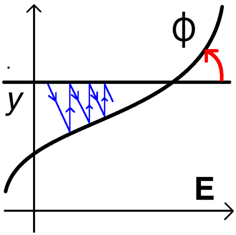
Fundamental result

Suppose graph Φ closed and $\bar{x} \in \Phi^{-1}(\bar{y}) \subset \mathbf{E}$. Key measures...

- **Modulus** of regularity.
- **Radius** (Dontchev-L-Rockafellar '03):

$$\inf\Big\{\|\text{linear } G\| : \Phi + G \text{ not regular at } (\bar{x}, \bar{y} + G\bar{x})\Big\}.$$

- **Angle** between graph Φ and $\mathbf{E} \times \{y\}$ (the coderivative criterion).



Then the quantity

$$\text{radius} = \frac{1}{\text{modulus}} = \tan(\text{angle}),$$

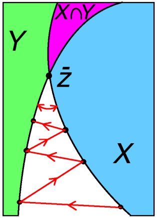
controls the linear convergence of a simple algorithm.
True for generic data y?

Example: set intersection, normals, and transversality

$$(0,0) \;\in\; \Phi(z) \;=\; (X - z) \times (Y - z).$$

Regularity is **transversality** of X and Y at $\bar{z}$: **normal cones** $N_X(\bar{z})$ and $-N_Y(\bar{z})$ intersect trivially (so, between them, **angle** > 0).

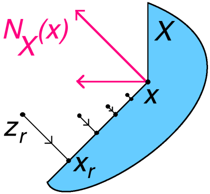
Alternating projections (von Neumann '33) then converges at linear rate depending on angle (Drusvyatskiy-Ioffe-L '13).



$N_X(x)$ at $x \in X$ consists of

$$\lim_r \lambda_r (z_r - x_r),$$

where $\lambda_r > 0$, $z_r \to x$, and $|z_r - x_r| = d_X(z_r)$.



A bad example

The problem $y \in \Phi(x)$ is **strongly regular** if

$$\Phi^{-1} \text{ single-valued and Lipschitz near } (\bar{y}, \bar{x})$$

(as in the Banach contraction mapping theorem).

But regularity can fail badly even for smooth convex optimization (and hence so does strong regularity).

There is a $\mathcal{C}^1$ strictly convex function $f \colon \mathbf{R} \to \mathbf{R}$ such that Legendre-Fenchel conjugation

$$f^*(y) = \max_x \left\{ yx - f(x) \right\},$$

or equivalently, solving $y = f'(x)$, is not regular for **any** y.

$(f')^{-1}$ is the Lebesgue singular function, so nowhere Lipschitz.
But what if f is more "concrete", or "tame" (Grothendieck)?

Semi-algebraic sets

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).

If $X, Y \subset \mathbf{R}^n$ are semi-algebraic, then, for **almost all** $z \in \mathbf{R}^n$, the intersection of $X - z$ and Y is everywhere transversal. Proof:

Theorem (Ioffe '07)... after Sard.
For almost all data y, the problem $y \in \Phi(x)$ is regular at every solution x, providing Φ has closed semi-algebraic graph.

Interlude: nonsmooth optimization via quasi-Newton

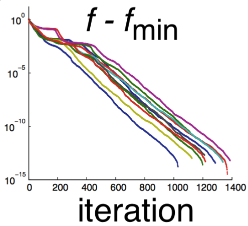
Generic regularity suggests linear convergence. Eg:
minimize **nonsmooth** Lipschitz $f : \mathbf{R}^n \to \mathbf{R}$ via Clarke criticality...

$$\text{solve} \quad 0 \in \partial f(x) = \mathrm{conv}\Big\{ \lim \nabla f(x_r) : x_r \to x \Big\}.$$

BFGS Algorithm: iterates x_k, approximate inverse Hessians H_k.

- x_{k+1} approximately minimizes f on $x_k - \mathbf{R}_+ H_k \nabla f(x_k)$.
- H_{k+1} minimizes $H \mapsto \operatorname{trace} H_k^{-1} H - \log\det H$
  subject to $H\big(\nabla f(x_{k+1}) - \nabla f(x_k)\big) = x_{k+1} - x_k$.

Popular since 1970 for smooth problems,
yet also effective when nonsmooth.
Why?? Example (L-Overton '13)
Minimize an eigenvalue product f
(nonsmooth, nonconvex, semi-algebraic,
$n = 190$), ten random initializations.
What controls the linear rate?



Generic strong regularity

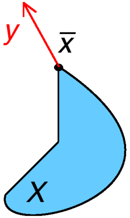
Theorem (Drusvyatskiy-Ioffe-L '13) Consider

$$\text{semi-algebraic } \Phi\colon \mathbf{E} \rightrightarrows \mathbf{F} \quad \text{with} \quad \dim(\operatorname{graph}\Phi) \leq \dim \mathbf{F}.$$

Then, for almost all data y, the problem $y \in \Phi(x)$ is strongly regular at every solution x.

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

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



Suppose X is semi-algebraic.
Then (Drusvyatskiy-L '13)

$$\dim(\operatorname{graph} N_X) \leq \dim \mathbf{E},$$

so, for almost all y, strong regularity holds for all $\bar{x}$. Hence...

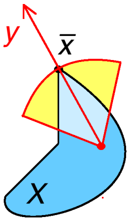
Consequences of strong regularity

For semi-algebraic optimization $\max_X \langle y, \cdot \rangle$ with generic data y, the condition $y \in N_X(x)$ is strongly regular at every maximizer $\bar{x}$.
Two consequences, with classical flavor...

Quadratic growth
(Bonnans-Shapiro '00):
there exists $\kappa > 0$ so

$$\langle y, x \rangle \leq \langle y, \bar{x} \rangle - \kappa |x - \bar{x}|^2$$

for $x \in X$ near $\bar{x}$.



Second-order condition (Mordukhovich '92)

$$(z, w) \in N_{\text{graph}\, N_X}(\bar{x}, y) \ \text{ and } \ w \neq 0 \quad \Rightarrow \quad \langle z, w \rangle < 0.$$

But we can say more...

Identifiability and "active set" philosophy

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

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

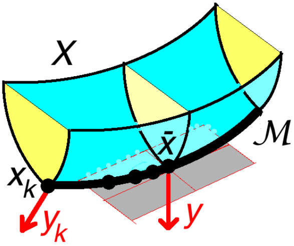
Example. Proximal point: $\rho(x_k - x_{k+1}) + y \in N_X(x_{k+1})$.

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 $\mathcal{M}$.

Hence the problem reduces to

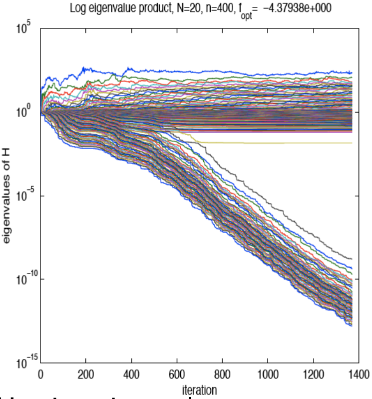
$$\max_{\mathcal{M}} \langle y, \cdot \rangle.$$

Low-dimensional and smooth.

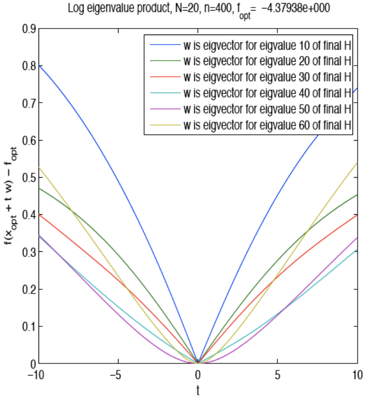


“Blind” algorithms reveal identifiable manifolds

Recall: BFGS (L-Overton '13) on an eigenvalue product problem (Anstreicher-Lee '04): $\min\left\{\prod \lambda_i(A \circ X) : X \in \mathbf{S}^{20}_{+},\ X_{ii} = 1\ \forall i\right\}$.



Hessian eigenvalues: smoothness versus sharpness



Eigenvectors predict identifiable manifold.

A prox-linear algorithm (L-Wright '08)

$\min_x \left\{ f(x) : G(x) \in Y \right\}$, where f and G are $\mathcal{C}^2$ and Y is **simple**.
Example (LASSO and LARS): $\min \left\{ |Ax - b|^2 : |x|_1 \le 1 \right\}$.

Subproblem: form linear approximations $\widetilde{f}(d) \approx f(x+d)$ and $\widetilde{G}$ at current feasible x. Since Y simple, easy to solve

$$\min_d \left\{ \widetilde{f}(d) + \mu |d|^2 : \widetilde{G}(d) \in Y \right\}.$$

Update: $x \leftarrow x^+ \approx x + d$. Specifically, x^+ feasible, with

$$\left| x^+ - (x+d) \right| \le \frac{|d|}{2} \quad \text{and} \quad \frac{f(x) - f(x^+)}{f(x) - \widetilde{f}(d)} \ge \frac{1}{2}.$$

If success, **repeat**; if not, reset $\mu \leftarrow 2\mu$ and try again.

Typically, at optimality, Y has an **identifiable manifold** $\mathcal{M}$ at $g(\bar{x})$: eventually $\widetilde{G}(d) \in \mathcal{M}$, and $\inf\{ f(x) : G(x) \in \mathcal{M} \}$ easier.

Summary

- Variational-analytic insights into computational inversion.
- Key tools: the normal cone and regularity/transversality.
- Sensitivity, error bounds, robustness, and linear convergence.
- Semi-algebraic optimization: generic regularity and identifiability.
- Quasi-Newton and prox-linear methods for nonsmooth optimization.