Nonsmooth optimization and semi-algebraic sets

Adrian Lewis

ORIE Cornell

Fields Institute Workshop on Optimization September 29, 2011

Outline

- Examples of “composite” optimization:
  - exact penalties
  - compressed sensing
  - low-rank matrix completion...
- A general-purpose proximal algorithm
- Acceleration and “partly smooth” geometry
- Semi-algebraic sets and generic variational geometry
- Wild versus tame optimization: examples
- Nonsmooth optimization via BFGS.

Composite optimization: the framework

Solve

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

for given functions

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

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 $\tilde{c}$.

A proximal algorithm

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

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

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).

Convergence theory

Subgradients: $v \in \partial h(u)$ means 0 minimizes

$$d \mapsto h(u+d) - v^T d.$$

Theorem

Minimizers $\bar{x}$ for $h \circ c$ are critical: for some Lagrange multiplier y,

$$y \in \partial h\big(c(\bar{x})\big) \quad \textit{and} \quad \nabla c(\bar{x})^* y = 0.$$

For some $\rho > 0$, the proximal step satisfies

$$\|d(x,\mu)\| \leq \rho \|x - \bar{x}\|$$

for all x near $\bar{x}$ and $\mu > 0$.

Theorem (L-Wright '09)

Limit points of the proximal algorithm are critical.

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 second-order model.

Geometry for acceleration

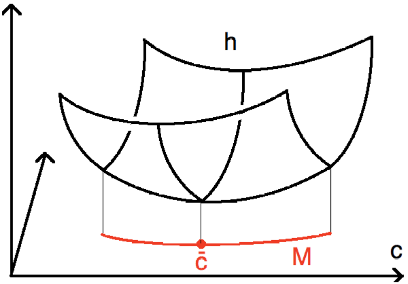
- The critical point $\bar{x}$ is nondegenerate:

$$y \in \operatorname{ri} \partial h\big(c(\bar{x})\big) \quad \text{and} \quad \nabla c(\bar{x})^* y = 0.$$

The function h is partly smooth (Lewis '03, Wright '93) relative to an active manifold M around $c(\bar{x})$:

- h is smooth on M;
- ∂h is continuous on M, and orthogonal to it (sharpness).

Eg: $h = \mathrm{dist}_P$ for P polyhedral; $M \subset P$ open facet.



Acceleration

Theorem (Hare-L '05)

Assuming nondegeneracy and partial smoothness, if the proximal algorithm generates $x_r \to \bar{x}$ and steps d_r, then eventually it identifies M:

$$c_r \;=\; c(x_r) + \nabla c(x_r) d_r \;\in\; M.$$

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

So we

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

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

Sensitivity

Partial smoothness gives nice sensitivity analysis.

Theorem (L '03)

Assume nondegeneracy, partial smoothness,

- *transversality:*

$$z \perp M \text{ at } c(\bar{x}) \;\; \text{and} \;\; \nabla c(\bar{x})^* z = 0 \quad \Rightarrow \quad z = 0,$$

- $h\big(c(\cdot)\big)$ *grows quadratically on* M *around* $\bar{x}$.

Then there's a unique local minimizer of

$$h(c(x)) - v^T x$$

near $\bar{x}$*, lying on* $c^{-1}(M)$*, and depending smoothly on* v.

Structure versus intrinsic geometry

Explicit structure in the presentation of h may help us

- implement acceleration ideas
- check second-order conditions for sensitivity analysis.

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

So, how typical is

- nondegeneracy
- partial smoothness
- quadratic growth?

For simplicity, fix $c = \mathrm{id}\ldots$

Generic optimality conditions

Generic strict complementarity, primal-dual nondegeneracy for

- nonlinear programs (Spingarn-Rockafellar '79)
- complementarity problems (Saigal-Simon '73)
- semidefinite programs (Alizadeh-Haeberly-Overton '97, Shapiro '97)
- conic convex programs (Pataki-Tunçel '01).

In our setting, given data $v \in \mathbf{R}^n$, consider $\min_x \left\{ h(x) - v^T x \right\}$.

Theorem (Mazur '33)

For convex coercive h and generic v, the optimal solution is unique.

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 $\quad$ 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*

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

For generic v *in*

$$\{v : \textit{optimal value finite}\} \quad (= \ \mathit{dom}\, h^*)$$

there's a unique optimal solution, and it satisfies

- *nondegeneracy*
- *partial smoothness relative to a unique manifold*
- *quadratic growth*
- *smooth dependence on* v.

Semi-algebraic assumptions rule out many more pathologies. . .

Example: Sard's theorem and metric regularity

Set-valued $F\colon \mathbf{R}^n \rightrightarrows \mathbf{R}^m$ is metrically regular at $\bar{x}$ for $\bar{y} \in F(\bar{x})$ if an error bound holds:

$$\frac{\operatorname{dist}(x, F^{-1}(y))}{\operatorname{dist}(y, F(x))} \quad \text{bounded near } (\bar{x}, \bar{y}).$$

Otherwise, $\bar{y}$ is critical. (Key to sensitivity/convergence analysis.)

Theorem (Sard '42)

Sufficiently smooth $F\colon \mathbf{R}^n \to \mathbf{R}^m$ have almost no critical values.

Theorem (Bolte-Daniiliidis-L '06)

Semi-algebraic $f : \mathbf{R}^n \to \mathbf{R}$ have only finitely many critical values.

Much more generally. . .

Theorem (Ioffe '07)

Noncritical values are generic for any semi-algebraic $F\colon \mathbf{R}^n \rightrightarrows \mathbf{R}^m$.

Example: 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.)

For continuous $f : \mathbf{R}^n \to \mathbf{R}$, we say $y \in \partial f(x)$ if

$$0 \ \text{ minimizes } \ d \mapsto f(x+d) - \langle y, d \rangle + o(d).$$

∂f typically 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, ∂f has everywhere n-dimensional graph.

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

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

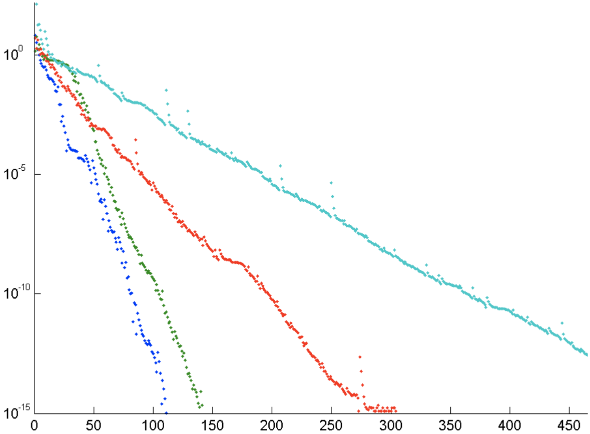
where step $\bar{\alpha} > 0$ chosen by line search (eg doubling and bisection) on $\phi(\alpha) = f(x + \alpha p)$ to satisfy Wolfe conditions:

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

Update H and **repeat.**

- In practice, if feasible, BFGS is often most popular.
- In theory, BFGS converges for convex coercive f but may fail for $\mathbf{C}^\infty$ nonconvex f (Powell '76, '84).
- BFGS often works well for nonsmooth f (Lemaréchal '82)!

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 semi-algebraic 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.

(There are small convex combinations of nearby gradients.)

Summary

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