PARTIAL SMOOTHNESS AND FAST LOCAL ALGORITHMS

Calvin Wylie

ICCOPT 2019 Berlin

Department of ORIE
Cornell University

Joint work with Adrian S. Lewis

Question 1: How to understand Newton algorithms exploiting smooth strucure for generalized equations? E.g.

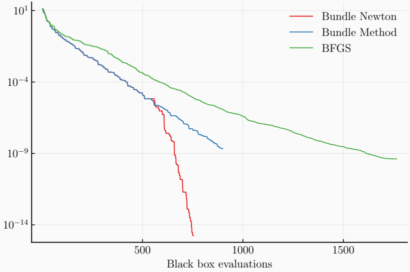
$$-\nabla f(x) \in N_Q(x)$$

The basic projection algorithm

$$x \leftarrow \mathrm{Proj}_Q(x - \gamma \nabla f(x))$$

typically identifies smooth substructure in Q. Newton acceleration?

Question 2: Superlinear convergence for black box (unstructured) nonsmooth minimization?



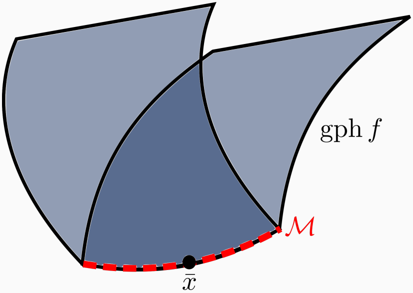
Functions and sets arising in nonsmooth optimization are typically highly structured. Around a solution is a smooth manifold of solutions to nearby problems.

- $\{\text{smooth } g_i(x) \leq 0\}$ relative to active set $\{x : g_j(x) = 0\}$
- PSD matrices $\mathbf{S}^n_+$ relative to $\{X \in \mathbf{S}^n_+ : \operatorname{rank}(X) = k\}$

Generalized notion of active constraint identification in nonlinear programming.
(Burke Moré '88, Al-Khayyal Kyparisis '91, ...)

Identifiable surfaces (Wright '93), $\mathcal{VU}$ decompositions (Mifflin Stagastizábal '00-), Partial Smoothness (Lewis '02-)...

- Partial smoothness is common – especially in the semi-algebraic case.
- Diverse first-order algorithms identify the manifold... which drives the local convergence.

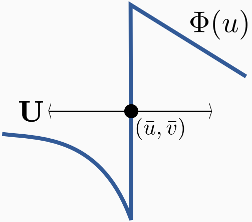


E.g. Proximal point algorithm

$$x^+ = \operatorname*{arg\,min}_y \left\{ f(y) + \frac{\rho}{2}|y - x|^2 \right\}$$

iterates eventually stay on $\mathcal{M}$.

More generally – Proximal/Projected gradient, Douglas Rachford, Primal-Dual splitting... (Liang-Fadili-Peyré ’18)



Set valued $\Phi : \mathbf{U} \rightrightarrows \mathbf{V}$ is partly smooth at $\bar{u}$ for value $\bar{v}$ if

- $\operatorname{gph}\Phi = \{(u, v) : v \in \Phi(u)\}$ is a manifold around $(\bar{u}, \bar{v})$.
- $\operatorname{proj} : \operatorname{gph}\Phi \to \mathbf{U} : (u, v) \mapsto u$ is constant rank around $(\bar{u}, \bar{v})$.

Active manifold $\mathcal{M} = \operatorname{proj}(\operatorname{gph}\Phi \text{ near } (\bar{u}, \bar{v}))$

Identification property

$$v_k \in \Phi(u_k), \quad u_k \to \bar{u}, \quad v_k \to \bar{v} \quad \Rightarrow \quad u_k \in \mathcal{M} \text{ for all large } k.$$

(Lewis-Liang ’18)

EXAMPLE: VARIATIONAL INEQUALITIES

Variational inequality with smooth F, convex partly smooth Q:

$$0 \in F(x) + N_Q(x)$$

Around a nondegenerate solution

$$-F(\bar{x}) \in \mathbf{ri}\, N_Q(\bar{x})$$

N_Q is partly smooth and

$$\mathbf{gph}\, N_Q = \mathbf{gph}\, N_{\mathcal{M}} \text{ near } (\bar{x}, -F(\bar{x}))$$

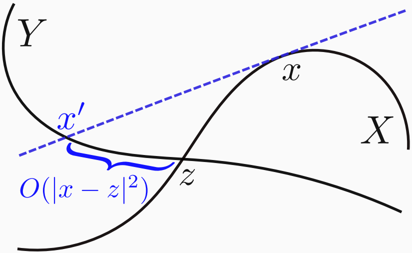
Basic Projection Algorithm

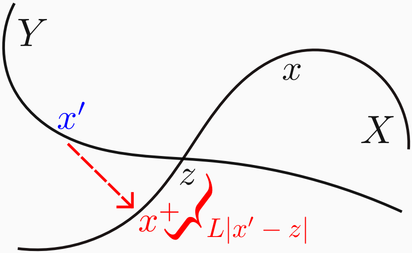
$$x^+ = \mathrm{Proj}_Q(x - F(x)) \;\Leftrightarrow\; (x - x^+) - F(x) \in N_Q(x^+)$$

$x \to \bar{x} \Rightarrow x \in \mathcal{M}$ eventually. Acceleration?

Frame $0 \in \Phi(u)$ as a manifold intersection problem

$$X = \operatorname{gph} \Phi \quad Y = \mathbf{U} \times \{0\}$$





1. Linearize X and intersect with Y.

2. Restore to X with a Lipschitz map that fixes z.

Transversality: $N_X(z) \cap N_Y(z) = \{0\}$

$u^+ = \mathrm{Proj}_{\mathcal{M}}(u') \quad v^+ = \mathrm{Proj}_{\Phi(u^+)}(0)$

Special case

$$\Phi = \text{smooth } F + \text{maximal monotone } \Psi$$

with iterate $(u, v) \in \operatorname{gph} \Phi$

- In the language of graphical derivatives (Aubin '81, Mordukhovich '80, Rockafellar-Wets), solve the tangent problem

$$v - \nabla F(u)(u' - u) \in D\Psi(u|v - F(u))(u' - u)$$

- Restore the the graph with forward-backward iteration

$$\begin{aligned} T(u) &= (I + \gamma\Psi)^{-1}(u - \gamma F(u)) \\ u^+ &= T(u') \\ v^+ &= (1/\gamma)(u' - T(u')) - F(u') + F(T(u')) \end{aligned}$$

Minimize nonsmooth $f : \mathbf{R}^n \to \mathbf{R}$ and satisfy Clarke stationarity

$$0 \in \partial f(\bar{x}) = \mathrm{conv}\left\{ \lim_{i\to\infty} \nabla f(x_i) : x_i \to \bar{x} \right\}$$

Black box setting:

- Structure of $\mathrm{gph}\, \partial f$ is unknown.
- But f is $\mathcal{C}^{(2)}$ almost everywhere with oracle returning $f(x), \nabla f(x), \nabla^2 f(x)$.

Seek a bundle $S = \{s_1, \ldots, s_k\}$ with small diameter

$$\mathrm{diam}(S) \;=\; \max\{|s - s'| : s, s' \in S\}$$

and small optimality measure

$$\Theta(S) \;=\; \mathrm{dist}\left(0, \mathrm{conv}(\nabla f(S))\right)$$

Define the linear and quadratic approximations

$$l_s(x) = f(s) + \langle \nabla f(s), x - s \rangle$$
$$q_s(x) = l_s(x) + \frac{1}{2}\langle x - s, \nabla^2 f(s)(x - s) \rangle$$

and simplex $\Delta = \left\{\lambda \geq 0 : \sum_{s \in S} \lambda_s = 1\right\}$

Bundle Newton Algorithm

- Choose $\lambda \in \Delta$ such that $|\sum_s \lambda_s \nabla f(s)| = \Theta(S)$
- Choose $\hat{x} \in \operatorname{arg\,min}\{\sum_s \lambda_s q_s(x) : l_s(x) \text{ equal for all } s\}$
- Replace point in S with $\hat{x}$ to minimize $\Theta(S)$

Reduces to classical Newton's method when $|S| = 1$

Consider a max function

$$f(x) = \max_{i=1,\ldots,k} f_i(x)$$

for some $\mathcal{C}^{(2)}$ functions $f_i : \mathbf{R}^n \to \mathbf{R}$. Black box returns function values, gradients, Hessians but **no knowledge of underlying functions** f_i.

Suppose f is partly smooth relative to

$$\mathcal{M} = \{x : f_i(x) \text{ equal for all } i\}$$

at a nondegenerate minimizer $\bar{x} \in \mathcal{M}$. (Classical second order conditions)

$$\Rightarrow \text{Local quadratic convergence}$$

A SIMPLE BUNDLE NEWTON METHOD

Weakly convex objective

$$f + \frac{\eta}{2}|\cdot|^2 \text{ is convex for large } \eta,$$

starting from a full bundle $S = (s_1, \ldots, s_k)$ where

$$\text{each } s_i \in \{x : f_i(x) > f_j(x)\ (j \neq i)\},$$

algorithm converges to $\bar{x}$ at a k-step quadratic rate.

- Partial smoothness $\Rightarrow \hat{x} - \bar{x} = O(|\bar{x} - S|^2)$.
- Nondegeneracy $\Rightarrow \Theta(\cdot)$ identifies the activity regions, and we maintain full bundles.
- Weak convexity $\Rightarrow$ every point in S will be updated after at most k iterations.

When f is a max function,

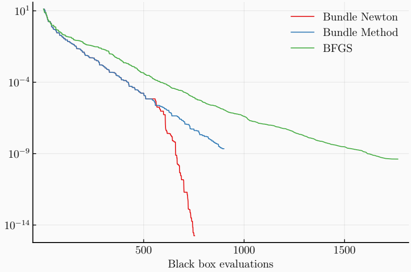
$$k = \dim(\partial f(\bar{x})) + 1$$

Nonsmooth optimization methods suggest subdifferential information as they progress. Apply standard global algorithm (e.g. Bundle method, BFGS, ...) to find a set of points Ω near minimizer $\bar{x}$.

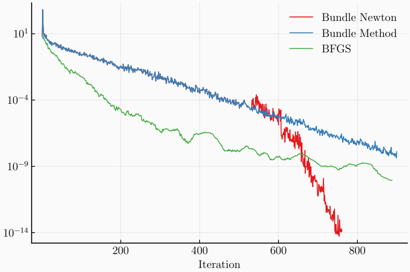
- $\partial f(\bar{x}) \approx \operatorname{conv}(\nabla f(\Omega))$
- $\hat{k} = \operatorname{rank}\left(\begin{bmatrix} \nabla f(x) \\ 1 \end{bmatrix} : x \in \Omega\right)$
- Choose $S \subset \Omega$ with $|S| = \hat{k}$.

EXAMPLE: EIGENVALUE OPTIMIZATION

$f - \min f$



$\lambda_1(A(x)) - \lambda_6(A(x))$



$$f(x) = \lambda_{\max}(A(x)), \quad A(x) = A_0 + \textstyle\sum_{i=1}^{50} x_i A_i, \quad A_0, \ldots, A_m \in \mathbf{S}_+^{25}$$

$$\mathcal{M} = \{x : \lambda_{\max}(A(x)) \text{ has multiplicity } 6\} \quad \dim \mathcal{M} = 30$$

- A.S. Lewis, C.J.S. Wylie. “Active-set Newton methods and partial smoothness”. *http://arxiv.org/abs/1902.00724*
- A.S. Lewis, C.J.S. Wylie. “A simple Newton method for local nonsmooth optimization”. *http://arxiv.org/abs/1907.11742*

Thank you!