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 structure for generalized equations? E.g.

\[-\nabla f(x) \in N_Q(x)\]

The basic projection algorithm

\[x \leftarrow \text{Proj}_Q(x - \gamma \nabla f(x))\]

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

Generalized notion of active constraint identification in nonlinear programming.

(Burke Moré ’88, Al-Khayyal Kyparisis ’91, ...)

Identifiable surfaces (Wright ’93), 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

\[
x^+ = \arg\min_y \left\{ f(y) + \frac{\rho}{2} |y - x|^2 \right\}
\]

iterates eventually stay on \( \mathcal{M} \).

More generally – Proximal/Projected gradient, Douglas Rachford, Primal-Dual splitting... *(Liang-Fadili-Peyré ’18)*
Set valued $\Phi : U \Rightarrow V$ is partly smooth at $\bar{u}$ for value $\bar{v}$ if

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

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

Identification property

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

(Lewis-Liang ’18)
Variational inequality with smooth $F$, convex partly smooth $Q$:

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

Around a nondegenerate solution

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

$N_Q$ is partly smooth and

$$\text{gph } N_Q = \text{gph } N_M \text{ near } (\bar{x}, -F(\bar{x}))$$

Basic Projection Algorithm

$$x^+ = \text{Proj}_Q(x - F(x)) \iff (x - x^+) - F(x) \in N_Q(x^+)$$

$x \to \bar{x} \Rightarrow x \in M$ eventually. Acceleration?
Frame $0 \in \Phi(u)$ as a manifold intersection problem

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

1. Linearize $X$ and intersect with $Y$.

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

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

$$u^+ = \text{Proj}_M(u') \quad v^+ = \text{Proj}_{\Phi(u^+)}(0)$$
ACCELERATED SPLITTING

Special case

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

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

- 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

\[
\begin{align*}
T(u) &= (I + \gamma \Psi)^{-1}(u - \gamma F(u)) \\
u^+ &= T(u') \\
v^+ &= (1/\gamma)(u' - T(u')) - F(u') + F(T(u'))
\end{align*}
\]
Minimize nonsmooth \( f : \mathbb{R}^n \to \mathbb{R} \) and satisfy Clarke stationarity

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

Black box setting:

• Structure of \( \text{gph} \partial f \) is unknown.
• But \( f \) is \( C^2 \) 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

\[
\text{diam}(S) = \max \{|s - s'| : s, s' \in S\}
\]

and small optimality measure

\[
\Theta(S) = \text{dist} (0, \text{conv}(\nabla f(S)))
\]
A SIMPLE BUNDLE NEWTON METHOD

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 \( \Delta = \{ \lambda \geq 0 : \sum_{s \in S} \lambda_s = 1 \} \)

Bundle Newton Algorithm

- Choose \( \lambda \in \Delta \) such that \( |\sum_{s} \lambda_s \nabla f(s)| = \Theta(S) \)
- Choose \( \hat{x} \in \text{arg min}\{\sum_{s} \lambda_s q_s(x) : l_s(x) \text{ equal for all } s\} \)
- 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 \( C^{(2)} \) functions \( f_i : \mathbb{R}^n \to \mathbb{R} \). Black box returns function values, gradients, Hessians but \textbf{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 \( \bar{x} \in \mathcal{M} \). \textbf{(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,

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

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

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

\[ \lambda_1(A(x)) - \lambda_6(A(x)) \]

\[ f(x) = \lambda_{\max}(A(x)), \quad A(x) = A_0 + \sum_{i=1}^{50} x_i A_i, \quad A_0, \ldots, A_m \in S_{+}^{25} \]

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


Thank you!