Science Fiction in Nonsmooth Optimization

Robert Mifflin

http://www.math.wsu.edu/faculty/mifflin

Work with C. Sagastizábal dedicated to Claude Lemarechal who once said, ”superlinear convergence in nonsmooth optimization is science fiction”

SIOPT 2011, Darmstadt

Grant support: NSF DMS 0707205, AFOSR FA9550-08-1-0370 and SOARD
<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1  Motivation with infinite max functions</td>
<td>4</td>
</tr>
<tr>
<td>2  $\nuU$-theory, $\nuU$-Lagrangians &amp; primal-dual tracks</td>
<td>5</td>
</tr>
<tr>
<td>2.1 Primal track to $\bar{x}$</td>
<td>6</td>
</tr>
<tr>
<td>2.2 Dual track to $0 \in \partial f(\bar{x})$</td>
<td>7</td>
</tr>
<tr>
<td>3  Approximating primal-dual tracks</td>
<td>8</td>
</tr>
<tr>
<td>3.1 Bundle approximation and relation to proximal points</td>
<td>9</td>
</tr>
<tr>
<td>4  Newton-like corrector-predictor $\nuU$ algorithm</td>
<td>10</td>
</tr>
<tr>
<td>4.1 Ideal iteration and line search</td>
<td>11</td>
</tr>
<tr>
<td>4.2 Convergence properties</td>
<td>16</td>
</tr>
<tr>
<td>4.3 Numerical results for a quasi-Newton version</td>
<td>17</td>
</tr>
</tbody>
</table>
1 Motivation

\[ \min_{x \in \mathbb{R}^n} f(x), \quad f \text{ convex} \quad (\text{lower } C^2 \text{ in the future}), \]

know only one subgradient of \( f \) at each \( x \).

Fast algorithms need to identify some "curvature";

only possible if Smooth Substructure exists.

Goal is to exploit natural structure implicitly,

including nonsmoothness,

without adding extraneous structure

from barrier or smoothing functions.
Convex nonsmooth function [Lewis+Overton, 2008], n=8
\[ \sqrt{x^TAx} + x^TBx \] \( A = \text{diag}(1,0,1,0,...), \ B = \text{diag}(1,...,1/n^2) \)

smooth BFGS algorithm
249 evaluations
2 \( \mathcal{VU} \)-theory & primal-dual tracks

A pdg-structured example

\[
f(x_1, x_2, x_3) = \frac{1}{2}x_1^2 + \frac{1}{2}\sqrt{(x_1^2 - 2x_2)^2 + (x_3 - x_2)^2}
\]

minimizer \( \bar{x} = (0, 0, 0) \)

\[
\partial f((0, 0, 0))
\]

For \( \bar{g} \in \partial f(\bar{x}) \) \( \mathcal{V} = \text{lin}(\partial f(\bar{x}) - \bar{g}) \) and \( \mathcal{U} := \mathcal{V}^\perp \)
A view of $f$ on $\mathcal{V}$-space
2.1 Primal track to $\bar{x}$

$U$-Lagrangian:  

$$L^\bar{g}_U(u) := \inf_{v \in V} \{ f(\bar{x} + u \oplus v) - \langle \bar{g}, v \rangle \} = f(\bar{x} + u \oplus v(u)) - \langle \bar{g}, v(u) \rangle$$

[LeMarechal, Oustry, Sagastizabal, 2000]

$f$ and $L^0_U$ on the $U$-space

$\rightarrow L^\bar{g}_U(0) = f(\bar{x}), \quad \rightarrow L^\bar{g}_U(u) \in C^1(U)$

$\rightarrow$ minimizer $v = v(u)$ generates trajectory smooth tangent to $U$
if $\forall \bar{g} \in ri \partial f(\bar{x})$ the minimizer $v(u)$ is the same

$\exists$ primal track $\chi(u) := \bar{x} + u \oplus v(u)$

with $f(\chi(u)) = L_\mathcal{U}^0(u)$

?what about $\nabla L_\mathcal{U}^0(u)$?
2.2 Dual track to $0 \in \partial f(\bar{x})$

For $\chi(u) = \bar{x} + u \oplus \nu(u)$

$\gamma(u) := \text{argmin} \left\{ |g|^2 : g \in \partial f(\chi(u)) \right\}$

$\nabla L_0^U(u) = U(\chi(u))$-component of $\gamma(u)$

$\left(\chi(u), \gamma(u)\right) \to (\bar{x}, 0)$ as $u \to 0$

Good primal-dual track $\leftrightarrow L_0^U \in C^2 + 0 \in ri \partial f(\bar{x})$

allows for a $\chi(u)$-restricted Newton method to minimize $f$
3 Approximating primal-dual tracks

Fundamental theoretical result:

Proximal Points are on the primal track

If $\bar{g} = 0 \in ri\partial f(\bar{x})$, then for all $x \approx \bar{x}$ there exists $u(x) :$

$$p(x) := \arg\min \left\{ f(y) + \frac{1}{2} \mu |y - x|^2 \right\} = \chi(u(x))$$

even with $\mu = \mu(x) : \mu(x)|x - \bar{x}| \to 0$ as $x \to \bar{x}$

$\Rightarrow$ use a bundle subroutine
to approximate the prox
and estimate the pair
$(\chi(u), \gamma(u))$ by $(p, s)$
3.1 Bundle approximation

With bundle \((y_i, f_i, g_i)\), recursively build \(\tilde{f}\), a \(V\)-model for \(f\) near \(x\), and find

\[
\chi\text{-qp solution: } p := \arg\min \left\{ \tilde{f}(y) + \frac{1}{2} \mu |y - x|^2 \right\} \approx \chi(u(x))
\]

\[
\gamma\text{-qp solution: } s := \arg\min \left\{ |g|^2 : g \in \partial \tilde{f}(p) \right\} \approx \gamma(u(x))
\]

UNTIL “good enough”: \(\varepsilon \leq (\sigma/\mu)|s|^2\)

By-product: local \(\mathcal{V}\mathcal{U}\)-decomposition, \(\mathcal{V}\mathcal{U}\)
4 Newton-like corrector-predictor $\mathcal{VU}$ algorithm

Given $x$ and a bundle:

- **Corrector step:** Solve $(\chi - \text{and } \gamma\text{-qp})$’s ending
  with $p, s, \mathcal{V}$, where $\mathcal{V}^T s \approx \nabla L_0^\mathcal{V}$, and
determine $H$, a $\mathcal{U}$-Hessian $\approx \nabla^2 L_0^\mathcal{U}$

- **Predictor step:** Solve $H \Delta u = -\mathcal{V}^T s \Rightarrow x^+ = p + \mathcal{V} \Delta u$
4.1 Ideal iteration

\[ \chi(u) \]

\[ x \]

\[ U - \text{Newton step} \]

\[ bundle \]

\[ steps \]
4.1 Candidates for $x^+$, $p^+$

Line search

if $f(p_{cand}^+) < f(p)$
4.1 Line search if candidates fail $f$-descent

\[ \chi(u) \]

\[ x \]

\[ x^+ \]

\[ x_{\text{cand}}^+ \]

\[ p_{\text{cand}} \]

\[ x^+ \]

\[ p \]

\[ x \]

To find $f(x^+ \le f(p)$
4.1 Line search to good $x^+$

\[ \chi(u) \]

...and repeat the process at most once
4.1 Second bundle run to good $p^+$

convergent

even if no $\chi(u)$
4.2 Convergence properties

1. If infinite number of inner bundle steps, this sequence converges to a minimizer of $f$

2. If the decreasing sequence $\{f(p)\}$ is infinite, then
   - either $f$ unbounded below,
   - or $\{s\} \to 0$ and any $\text{acc}(\{p\})$ minimizes $f$

3. If a primal-dual track to a strong minimizer pair $(\bar{x}, 0)$ exists and
   - $\frac{\sigma}{\mu^2} = O(|s^-|^2)$,
   - bounded $\{H^{-1}\}$,
   - $\text{acc}(\{U\})$ → a basis for $\mathcal{U}$ ([Daniilidis, Sagastizabal, Solodov, 2009]),
   - Dennis-Moré-like condition for $\{H\}$,
   - $s - \gamma = o(|s|)$ or $o(|\gamma|)$

then $\{p\}$ converges superlinearly to $\bar{x}$
4.3 Preliminary numerical results for a quasi-Newton version

\[ H = \mathbf{u}^\top H_{qN} \mathbf{u} \]

where \( H_{qN} = BFGS(p - p^-, s - s^-) \) is an \( n \times n \) matrix, with updating started at or after iteration 3 when a sufficiently large curvature is found for initial scaling of the identity
## Summary of results

<table>
<thead>
<tr>
<th></th>
<th>2d-U1</th>
<th>3d-EX</th>
<th>3d-U2</th>
<th>3d-U1</th>
<th>3d-U0</th>
<th>MAXQUAD</th>
</tr>
</thead>
<tbody>
<tr>
<td>f/g Ac</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>N1CV2</td>
<td>38 7</td>
<td>103 7</td>
<td>55 7</td>
<td>61 7</td>
<td>30 7</td>
<td>156 8</td>
</tr>
<tr>
<td>VU.qN</td>
<td>19 16</td>
<td>35 14</td>
<td>36 17</td>
<td>27 16</td>
<td>32 16</td>
<td>72 12</td>
</tr>
</tbody>
</table>

**VU.qN**: BFGS version with $\sigma = 0.5$, heuristic $\mu$-update rules when $\mu$ too large, adequate $V$-approximation line search, last $p$ replaced by best $p$ at bundle termination and a possible line search along $v\Delta u$ to satisfy a Wolfe directional derivative increase test or to not use $p + v\Delta u$ as the next bundle center if $f$ there is relatively too large
Nonsmooth Science Fiction
Smooth-BFGS 249  VU-BFGS 87

Lewis&Overton example, 2008
\(\sqrt{x'Ax} + x'Bx; A = \text{diag}(1, 0, 1, 0, \ldots), B = \text{diag}(1, \ldots, 1/n^2), \text{dim} V = \text{dim} U = 4\)
Conclusion

Important to have BOTH $V$ and $U$ models and alternating $V$ and $U$ steps dependent on each other.


Current work: quasi-Newton, $\mu$-adjustment and extension to semismooth functions using negative curvature estimates in appropriate $V$-models.