A First Step in Designing a VU-algorithm for Nonconvex Minimization

Robert Mifflin

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

Work with Claudia Sagastizábal

ISMP 2012, Berlin

Grant support: NSF DMS 0707205, AFOSR FA9550-11-1-0139 and SOARD
Introduction

\[ \min_{x \in \mathbb{R}^n} f(x); \text{ } f \text{ locally Lipschitz} \]

with Clarke subdifferential \( \partial f \) and only one generalized gradient of \( f \)
computed by a black box at each \( x \).

Ultimate Goal: Design a VU-algorithm for such \( f \).

A VU-algorithm implicitly exploits underlying nonsmooth/smooth structure to achieve rapid convergence.

Current Goal: Define a bundle method that achieves convergence to stationary points and produces good V-models for \( f \).
Outline

- Introduction to minimization with VU-algorithms

- Bundle algorithm concepts and notation for V-models with negative curvature estimates from line search

- General definitions of null and serious points

- Three general V-model subfunction conditions for achieving stationarity results in the limit

- Single variable minimization for motivating the n-variable case and for use in a finite line search

- V-model subfunctions for nonconvex objectives on $\mathbb{R}^n$

- Future work for a complete VU-algorithm
\( \mathcal{V} \) and \( \mathcal{U} \) subspaces & graphs of \( f \) thereon

A nonconvex pdg-structured example

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

\( x^* = (0, 0, 0) \) is a stationary point (minimizer)

zero subgradient \( \in \partial f(x^*) \)

In general, for any \( \bar{x} \)

\( \bar{g} \in \partial f(\bar{x}), \quad \mathcal{V}(\bar{x}) := \text{lin}(\partial f(\bar{x}) - \bar{g}) \quad \text{and} \quad \mathcal{U}(\bar{x}) := \mathcal{V}(\bar{x})^\perp \)
A view of $f$ on $\mathcal{V}(x^*)$

$$f(u, 0, 0)$$

and for $u \in \mathcal{U}(x^*)$

'\mathcal{U}$-Lagrangian' $L_\mathcal{U}^0(u) = f(\chi(u))$

with $\chi(u)$ a primal track

[MS, Primal-dual gradient structured functions, SIOPT 13(4):1174-1194, 2003]
Lewis and Overton 8-variable half-and-half function

[MS, A VU-algorithm for convex minimization, Math. Prog. 104(2-3), 583-608, 2005]

Sublinear, linear, and superlinear convergence
Bundle algorithm concepts and notation

\( g(\cdot) \): one (generalized) gradient in \( \partial f(\cdot) \)

\( H(\cdot) \): low rank \( n \times n \) Hessian matrix

\( s(\cdot) \): safeguard scalar for convergence

last two are for down-shifting; are zero if \( f \) is convex

\( B \): bundle of data items \( (y_i, f(y_i), g(y_i), H(y_i), s(y_i)) \)

\( x \): bundle center = some \( y_j \) in \( B \)

\( \tilde{g}(x, y_i) \): estimate of a gradient at \( x \)

\( \tilde{\epsilon}(x, y_i) \geq 0 \), so that \( f(x) - \tilde{\epsilon}(x, y_i) \) underestimates \( f \) at \( x \)

both based on data at \( y_i \)

\( \phi_B(x + d) \): V-model (cutting-plane) function for \( d \in \mathbb{R}^n \)

\[
max \{ f(x) - \tilde{\epsilon}(x, y_i) + \tilde{g}(x, y_i)^T d : y_i \in B \}
\]
\( \mu \): parameter for proximal subproblem

\( P(\mu, x, B) : \min \{ \frac{1}{2} \mu |d|^2 + \phi_B(x + d) - f(x) : d \in \mathbb{R}^n \} \)
equivalent to quadratic program

\( d(x) \): subproblem solution \( d = d(x) \)

\( -\delta(x) \): subproblem minimal objective value
both also also dependent on \( \mu \) and \( B \)

serious step: gives adequate \( f \)-descent for an \( x \)-change
null step: gives adequate improvement of V-model \( \phi_B \)
line search needed to generate if \( f \) is nonconvex

If \( f \) is known to be convex then

\( (Cg) \quad \tilde{g}(x, y_i) := g(y_i), \text{ independent of } x, \)
\( (Ce) \quad e(x, y_i) = e(x, y) := f(x) - (f(y_i) + g(y_i)^T(x - y_i)) \geq 0, \)
and \( x + d(x) \) is either a null or serious point.
Bundle algorithm iteration

Loop: Solve subproblem $P(\mu, x, B)$ for $d(x)$, $\delta(x)$.

If $\delta(x) = 0$, stop with $x$ stationary.
Else ($\delta(x) > 0$ and $d(x) \neq 0$), call for a line search from $x$ along $d(x)$ with stepsize $t > 0$ so that either $t \uparrow \infty$ and $f(x + td(x)) \downarrow -\infty$
or it stops with $t$ such that the point $y_+ := x + td(x)$ is either null or serious.

Append $(y_+, f(y_+), g(y_+), H(y_+), s(y_+))$ to $B$, delete inactive $y_i$-data from $B$ and if $y_+$ is serious, then choose $\mu_+ > 0$ and replace $(x, f(x), \mu)$ by $(y_+, f(y_+), \mu_+)$. Update all relevant V-model functions $f(x) - \tilde{e}(x, y_j) + \tilde{g}(x, y_j)^T d$
and go to Loop.
Definition of $H(y(t))$ with $y(t) := x + td(x)$

Except for when the first line search $t$-value 1 does not satisfy Arimijo descent, each $t$-value has an associated $2^{nd}$ derivative estimate computed from two difference vectors

$\Delta y := (x + td(x)) - (x + t^-d(x)) = (t - t^-)d(x),$

$\Delta g := g(x + td(x)) - g(x + t^-d(x))$

where $t^-$ is a particular previous $t$-value.

For the nonexceptional cases $H(y(t))$ is a symmetric rank one update of the zero matrix using $\Delta y$ and $\Delta g$ if $\Delta g^T\Delta y < 0$.

The safeguard $s(y(t))$ is discussed below after $n = 1$ case.
Definitions of null and serious points

The null point definition is chosen to be the weakest condition known with the property that if there is an infinite number of consecutive null steps then the corresponding $\delta(x)$-sequence converges to zero.

With the general $\tilde{g}$ and $\tilde{e}$ properties assumed below the above result then implies that the last center $x$ is stationary for $f$.

The null step point definition involves parameters $m_N \in (0, 1)$ and $m_V \in [0, 1]$. 
\[ y_+ = x + td(x) \text{ with } t > 0 \]

is a null step point if
\[-\tilde{e}(x, y_+) + \tilde{g}(x, y_+)^T d(x) \geq -m_N \delta(x) - m_V \frac{1}{2} \mu |d(x)|^2; \]

is a serious step point if
\[
\frac{f(y_+) - f(x)}{t} \leq -m_A \delta(x) - m_V \frac{1}{2} \mu |d(x)|^2
\]
and
\[ t \geq 1 \text{ or } \tilde{e}(x, y_+) \geq m_S \delta(x). \]

The serious def. has a complementary Armijo-type \( f \)-descent condition with parameter \( m_A \in (0, m_N) \).

The inequality with parameter \( m_S \in (0, m_N - m_A) \) results from the need for (1) a finite line search assuming \( f \) is semismooth and (2) proving stationarity convergence results if there is an infinite sequence of serious steps, possibly with a serious \( t \)-sequence converging to zero.
Three general conditions for \( \tilde{g} \) and \( \tilde{\varepsilon} \) to give stationarity convergence results

For general \( f \) assume \( \tilde{g} \) and \( \tilde{\varepsilon} \) satisfy

(G1) \( \tilde{\varepsilon}(x, y) \geq 0 \);

(G2) if \( \{(x, y)\} \to (\bar{x}, \bar{x}) \) and \( \{\tilde{g}(x, y)\} \) is bounded
then \( \{\tilde{\varepsilon}(x, y)\} \to 0 \);

[Special case: \( \tilde{\varepsilon}(\bar{x}, \bar{x}) = 0 \)]

(G3) if \( \{(x, y)\} \to (\bar{x}, \bar{y}) \), \( \{\tilde{g}(x, y)\} \to \bar{g} \) and \( \{\tilde{\varepsilon}(x, y)\} \to 0 \)
then \( \bar{g} \in \partial f(\bar{x}) \).

[Special case: \( \tilde{g}(\bar{x}, \bar{y}) \in \partial f(\bar{x}) \) if \( \tilde{\varepsilon}(\bar{x}, \bar{y}) = 0 \)]

Lemma: If \( f \) is convex with \( \tilde{g} \) and \( \tilde{\varepsilon} \), respectively, defined by (Cg) and (Ce) then conditions (G1), (G2) and (G3) hold.
Let \{y_\ell\} be the bundle algorithm generated sequence of \(y_+\) points and \{x_k\} be the subsequence of these points that are also serious points and, hence, bundle centers.

**Theorem (Stationarity implied by (G1), (G2) and (G3))**:

(i) Suppose \(x_k\) is the last serious point generated and \(\{\tilde{g}(x_k, y_\ell)\}\) is bounded. Then \(x_k\) is stationary.

(ii) Suppose \{\(x_k\)\} is infinite, \{\(y_\ell\)\} and \(\{\tilde{g}(x_k, y_i) : y_i \in B_{\ell_k}\}\) are bounded and \(0 < \mu_{min} \leq \mu_k \leq \mu_{max} < \infty\) where \((\mu_k, B_{\ell_k-1})\) generates \(x_k\). Then every accumulation point of \{\(x_k\)\} is stationary.

What should \(\tilde{g}\) and \(\tilde{e}\) be if \(f\) is nonconvex?

(Cg+) \(\tilde{g} = g + \) nonconvexity correction?

(Ce+) \(\tilde{e} = e + \) nonconvexity correction + safeguard?
Geometry of single variable $VU$-minimization

$(n = 1)$

Convex $f$

$L \ x \ y$

The next iterate is the minimizer of the $V$-model
(closer to $x$ than the $U$-model minimizer)
The $n = 1$ algorithm generates intervals, $[x, y]$ or $[y, x]$, such that $f(x) \leq f(y)$ and $g(x)^T(y - x) \leq 0$.

It updates $x$- and $y$-side $2^{nd}$ derivative estimates and two associated quadratic $f$-approximates using endpoints of a previous interval $[L, R]$ strictly containing $x$ and $y$.

A $U$-model is defined by the $x$-side quadratic function $q_x$ and the $x$-side of the $V$-model is always defined by the linearization depending on $f(x)$ and $g(x)$.

The $y$-side of the $V$-model is defined as illustrated next if the $y$-side quadratic function $q_y$ is strictly concave; else, by the linearization depending on $f(y)$ and $g(y)$ as in previous figure.
superlinearly convergent for certain piecewise $C^2$ functions when safeguarded properly
Initially define a safeguard scalar $s := \frac{1}{2}/|R - L|$ where $[L, R]$ is the first interval of uncertainty.

If $q_y(x) > f(x)$, i.e. no down-shift, reset $s$ to $\frac{1}{2}/|y - x|$.

Choose $d$ so that $x + d$ is the $V$- or $U$-model minimizer that is closer to $x$. Set the next iterate to be the projection of $x + d$ into the safeguard interval $[\min(x, y) + s|y - x|^2, \max(x, y) - s|y - x|^2]$.

In order to not prevent the possibility of superlinear convergence, $s$ is not reset as long as $q_y(x) \leq f(x)$, since a reset causes a bisection step.
Algebraic definitions of \( \tilde{g} \) and \( \tilde{e} \) for \( f \) on \( \mathbb{R}^n \)

Given a bundle center \( x \) and a point \( y \) with associated data \( g(y), H(y) \) and \( s(y) \) let

\[
h(x, y) := (x - y)^T H(y)(x - y),
\]

\[
\tilde{g}(x, y) := g(y) \quad \text{if } h(x, y) \geq 0
\]
\[
= g(y) + H(y)(x - y) \quad \text{otherwise},
\]

\[
\tilde{e}(x, y) := \max\{e(x, y) - \frac{1}{2} \min(0, h(x, y)), s(y)|x - y|^2\}
\]

where

\[
e(x, y) := f(x) - (f(y) + g(y)^T(x - y))
\]

and

\[
s(y) \geq s_{\min} \geq 0
\]

with safeguard

\[
s_{\min} > 0 \text{ if } f \text{ is not convex.}
\]
Lemma: If $\tilde{g}(\cdot, \cdot)$ and $\tilde{e}(\cdot, \cdot)$ are so defined then
(G1) holds and
(G2) holds if $\{s(y)\}$ is bounded from above as $\{y\} \to \bar{x}$ and
(G3) holds if $\{H(y)\}$ is bounded as $\{y\} \to \bar{y}$.

This Lemma and the definition of $\tilde{g}(\cdot, \cdot)$ then imply that
the stationarity results of the Theorem hold
if $0 < \mu_{\min} \leq \mu_k \leq \mu_{\max} < \infty$ and $\{y_\ell\}$, $\{s(y_\ell)\}$ and $\{H(y_\ell)\}$
are all bounded.
Future research

(i) For the exceptional case when \( y(1) = x + d(x) \) does not satisfy an Armijo descent test, determine conditions for when \( H(y(1)) \) can be an SR1 update of \( H(y_j) \) for some \( y_j \) active in the bundle that generated \( d(x) \). This would include the angle between \( d(x) \) and \( y_j - x \) being small.

(ii) Determine choices for \( s(y) \) in the \( n \)-variable case; dependence on \( x \) too?

(iii) Using the above bundle algorithm develop a VU-algorithm for lower-\( C^2 \) functions[Rockafellar, 1982]. This involves choosing values for \( m_V \) to generate very good V-models.
Recall, $y_+ = x + td(x)$ with $t > 0$

is a null step point if

$$-\tilde{\epsilon}(x, y_+) + \tilde{g}(x, y_+)^T d(x) \geq -m_N \delta(x) - m_V \frac{1}{2} \mu |d(x)|^2;$$

is a serious step point if

$$\frac{[f(y_+) - f(x)]}{t} \leq -m_A \delta(x) - m_V \frac{1}{2} \mu |d(x)|^2$$

and

$$t \geq 1 \quad \text{or} \quad \tilde{\epsilon}(x, y_+) \geq m_S \delta(x).$$
Line Search with variable \( t \)

The search generates a sequence of nested intervals \([t_L, t_R]\) where \( x + td(x) \) with \( t = t_L(t_R) \) does (does not) satisfy the serious point Armijo \( f \)-descent condition.

Start with \( t = 1 \) in the initial interval \([t_L, t_R) = [0, \infty)\).

If \( t = 1 =: t_R \) then enter the interpolation

Loop: If \( x + td(x) \) is a serious or null point, exit.

If \([t_L, t_R]\) is a VU-model compatible interval compute the next value of \( t \) as in the single variable algorithm.
Else replace \( t \) by the bisector of \([t_L, t_R]\).
Replace the appropriate endpoint of \([t_L, t_R]\) by \( t \) and go to Loop.

Else \( (t = 1 =: t_L) \),
sequentially increase $t$ until there is an exit with $t =: t_L$ and $g(x + td(x))^T d(x)$ satisfying a Wolfe test, or $t =: t_L$ and $t$ being too large, or $t =: t_R$.

In the last case an interpolation phase as above could be entered to find a serious point, possibly better than the one given by the current $t_L$ value.

*Lemma*: If $f$ is semismooth then the above line search is finite.