Convex Models in Bundle Methods for Nonsmooth Nonconvex Minimization: Prerequisite for a VU-algorithm

Robert Mifflin

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

Work with Claudia Sagastizábal

UBCO 2014, Kelowna

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

\[ \min_{x \in \mathbb{R}^n} f(x); \quad f \text{ locally Lipschitz,} \]
\[ \text{only one (Clarke) gradient } g(x), \]
\[ \text{computed by a black box at each } x. \]

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

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

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 - convex case
Lewis and Overton 8-variable half-and-half function

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

Sublinear, linear, and superlinear convergence -
What can be done for nonconvex case?
**Introduction**

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

only one (Clarke) gradient \( g(x) \),
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. To do so, a suitable V-model for \( f \) is needed.

**This talk:** Define a bundle method that achieves convergence to stationary points and produces good V-models for \( f \). We refer to this as a viable algorithm with a viable V-model (i.e. polyhedral model). Also consider more general convex models.
Outline

- Difficulty on V-space due to nonconvexity

- Four general conditions for a viable bundle algorithm

- Framework with model functions $M$, centers $x$, line searches generating null or serious (next center) points

- General stationarity theorem

- Specific V-model bundle algorithm with safeguarded negative curvature corrections generated by line searches

- Specific null and serious point definitions for obtaining finite line searches for semismooth objectives and viability to imply asymptotic stationarity
- Future work for a complete VU-algorithm
\( \mathcal{V} \) and \( \mathcal{U} \) subspaces and graph of \( f \) on \( \mathcal{V} \)

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 $\mathcal{V}(x^*)$
Bundle iteration elements

For a prox-parameter $\mu > 0$ and a convex model function $M(\approx f$ near a center $x)$ define

search direction $d(x) := \arg \min M(x + \cdot) + \frac{1}{2}\mu |\cdot|^2,$

aggregate gradient $G(x) := -\mu d(x) \in \partial M(x + d(x)),$

aggregate error $E(x) := M(x) - M(x + d(x)) - \mu |d(x)|^2 \geq 0$ (nonnegative by subgradient inequality for convex $M$),

progress measure $D(x) := f(x) - M(x + d(x)) - \frac{\mu}{2} |d(x)|^2,$

$D(x) = f(x) - M(x) + E(x) + \frac{1}{2\mu} |G(x)|^2,$

so $D(x) \geq E(x) + \frac{1}{2\mu} |G(x)|^2 \geq 0$

if $f(x) \geq M(x)$
Bundle iteration elements

For a prox-parameter $\mu > 0$ and a model function $M(\approx f \text{ near a center } x)$ define

search direction $d(x) := \arg \min M(x + \cdot) + \frac{1}{2} \mu |\cdot|^2$,

aggregate gradient $G(x) := -\mu d(x) \in \partial M(x + d(x))$,

aggregate error $E(x) := M(x) - M(x + d(x)) - \mu |d(x)|^2 \geq 0$ (nonnegative by subgradient inequality for convex $M$),

progress measure $D(x) \geq E(x) + \frac{1}{2\mu} |G(x)|^2 \geq 0$ if $f(x) \geq M(x)$

What is a viable model?
Viable Models

A model function $M$ is viable if it is convex and satisfies

$\textbf{V1}$ The model is lower at its center $x$: $M(x) \leq f(x)$

$\textbf{V2}$ Finite version

Zero aggregate error implies aggregate gradient is an $f$-subgradient:

$E(x) = 0 \implies G(x) \in \partial f(x)$
Viable Models

A model function $M$ is viable if it is convex and satisfies conditions

**V1** The model is lower at its center $x$: $M(x) \leq f(x)$

$V1$ implies $D(x) \geq E(x) + \frac{1}{2\mu}|G(x)|^2 \geq 0$

**V2** Finite version

Zero aggregate error implies
aggregate gradient is an $f$-subgradient:

$E(x) = 0 \implies G(x) \in \partial f(x)$

$V1+V2$ and $D(x) = 0$ imply stationarity of $x$
A model function $M$ is viable if it is convex and satisfies conditions

**V1** The model is lower at a center $x$: $M(x) \leq f(x)$

$V1$ implies $D(x) \geq E(x) + \frac{1}{2\mu}|G(x)|^2 \geq 0$

**V2** Finite version

Zero aggregate error implies aggregate gradient is an $f$-subgradient:

$E(x) = 0 \implies G(x) \in \partial f(x)$

$V1+V2$ and $D(x) = 0$ imply stationarity of $x$

For $f$ convex, a cutting-plane model ensures both conditions hold, even an asymptotic version of $V2$. 
Viable Algorithms

A bundle algorithm is viable if its model $M$ is viable and it satisfies asymptotic conditions $V2$ and $V3$ and line search condition $V4$, depending on $D$-decreasing null and $f$-decreasing serious point definitions:

**V2 Algorithmic (asymptotic) version**

Zero asymptotic aggregate error implies associated asymptotic aggregate gradient is an $f$-subgradient:

$$x_k \to \bar{x} \text{ and } E(x_k) \to 0 \implies G(x_k) \to \varepsilon \partial f(\bar{x})$$

$V1+V2$ and associated $D(x_k) \to 0$ imply stationarity of $\bar{x}$
**Viable Algorithms**

A bundle algorithm is viable if its model $M$ is viable and it satisfies asymptotic conditions $V2$ and $V3$ and line search condition $V4$, depending on $D$-decreasing null and $f$-decreasing serious point definitions:

**V2 Algorithmic (asymptotic) version**

Zero asymptotic aggregate error implies associated asymptotic aggregate gradient is an $f$-subgradient:

$$x_k \to \bar{x} \text{ and } E(x_k) \to 0 \implies G(x_k) \to \partial f(\bar{x})$$

**V3** Zero asymptotic progress measure:

$$x_k \to \bar{x} \implies D(x_k) \to 0$$

**V4** viable line search at each iteration: defined next to find a null or serious point


**Bundle algorithm with viable line search**

Input null/serious point defs. and $x_0, \mu_0 > 0, M_0$;
initialize $\ell := 0$, $k := 0$, $x := x_0$

Loop: Solve subproblem with $x, \mu_\ell, M_\ell$ for $d_\ell(x)$, $D_\ell(x)$.
    If $D_\ell(x) = 0$, stop with $x$ stationary.
    Else, call for a line search from $x$ along $d_\ell(x)$
    with stepsize $t > 0$ so that ...
Bundle Algorithm with **viable** line search

either $t \uparrow \infty$ and $f(x + td_\ell(x)) \downarrow -\infty$
or it stops with $t = t_\ell$ such that the point

$$y_{\ell+1} := x + t_\ell d_\ell(x)$$
is either null or serious.

If $y_{\ell+1}$ is serious, set $x_{k+1} := y_{\ell+1}$, $\ell(k) := \ell$ and replace $x$ by $x_{k+1}$ and $k$ by $k + 1$.

Choose $\mu_{\ell+1} > 0, M_{\ell+1}$ based on bundled $M_\ell$-data, $y_{\ell+1}$ and other data generated at iteration $\ell$.
Replace $\ell$ by $\ell + 1$ and go to Loop.
**Theorem (Stationarity).**

Suppose

- the bundle algorithm does not terminate,
- the prox-parameters are in a positive interval 
  \([\mu_{\text{min}}, \mu_{\text{max}}]\), and
- the progress measure sequence \(\{D_\ell(x_k)\}\) is bounded.

If conditions **V1** to **V4** hold then any \(\bar{x}\) that is a limit point of \(\{x_k\}\) is stationary for \(f\).

If \(x_k\) is finite with \(\bar{x}\) being the last \(x_k\) then **V3** is written \(D_\ell(\bar{x}) \rightarrow 0\) instead of \(D_{\ell(k)}(x_k) \rightarrow 0\).

Now, for nonconvex \(f\), we define a specific algorithm with viable polyhedral model, line search and null/serious definitions.
Specific viable model for nonconvex $f$

Polyhedral (V-model) function

$$M(x + d) = \max\{f(x) - \tilde{e}(x, y_i) + \langle \tilde{g}(x, y_i), d \rangle : y_i \in B\}.$$ 

Gradient estimates $\tilde{g}(x, y_i)$ and linearization error estimates $\tilde{e}(x, y_i)$ depend on the center $x$, a bundle $B$ of previous iterates $y_i$ and associated data.
Specific viable model for nonconvex $f$

Polyhedral (V-model) function

$$M(x + d) = \max \{ f(x) - \tilde{e}(x, y_i) + \langle \tilde{g}(x, y_i), d \rangle : y_i \in B \}.$$  

Gradient estimates $\tilde{g}(x, y_i)$ and linearization error estimates $\tilde{e}(x, y_i)$ depend on the center $x$, a bundle $B$ of previous iterates $y_i$ and associated data

\textbf{V1} ensured by forcing $\tilde{e}(x, y_i) \geq 0$ via sufficient curvature terms or safeguards

Also beneficial to keep center $x$ in $B$ and to have

$\tilde{e}(x, x) = 0$, $\tilde{g}(x, x) = g(x)$
Specific viable model for nonconvex $f$

Polyhedral (V-model) function

$$ M(x + d) = \max\{f(x) - \tilde{e}(x, y_i) + \langle \tilde{g}(x, y_i), d \rangle : y_i \in B\}. $$

Gradient estimates $\tilde{g}(x, y_i)$ and linearization error estimates $\tilde{e}(x, y_i)$ depend on the center $x$, a bundle $B$ of previous iterates $y_i$ and associated data

$$(y_i, f(y_i), g(y_i), H(y_i), s(y_i))$$

with low rank Hessian matrix and safeguard scalar for down-shifting (both zero if $f$ is convex)

V1 ensured by forcing $\tilde{e}(x, y_i) \geq 0$ via sufficient curvature terms or safeguards

Also beneficial to keep center $x$ in $B$ and to have the center linearization $f(x) + \langle g(x), d \rangle$ in the model-max
Convex \( f \)

\[
\tilde{g}(x, y) := g(y), \text{ independent of } x \neq y,
\]
\[
\tilde{e}(x, y) := e(x, y) := f(x) - (f(y) + \langle g(y), (x - y) \rangle);
\]
with \( y = y_i \) this gives lower cutting planes:  \( e(x, y_i) \geq 0 \)
and further convex analysis gives \( V_2 \)

Nonconvex \( f \)

Keep \( M \) polyhedral and simply modify \( \tilde{e}, \tilde{g} \) via negative curvature corrections from \( H \), computed during line search. When corrections are not large enough there is a safeguard using \( s \) to make \( \tilde{e} \) large enough to obtain \( V_2 \), depending on outer semicontinuity of \( \partial f(\cdot) \).
Convex $f$

\[ \tilde{g}(x, y) := g(y), \text{ independent of } x \neq y, \]
\[ \tilde{e}(x, y) := e(x, y) := f(x) - (f(y) + \langle g(y), (x - y) \rangle); \]
with $y = y_i$ this gives lower cutting planes: $e(x, y_i) \geq 0$
and further convex analysis gives $V2$

Nonconvex $f$

Keep $M$ polyhedral and simply modify $\tilde{e}, \tilde{g}$ via negative curvature corrections from $H$, computed during line search. When corrections are not large enough there is a safeguard using $s$ to make $\tilde{e}$ large enough to obtain $V2$, depending on outer semicontinuity of $\partial f(\cdot)$.

Solving the proximal subproblem with polyhedral $M$ gives $G(x) [\mathcal{E}(x)]$ as a convex combination of $\tilde{g}(x, y_i) [\tilde{e}(x, y_i)]$. 
Geometry of single variable $VU$-minimization
($n = 1$)

(convex $f$)

The next iterate is the minimizer of the $V$-model
(closer to $x$ than the $U$-model minimizer)
superlinearly convergent for certain piecewise $C^2$ functions when safeguarded properly
An interval $[x, y]$ or $[y, x]$ is called compatible if $f(x) \leq f(y)$ and $\langle g(x), (y - x) \rangle \leq 0$.

The $n = 1$ algorithm generates such intervals.

However, the viable line search of the $n$-variable algorithm determines the endpoints of its $t$-interval of uncertainty based on satisfaction (or not) of an Armijo $f$-descent test dictated by the serious point definition given below.

For a compatible $t$-interval the line search defines its next iterate as in the above algorithm; otherwise the next iterate is the bisector of the $t$-interval.
The $n = 1$ algorithm updates two $2^{nd}$ derivative estimates and associated *quadratic* $f$-approximates using previous interval endpoints.

The $n$-variable algorithm proceeds similarly, with respect to $V$-models, using matrices $H(x + td(x))$, updated by an SR1 formula during line search on $t$, to employ if negative curvature is found.
Specific Viable Model Definition

Given a center $x$ and a bundle point $y$ with associated data $g(y)$, $H(y)$ and $s(y)$ compute the curvature

$$h(x, y) := \langle x - y, H(y)(x - y) \rangle$$

Nonnegative curvature $h$:

$$\tilde{g}(x, y) := g(y) \quad \tilde{e}(x, y) := \max(e(x, y), s(y)|x - y|^2)$$

Negative curvature $h$:

$$\tilde{g}(x, y) := g(y) + H(y)(x - y) \quad \tilde{e}(x, y) := \max(e(x, y) - \frac{1}{2} h, s(y)|x - y|^2)$$
Specific Viable Model Definition

Given a center $x$ and a bundle point $y$ with associated data $g(y)$, $H(y)$ and $s(y)$ compute the curvature

$$h(x, y) := \langle x - y, H(y)(x - y) \rangle$$

Nonnegative curvature $h$:

$$\tilde{g}(x, y) := g(y) \quad \tilde{e}(x, y) := \max(e(x, y), s(y)|x - y|^2)$$

Negative curvature $h$:

$$\tilde{g}(x, y) := g(y) + H(y)(x - y) \quad \tilde{e}(x, y) := \max(e(x, y) - \frac{1}{2}h, s(y)|x - y|^2)$$

$s(y) \in [s_{\text{min}}, s_{\text{max}}]$ with safeguard $s_{\text{min}} > 0$ if $f$ is not convex, to obtain $V2$.

These definitions immediately imply $V1$. 
Specific viable null/serious point definitions

The null point definition is the weakest one known such that infinite number of consecutive null steps with $\mu$ nondecreasing and $x_k = \bar{x}$ fixed make the $D_\ell(\bar{x})$-sequence converge to zero (V3 null version).

Four parameters for linear combinations of $D(x)$ and $\mu |d(x)|^2 = |G(x)|^2 / \mu = - \langle G(x), d(x) \rangle$, with bounds for obtaining V3, V4.

- null point $D$-decrease: $m_N \in (0, 1)$
- serious point Armijo-type $f$-decrease: $m_A \in (0, m_N)$
- serious point small stepsize $D$-decrease: $m_S \in (0, m_N - m_A)$
- both points V-model improvement: $m_V \in [0, 1]$ (V5)
- a fifth parameter is possible for even more serious flexibility
Specific viable null/serious point definitions

\( y_+ = x + td(x) \) with \( t > 0 \)

is a null step point if

\[-\tilde{e}(x, y_+) + \langle \tilde{g}(x, y_+), d(x) \rangle \geq -m_N D(x) - m_V \frac{1}{2} \mu |d(x)|^2; \]

is a serious step point if

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

and

\[
t \geq 1 \quad \text{or} \quad \tilde{e}(x, y_+) \geq m_S D(x).
\]

Lemma. If \( f \) is convex and \( s_{\text{max}} = 0 \) then \( \tilde{g} = g \), \( \tilde{e} = e \) and \( t = 1 \) gives either a null or a serious point.

For \( t < 1 \) the inequality with parameter \( m_S \) ensures \( V_3 \) for a serious point sequence when its corresponding \( t \)-sequence converges to zero.
**Convergence of Specific Viable Algorithm**

**Theorem (Stationarity).**

Suppose $f$ is semismooth and

- the bundle algorithm does not terminate,
- the prox-parameters are in a positive interval $[\mu_{\text{min}}, \mu_{\text{max}}]$, and
- the sequences of centers $\{x_k\}$ and matrices $\{H(y_\ell)\}$ are bounded.

Then any $\bar{x}$ that is a limit point of $\{x_k\}$ is stationary for $f$.

Proof shows satisfaction of $\textbf{V1}$ to $\textbf{V4}$ with the asymptotic conditions based on boundedness of $\{y_\ell\}$, which follows from $\{x_k\}$ bounded and $\textbf{V4}$ depending on semismoothness of $f$. \qed
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 $y_j - y(1)$ being small.

(ii) Determine choices for $s(y)$ in the $n$-variable case; dependence on $x$ too as in the 1-variable case?

(iii) Using the above bundle algorithm develop a VU-algorithm for lower-$C^2$ functions [Janin, 1974], [Rockafellar, 1982]. This involves choosing values for $m_V \leq 1$ to generate very good V-models.
Recall, \( y_+ = x + td(x) \) with \( t > 0 \)

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

is a serious step point if
\[
\frac{[f(y_+) - f(x)]}{t} \leq -m_A D(x) - m_V \frac{1}{2} \mu |d(x)|^2
\]
and
\[
t \geq 1 \quad \text{or} \quad \bar{e}(x, y_+) \geq m_S D(x).
\]
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 A-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 $\langle g(x + td(x)), d(x) \rangle$ satisfying a Wolfe test, or $t =: t_L$ and $t$ being too large (i.e. $f(x + td(x))$ too small), 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 interpolation entering $t_L$ value.

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