Computing a subgradient by forward/tangent AD for functions of two variables

Kamil A. Khan, Yingwei Yuan
kamilkhan@mcmaster.ca

Department of Chemical Engineering, McMaster University

23rd EuroAD Workshop

August 12, 2020
Outline

Motivation: nonsmoothness in applications

Capturing local sensitivity information
  Directional derivatives
  Convex subgradients
  Clarke subdifferentials

New approach for bivariante functions
  Constructing compass differences
  Compass differences are subgradients

Examples
Nonsmoothness in applications

▶ Qualitative changes in behavior:
  ▶ Transitions in thermodynamic phase
  ▶ Safety mechanisms
  ▶ Multi-stage processes: pressure-swing, semi-batch, startup/shutdown

▶ Penalty methods for constrained optimization; regularization methods for PDEs and ML

▶ Pinch analysis for heat integration
  [Duran & Grossmann (1986), Watson, Khan, & Barton (2015), Vikse et al. (2018)]

▶ Convex relaxations and envelopes for global optimization
  [McCormick (1976), Khajavirad & Sahinidis (2012), Tsoukalas & Mitsos (2015)]
Directional derivatives

Consider a scalar-valued function $f : \mathbb{R}^n \to \mathbb{R}$.

One-sided directional derivative

- If $f$ is not differentiable at $x$, it might still have a directional derivative $f'(x; d)$ in any direction $d$:

$$f'(x; d) = \lim_{t \to 0^+} \frac{f(x + td) - f(x)}{t}.$$

- Computed e.g. by forward/tangent AD [Griewank (1994)]
- Virtually all continuous models are directionally differentiable
- If $f$ is differentiable at $x$, then $f'(x; d) = \nabla f(x)^T d$
Subgradients and subdifferentials

- Suppose \( f : \mathbb{R}^n \to \mathbb{R} \) is convex
- \( f \) has at least one subgradient \( s \) at each point \( x \):

\[
f(y) \geq f(x) + s^T(y - x) \quad \text{for all } y
\]

- Useful in nonsmooth convex optimization
  [e.g. Hiriart-Urruty & Lemaréchal (1996), Nesterov (2018)]

- Useful in global optimization
  [Khan (2018), Cao, Song, & Khan (2019)]

- The subdifferential \( \partial f(x) \) collects all subgradients of \( f \) at \( x \)
- If \( f \) is differentiable at \( x \), then \( \partial f(x) = \{\nabla f(x)\} \)
Nonconvex subdifferentials

Clarke (1973)

- Suppose $f : \mathbb{R}^n \rightarrow \mathbb{R}$ is locally Lipschitz continuous

Clarke subdifferential

- $\partial f(x) := \text{convex hull of gradients of } f \text{ close to } x$
- Coincides with subdifferential when $f$ is convex
- Useful in nonsmooth nonconvex optimization
  [e.g. Facchinei and Pang (2008)]
- Does not obey a sharp chain rule:

$$f(x) \equiv \max\{0, x\}, \quad g(x) \equiv \min\{0, x\}, \quad h(x) \equiv x = f(x) + g(x)$$

- $0 \in \partial f(0) = [0, 1]$, and $0 \in \partial g(0) = [0, 1]$, but $0 \notin \partial h(0) = \{1\}$
For many nonsmooth systems, ... 

... directional derivatives are easy to evaluate

- Nonsmooth forward/tangent AD
  - [Griewank (1994)]
- Nonsmooth dynamical systems
  - [Pang & Stewart (2009)]
- Optimal-value functions for NLPs
  - [Danskin (1966), Hogan (1973)]

... but subgradients and Clarke subgradients are tougher

- Nonsmooth composite functions
  - Nonsmooth vector forward AD [Khan & Barton (AD 2012, 2015)]
  - Branch-locking reverse AD [Khan (AD 2016)]
- Nonsmooth dynamical systems
  - LD-derivatives [Khan & Barton (2014), Stechlinski & Barton (2016)]
- Optimal-value functions for NLPs
  - [Tsoukalas & Mitsos (2015), Stechlinski, Khan, & Barton (2018)]
Idea: can we compute subgradients from directional derivatives?

- For univariate $f$, directional derivatives correspond to subgradients
- For differentiable $f$, we can use vector forward AD:
  \[ \nabla f(x) = (f'(x; e_1), f'(x, e_2), \ldots, f'(x; e_n)) \]
- For nonsmooth $f$, this is not necessarily a subgradient, e.g. $f(x, y) \equiv \max(x, y)$ at $(0, 0)$:

- For convex $f$, there’s a duality relationship: [Rockafellar (1970)]
  \[ \partial f(x) = \{ s : f'(x; d) \geq s^T d \quad \text{for all } d \} \]
- For abs-factorable functions $f : \mathbb{R}^n \to \mathbb{R}$ with $p$ absolute-values, we can use $np$ directional derivatives by “cone-squashing” [Khan & Barton (2012)]
Compass differences

- Suppose $f : \mathbb{R}^n \rightarrow \mathbb{R}$ is directionally differentiable
- Assemble directional derivatives in all $\pm$coordinate directions:

$$
\Delta ^\mp f(x) := \frac{1}{2} \begin{bmatrix}
  f'(x; e_{(1)}) - f'(x; -e_{(1)}) \\
  f'(x; e_{(2)}) - f'(x; -e_{(2)}) \\
  \vdots \\
  f'(x; e_{(n)}) - f'(x; -e_{(n)})
\end{bmatrix}
$$

- Centered finite difference of $f'(x; d)$ w.r.t. $d$ at $0$: 
In $\mathbb{R}^2$, compass differences are subgradients

Khan and Yuan (2020)

- Suppose $f : \mathbb{R}^2 \rightarrow \mathbb{R}$ is locally Lipschitz continuous and directionally differentiable at $x$.
  - Only two dimensions?
    - Yes.

- New result: if $f$ is convex, then $\Delta^{\oplus} f(x) \in \partial f(x)$.
  - Intuition: in $\mathbb{R}^2$, the midpoint of the interval hull of a compact convex set $C \subset \mathbb{R}^2$ is itself in $C$:

- New result: if $f$ is not convex, then $\Delta^{\oplus} f(x) \in \partial f(x)$ as well.
Why two dimensions?

- Intuition: in $n \geq 3$ dimensions, the center of the interval hull of a compact convex set $C \subset \mathbb{R}^n$ is not necessarily in $C$:

![Image of a hexahedron with a point inside]

- The following convex functions have these subdifferentials at $0$:

$$f(x) \equiv \max(x_1 - x_2 - x_3, \; x_2 - x_3 - x_1, \; x_3 - x_1 - x_2, \; -x_1 - x_2 - x_3)$$

$$\phi(x) \equiv \max(x_1 + x_2 - x_3, \; x_2 + x_3 - x_1, \; x_3 + x_1 - x_2, \; x_1 + x_2 + x_3)$$

- Here $\Delta^\oplus f(0) = \Delta^\oplus \psi(0)$, but $\partial f(0)$ and $\partial \phi(0)$ are disjoint.
Extensions

Finite difference approximations

- From its definition,
  \[
  \Delta^\oplus f(x_1, x_2) = \lim_{\epsilon \to 0^+} \frac{1}{2\epsilon} \left[ f(x_1 + \epsilon, x_2) - f(x_1 - \epsilon, x_2) \right] - \left[ f(x_1, x_2 + \epsilon) - f(x_1, x_2 - \epsilon) \right].
  \]
- So, for such \( f \),
  centered finite differences \( \approx \) subgradients
- Works for nonsmooth black-box functions \( f : \mathbb{R}^2 \to \mathbb{R} \)

Changing basis

- If \( u, v \) are linearly independent, and \( M := \begin{bmatrix} u & v \end{bmatrix} \), then
  \[
  \frac{1}{2} (M^T)^{-1} \begin{bmatrix} f'(x; u) - f'(x; -u) \\ f'(x; v) - f'(x; -v) \end{bmatrix} \in \partial f(x).
  \]
Extensions

[redacted summary of in-progress work]
Example: dynamic systems

Consider a parametric ODE in $x$:

$$
\dot{x}_1 = |x_1| + |x_2| \\
\dot{x}_2 = |x_2|,
$$

$x(0, p) = p$,

with the cost function $\phi(p) \equiv x_1(1, p)$.

Using Pang & Stewart’s directional derivatives (2013), we can compute $\Delta^\oplus \phi(0)$ in MATLAB:

This approach works for any number of state variables $x_i$. 
Example: optimal-value functions

- Danskin (1966): if
  - \( f : \mathbb{R}^n \times \mathbb{R}^m \rightarrow \mathbb{R} \) is smooth,
  - \( \phi(p) \equiv \min_x f(x, p) \), and
  - \( Y(p) := \operatorname{argmin}_x f(x, p) \),
then \( \phi'(p; d) = \min_{x \in Y(p)} \frac{\partial f}{\partial p}(x, p) d \).

- (Other variants too.)

- When \( m = 2 \) and sufficient conditions aren’t satisfied, compass differences are the only known way to exactly describe a subgradient of \( \phi \).

- Extends to Tsoukalas-Mitsos relaxations (2015) of bivariate intrinsic functions [c.f. Yingkai’s earlier talk]
  - e.g. relaxation of \( g(z) \equiv (z_1^2 - z_2^2 + 1)((z_1 - 1)^6 + z_2 + 1) \):
Conclusions

▷ Compass difference for bivariate nonsmooth functions:

\[ \Delta \oplus f(x) := \frac{1}{2} \left[ f'(x; e^{(1)}) - f'(x; -e^{(1)}) \right] - \frac{1}{2} \left[ f'(x; e^{(2)}) - f'(x; -e^{(2)}) \right] \]

▷ Often easy to evaluate by e.g. forward/tangent AD

▷ For such functions (even when nonconvex):
  ▷ compass differences are subgradients
  ▷ centered finite differences converge to a subgradient
  ▷ [Khan and Yuan (2020)]

▷ Especially useful for:
  ▷ black-box models
  ▷ dynamic models
  ▷ models with embedded optimization

▷ [redacted summary of in-progress work]

▷ Funding: NSERC, MACC