Bull. Austral. Math. Soc. 49J52, 90C48, 26E20
Vol. 49 (1994) [129–137]

A CHAIN RULE FOR NONSMOOTH COMPOSITE FUNCTIONS VIA MINIMISATION

D. Ralph

Nonsmooth calculus using the approximate subdifferential of Mordukhovich and Ioffe admits a sharper chain rule, hence sharper applications in optimisation, than does the generalised gradient of Clarke. We observe, however, that at a local minimum point of the composition of nonsmooth vector valued and real valued functions, the generalised gradient admits a special, relatively sharp chain rule, that yields sharper results than have been seen before in the context of the generalised gradient.

1. Introduction

Let X and Y be Banach spaces with continuous dual spaces X' and Y' respectively, and $\mathcal{L}(X,Y)$ be the space of bounded (continuous) linear mappings from X to Y. Let $f : Y \to \mathbb{R}$, $g : X \to Y$, and suppose that g is Lipschitz near a point $\overline{x} \in X$ and f is Lipschitz near $\overline{z} \overset{\text{def}}{=} g(\overline{x})$.

There are various notions of derivatives generalising the classical definition which is used for smooth functions. Perhaps the best known is the *generalised gradient*, Clarke [3], of a locally Lipschitz function, denoted by ∂. We mention that the generalised directional derivative of f at $y \in Y$ in the direction $v \in Y$,

$$f^\circ(y;v) \overset{\text{def}}{=} \limsup_{\substack{y' \to y \\ t \downarrow 0}} [f(y'+tv) - f(y')]/t,$$

is used to define the generalised gradient:

$$\partial f(y) \overset{\text{def}}{=} \{\lambda \in Y' \mid \lambda v \leqslant f^\circ(y;v),\ \forall v \in Y\}.$$

An alternative is the *approximate subdifferential* of Mordukhovich [13] and Ioffe [9], ∂_a, whose more complex definition is postponed to the Appendix, for brevity here. We note $\partial f(y)$ is always convex, though $\partial_a f(y)$ may not be, and that the former contains the latter for y near $\overline{z}$, sometimes strictly. For example, the function $f : \mathbb{R} \to \mathbb{R} : y \mapsto -|y|$ is such that $\partial_a f(0) = \{-1, 1\}$ whereas $\partial f(0) = [-1, 1]$. Nevertheless Ioffe [9,

Received 11th March, 1993
I thank Barney Glover for his generous contribution of ideas in discussions on this paper.

https://doi.org/10.1017/S0004972700016178 Published online by Cambridge University Press

Proposition 3.3] shows that $\operatorname{cl}\operatorname{co}\partial_a f(y) = \partial f(y)$ for y near $\bar{x}$, where cl co denotes the closed convex hull operation. An extension of classical calculus is that if $\bar{x}$ is a local minimiser of f then $0 \in \partial_a f(\bar{x}) \cap \partial f(\bar{x})$. Also, if f is continuously differentiable with gradient ∇f, then $\partial_a f(y) = \partial f(y) = \{\nabla f(y)\}$.

The approximate subdifferential has proven to be a slightly sharper tool, as evidenced not only by the inclusion $\partial_a f(y) \subset \partial f(y)$, but by the need for the closed convex hull operation in the second of the following two chain rules.

Theorem 1. *Suppose g is compactly Lipschitzian (see Appendix) at $\bar{x}$. Then*

$$\partial_a(f \circ g)(\bar{x}) \subset \bigcup_{\lambda \in \partial_a f(\bar{x})} \partial_a(\lambda g)(\bar{x}), \tag{1}$$

and

$$\partial(f \circ g)(\bar{x}) \subset \operatorname{cl}\operatorname{co} \bigcup_{\lambda \in \partial f(\bar{x})} \partial(\lambda g)(\bar{x}). \tag{2}$$

Proof: If Y is finite dimensional, the compactly Lipschitz hypothesis on g is superfluous because g is Lipschitz near $\bar{x}$, and the chain rule (1) is due to Mordukhovich [**13**]. More generally, the first chain rule is due to Jourani and Thibault [**12**]. It is closely related to of a result of Ioffe [**10**, Corollary 7.8.1], which assumes that $f \circ g$ has a strict prederivative at $\bar{x}$ with compact values. The second chain rule is due to Glover [**5**]. ◻

Both chain rules are fundamental in proving a number of closely related stability (surjectivity) results and optimality conditions for nonsmooth optimisation. Results using the generalised gradient can often be deduced as trivial corollaries of corresponding results using the approximate subdifferential, because the generalised gradient contains the approximate subdifferential. Some results using the generalised gradient, however, are actually sharper as corollaries than they appear to be if proved from first principles using only the generalised gradient. An example is the chain rule Proposition 2, below, which, though specialised to the case of minimum points, is clearly sharper than its often quoted counterpart (2). We highlight Proposition 2, rather than simply listing further such results, because it can be used in the context of the generalised gradient to directly generate many such results. See Section 2 for further examples and discussion.

We use the *generalised Jacobian*, of Ioffe [**8**] and Ralph [**14**], of g at $\bar{x}$:

$$\partial g(\bar{x}) \overset{\text{def}}{=} \bigcap_{\lambda \in Y'} \{A \in \mathcal{L}(X,Y) \mid \lambda A \in \partial(\lambda g)(\bar{x})\}.$$

Proposition 2. *Under the hypotheses of Theorem 1, if $\bar{x}$ is a local minimiser of $f \circ g$ then*

$$0 \in \partial f(\bar{x})\partial g(\bar{x}) \overset{\text{def}}{=} \bigcup_{\lambda \in \partial f(\bar{x})} \lambda \partial g(\bar{x}).$$

https://doi.org/10.1017/S0004972700016178 Published online by Cambridge University Press

PROOF: From Ioffe [9], $0 \in \partial_a (f \circ g)(\bar{x})$. So (1) gives

$$0 \in \bigcup_{\lambda \in \partial_a f(\bar{x})} \partial_a (\lambda g)(\bar{x}).$$

We have $\partial_a f(\bar{x}) \subset \partial f(\bar{x})$ and, for each $\lambda \in Y'$,

$$\partial_a (\lambda g)(\bar{x}) \subset \partial(\lambda g)(\bar{x}) = \lambda \partial g(\bar{x}).$$

The equality, primarily the existence of $\partial g(\bar{x})$, follows from both [**8**, Theorem 10.4] and the proof of [**6**, Proposition 2.2]. □

A proof of this result without using approximate subdifferentials would be of interest. We believe one of the successes of approximate subdifferentials is that formulae such as (1) become available, whereas, in the context of generalised gradients, Proposition 2 has possibly never even been conjectured.

2. APPLICATIONS

2.1 METRIC REGULARITY.

Consider the system

$$(*) \qquad g(x) = 0, \quad x \in D$$

where D is a nonempty closed subset of X. Suppose that $\bar{x}$ is feasible, that is, $g(\bar{x}) = 0$ and $\bar{x} \in D$, and g is compactly Lipschitzian near $\bar{x}$. Let $\Gamma \overset{\text{def}}{=} g^{-1}(0) \cap D$, the feasible region.

This system is *metrically regular* at $\bar{x}$ if

$$\alpha \|g(x)\| \geqslant \operatorname{dist}_\Gamma (x) \overset{\text{def}}{=} \inf_{x' \in \Gamma} \|x - x'\|,$$

for some $\alpha > 0$ and each x in D near $\bar{x}$. The function $\operatorname{dist}_\Gamma$ is called the *distance function of* Γ, and is Lipschitz with Lipschitz constant 1 [**3**, Proposition 2.4.1].

There are various conditions using the Clarke [3] tangent or normal cones to D, that ensure the existence of such a constant α. More general results, however, have been possible using the approximate subdifferential than the generalised gradient. The standard method of proof uses a result of the following type:

PROPOSITION 3. *If* $(*)$ *is not metrically regular at* $\bar{x}$, *then there are sequences* $(x_n) \to \bar{x}$ *in* D, $(y_n) \to \bar{y} \overset{\text{def}}{=} g(\bar{x})$ *in* Y, *and* $(\delta_n) \to 0$ *in* $(0, \infty)$, *such that for each* n, $y_n \neq g(x_n)$, *and the function*

$$\phi_n : X \to \mathbb{R} : x \mapsto \|g(x_n) - y_n\| + \delta_n \|x - x_n\|$$

https://doi.org/10.1017/S0004972700016178 Published online by Cambridge University Press

has a (global) minimum, over D, at x_n.

Proof: See the proof of Ioffe [7, Theorem 1]. The result is also a corollary of Borwein [1, Theorem 2.2]. $\square$

Let U be a neighbourhood of $\bar{x}$ in which g is Lipschitz with Lipschitz constant $K > 0$, and assume without loss of generality that $(x_n) \subset U$ and $(\delta_n) \subset (0,1]$. Thus ϕ_n has the Lipschitz constant $K+1$ on U, hence by Clarke [3, Proposition 2.4.3], x_n is an unconstrained local minimiser of $\phi_n + (K+1)\operatorname{dist}_D$. Now ϕ_n is the composition of the mappings $F : Y \times X \times \mathbb{R} \longrightarrow \mathbb{R} : (y,x,\alpha) \mapsto \|y\| + \delta_n \|x\| + (K+1)\alpha$ and $G : X \longrightarrow Y \times X \times \mathbb{R} : x \mapsto (g(x) - y_n, x - x_n, \operatorname{dist}_D(x))$. The chain rule Proposition 2 yields

$$0 \in \partial \|\cdot\| (g(x_n) - y_n)\, \partial g(x_n) + \delta_n \partial \|\cdot\| (0) + (K+1)\partial \operatorname{dist}_D (x_n),$$

where the first norm is on Y, and the second on X. In particular, there exist λ_n in the unit sphere of Y' and ξ_n in the closed unit ball of X', such that

$$(**) \qquad 0 \in \lambda_n \partial g(x_n) + \delta_n \xi_n + (1+K) \operatorname{dist}_D (x_n).$$

This can be used to deduce sufficient conditions for metric regularity, as stated below.

When using the generalised gradient, however, the usual chain rule (2) yields that

$$\partial \psi_n(x_n) \subset \operatorname{cl} \operatorname{co}[\dot{\partial} \|\cdot\| (g(x_n) - y_n)\, \partial g(x_n)] + \delta_n \partial \|\cdot\| (0) + (K+1)\partial \operatorname{dist}_D (x_n).$$

ruling out the relatively simple inclusion $(**)$. Thus until now [1, 6], in this framework the assumption that Y' has an equivalent smooth norm has been made, to guarantee that firstly $\partial \|\cdot\| (y_n - g(x_n))$ is a singleton $\{\lambda_n\}$, where $\lambda_n \in Y'$ is the gradient of $\|\cdot\|$ at $y_n - g(x_n)$, and secondly

$$\partial \psi_n(x_n) = -\lambda_n \partial g(x_n).$$

In this more restrictive case, $(**)$ holds.

The following results on metric regularity are proven in a straightforward manner using $(**)$. See for example, [7, 1, 11, 6] for related proofs. Denote the open unit ball in X by $\mathbb{B}_X$, and the Clarke tangent cone [3] to D at $x \in D$ by $T_D(x)$. The core (algebraic interior) of a set in Y is the subset of its points y such that for any direction $v \in Y$, $y + tv$ also lies in the set for sufficiently small $t > 0$; in fact, since Banach spaces are Baire spaces, substituting the interiority operation for the core operation in each of the following conditions gives equivalent conditions, respectively. The generalised Jacobian ∂g is *(strongly) upper semicontinuous at* $\bar{x}$ if for each neighbourhood $\mathcal{N}$ of $\partial g(\bar{x})$ (in the norm topology of $\mathcal{L}(X,Y)$), there is a neighbourhood U of $\bar{x}$ such that $\partial g(x) \subset \mathcal{N}$ if $x \in U$.

https://doi.org/10.1017/S0004972700016178 Published online by Cambridge University Press

Theorem 4. *Suppose g is compactly Lipschitzian in a neighbourhood of $\bar{x} \in D$, where D is a closed subset of X. The following are both sufficient conditions for $(*)$ to be metrically regular at $\bar{x}$.*

(1) *For some neighbourhood U of $\bar{x}$,*

$$0 \in \operatorname{core} \bigcap_{x \in U \cap D,\, A \in \partial g(x)} \operatorname{cl} A(\mathbb{B}_X \cap T_D(x)).$$

(2) *∂g is (strongly) upper semicontinuous at $\bar{x}$, D is convex and*

$$0 \in \operatorname{core} \bigcap_{A \in \partial g(\bar{x})} \operatorname{cl} A(\mathbb{B}_X \cap (D - \bar{x})).$$

The second condition is an extension to the nonsmooth case of the well known results of Robinson [**16**], which apply to a continuously differentiable function g and a closed convex set D.

2.2 Optimal Multipliers via Exact Penalty Functions.

Jourani and Thibault [**12**] emphasise that the chain rule (2) for the generalised gradient, since it requires the closed convex hull operation, is not as useful as the approximate subdifferential in obtaining first-order necessary conditions for nonsmooth constrained optimisation. The chain rule Proposition 2 remedies this shortcoming.

As an illustration suppose $\phi : X \to \mathbb{R}$ is Lipschitz near $\bar{x}$, and $\bar{x}$ is a local minimiser of

$$\min_x \phi(x) \text{ subject to } g(x) = 0, x \in D.$$

Assume the system $(*)$ is metrically regular at $\bar{x}$. Then $\bar{x}$ is also a local minimiser of the penalty function

$$p_\kappa(x) \overset{\text{def}}{=} \phi(x) + \kappa(\|g(x)\| + \operatorname{dist}_D(x)),$$

for some $\kappa > 0$. To see this, observe that metric regularity says that the increase in the penalty term $\|g(x)\| + \operatorname{dist}_D(x)$ is at least linear in the distance from the feasible region, for $x \in D$ near $\bar{x}$; this easily extends to all x near $\bar{x}$. Hence for some κ, the rate of increase of $\kappa(\|g(x)\| + \operatorname{dist}_D(x))$ caused by violating the constraints near $\bar{x}$ is greater than the Lipschitz constant of ϕ near $\bar{x}$, and it follows that $p_\kappa(x) > p_\kappa(\bar{x})$ if x is near $\bar{x}$ but is not feasible. The penalty function p_κ is called *exact* because $\bar{x}$ is a local minimiser of p_κ. See Burke [**2**] for further details.

Since $\bar{x}$ is a local minimiser of p_κ, we have $0 \in \partial_a p_\kappa(\bar{x}) \cap \partial p_\kappa(\bar{x})$. If g is compactly Lipschitzian at $\bar{x}$, (1) gives

$$\begin{aligned} 0 \in \partial_a p_\kappa &\subset \partial_a \phi(x) + \kappa\Big[\bigcup_{\lambda \in Y', \|\lambda\| \leqslant 1} \partial_a(\lambda g)(\bar{x})\Big] + \kappa \partial_a \operatorname{dist}_D(\bar{x}) \\ &\subset \partial_a \phi(x) + \bigcup_{\lambda \in Y'} \partial_a(\lambda g)(\bar{x}) + N_D^G(\bar{x}), \end{aligned}$$

where $N_D^G(\overline{x})$ is the G-normal cone [10] to to D at $\overline{x}$, that is, the weak* closure of the union of sets $\alpha\partial_a \operatorname{dist}_D(\overline{x})$, $\alpha > 0$. The functionals λ in Y' (and also those in $N_D^G(\overline{x})$) are immediate generalisations of the Lagrange or Karush-Kuhn-Tucker multipliers so familiar in smooth optimisation [4].

In light of (2), one might expect that the corresponding formula using the generalised gradient must involve

$$\operatorname{cl}\operatorname{co}\bigcup_{\lambda\in Y'}\partial(\lambda g)(\overline{x}),$$

which obscures the multipliers λ so neatly found above. But from Proposition 2 (or the above formulae involving the approximate subdifferential),

$$\begin{aligned}0 &\in \partial\phi(x) + \kappa\partial \operatorname{dist}_C(\overline{x})\partial g(\overline{x}) + \kappa\partial \operatorname{dist}_D(\overline{x})\\ &\subset \partial\phi(x) + \bigcup_{\lambda\in Y'}\lambda\partial g(\overline{x}) + N_D(\overline{x}),\end{aligned}$$

where $N_D(\overline{x})$, the Clarke normal cone to D at $\overline{x}$, coincides with the weak* closure of the union of all sets $\alpha\partial \operatorname{dist}_D(\overline{x})$, $\alpha > 0$ [**3**, Proposition 2.4.2]. This extends multiplier rules for classical constrained optimisation problems, using generalised gradients and generalised Jacobians in place of gradients and Jacobians of smooth functions.

2.3 A Chain-Mean-Value Rule.

We have

Lemma 5. *For any x_1, x_2 near $\overline{x}$ there exists $t \in (0,1)$ such that $x \overset{\text{def}}{=} tx_1 + (1-t)x_2$ satisfies*

$$f\circ g(x_1) - f\circ g(x_2) \in \partial f(g(x))\partial g(x)(x_1 - x_2).$$

Proof: Choose x_1, x_2 in a convex neighbourhood U of $\overline{x}$ in which $f\circ g$ is Lipschitz. Let $h : [0,1] \to \mathbb{R} : t \mapsto f\circ g(tx_1 + (1-t)x_2) - tf\circ g(x_1) - (1-t)f\circ g(x_2)$. By continuity of h, there exists $t \in (0,1)$ that minimises either h or $-h$ over $[0,1]$. In either case an application of Proposition 2 yields the result. □

3. Appendix

Let $\phi : X \to \mathbb{R}\cup\{\infty\}$ and $\operatorname{dom}\phi \overset{\text{def}}{=} \{x \in X \mid \phi(x) < \infty\}$. Let $x \in \operatorname{dom}\phi$. For $u \in X$, let

$$\phi^{\downarrow}(x;u) \overset{\text{def}}{=} \liminf_{\substack{u'\to u\\ t\downarrow 0}}[\phi(x+tu') - \phi(x)]/t,$$

and define

$$\partial^{\downarrow}\phi(x) \overset{\text{def}}{=} \{\xi \in X' \mid \xi u \leqslant \phi^{\downarrow}(x;u),\ \forall u \in X\}.$$

https://doi.org/10.1017/S0004972700016178 Published online by Cambridge University Press

If $x \notin \operatorname{dom}\phi$, $\partial^{\downarrow}\phi(x)$ is defined as the empty set.

For any set D in X, let

$$\phi_D(x) \overset{\text{def}}{=} \begin{cases} \phi(x) & \text{if } x \in D, \\ \infty & \text{otherwise.} \end{cases}$$

By $x \xrightarrow{\phi} \overline{x}$ we mean $x \to \overline{x}$ and $\phi(x) \to \phi(\overline{x})$. By

$$\limsup_{x \xrightarrow{\phi} \overline{x}} \partial^{\downarrow}\phi_D(x)$$

we mean the set of weak* limit points of all sequences (ξ_n) such that for some $(x_n) \xrightarrow{\phi} \overline{x}$, $\xi_n \in \partial^{\downarrow}\phi_D(x_n)$ for each n.

DEFINITION 6: [9, Definition 1]

Let $\mathcal{A}$ denote the collection of all finite dimensional subspaces of X. The set

$$\partial_a\phi(x) \overset{\text{def}}{=} \bigcap_{L \in \mathcal{A}} \limsup_{x \xrightarrow{\phi} \overline{x}} \partial^{\downarrow}\phi_{x+L}(x)$$

is called the a(pproximate)-subdifferential of ϕ at $\overline{x}$.

If ϕ is Lipschitz near $\overline{x}$ then the above can be simplified using the fact that

$$\phi^{\downarrow}{}_{x+L}(x; u) = \phi^{-}(x; u) \overset{\text{def}}{=} \liminf_{t \downarrow 0}[\phi(x + tu) - \phi(x)]/t,$$

for each x near $\overline{x}$ and u in L.

We also recall the definition of a compactly Lipschitzian mapping:

DEFINITION 7: [12]

The mapping g is said to be *(strongly) compactly Lipschitzian* at a point $a \in X$ if there is a mapping K from X into the set of nonempty (strongly) compact subsets of Y and a mapping r from $X \times X$ into $[0, \infty)$ such that:

(1) $\lim_{x \to a,\, h \to 0} r(x; h) = 0$,

(2) there is $\delta > 0$ such that for each $h \in \delta\mathbb{B}_X$, $x \in a + \delta\mathbb{B}_X$ and $t \in (0, \delta)$

$$t^{-1}[g(x + th) - g(x)] \in K(h) + \|h\|\ r(x; th)\mathbb{B}_Y;$$

(3) $K(0) = \{0\}$ and the set set-valued mapping K is upper semicontinuous (that is, for each $u \in X$ and $\varepsilon > 0$, there is $\delta > 0$ such that $K(u') \subset K(u) + \varepsilon\mathbb{B}_Y$ for each $u' \in u + \delta\mathbb{B}_X$).

https://doi.org/10.1017/S0004972700016178 Published online by Cambridge University Press

If g is compactly Lipschitzian at every $x \in X$ then g is said to be *locally compactly Lipschitzian.*

Note that if Y is finite dimensional and g is Lipschitz near $\overline{x}$ with Lipschitz constant $k > 0$, then we see that g is compactly Lipschitzian at each point near $\overline{x}$ by taking $K(h) \overset{\text{def}}{=} k\,\|h\|\,\mathrm{cl}\,\mathbb{B}_Y$. In infinite dimensions, if g is continuously differentiable with Jacobian $\nabla g(\overline{x})$ at $\overline{x}$, then by taking $K(h) \overset{\text{def}}{=} \nabla g(\overline{x})h$ we see that g is compactly Lipschitzian at $\overline{x}$. For more examples of compactly Lipschitzian functions see Thibault [17].

References

[1] J.M. Borwein, 'Stability and regular points of inequality systems', *J. Optim. Theory Appl.* **48** (1986), 9–52.

[2] J.V. Burke, 'An exact penalization viewpoint of constrained optimization', *SIAM J. Control Optim.* **29** (1991), 968–998.

[3] F.H. Clarke, *Optimization and nonsmooth analysis* (Wiley Interscience, New York, 1983).

[4] R. Fletcher, *Practical methods of optimization*, 2nd ed. (John Wiley and Sons, New York, 1987).

[5] B.M. Glover, 'Locally compactly Lipschitzian mappings in infinite dimensional programming', *Bull. Austral. Mathem. Soc.* **47** (1993), 395–406.

[6] B.M. Glover and D. Ralph, 'First order approximations to nonsmooth mappings with applications to metric regularity', in *Preprint Series No. 3 — 1993* (Department of Mathematics, University of Melbourne, 1993).

[7] A.D. Ioffe, 'Regular points of Lipschitz mappings', *Trans. Amer. Math. Soc.* **251** (1979), 61–69.

[8] A.D. Ioffe, 'Nonsmooth analysis: differential calculus of nondifferentiable mappings', *Trans. Amer. Math. Soc.* **266** (1981), 1–56.

[9] A.D. Ioffe, 'Approximate subdifferentials and applications II: Functions on locally convex spaces', *Mathematika* **33** (1986), 111–128.

[10] A.D. Ioffe, 'Approximate subdifferentials and applications III: The metric theory', *Mathematika* **36** (1989), 1–38.

[11] A. Jourani and L. Thibault, 'Approximations and metric regularity in mathematical programming in Banach spaces', *Math. Oper. Res.* (1992) (to appear).

[12] A. Jourani and L. Thibault, 'Approximate subdifferential of composite functions', *Bull. Austral. Math. Soc.* **47** (1992), 443-455.

[13] B.S. Mordukhovich, 'Nonsmooth analysis with nonconvex generalized differentials and dual maps', *Dokl. Akad. Nauk. BSSR* **28** (1984), 976–979.

[14] D. Ralph, *Rank-1 Support functionals and the generalized Jacobian, piecewise linear homeomorphisms*, Ph.D Thesis (Computer Sciences Department, University of Wisconsin, Madison, 1990).

https://doi.org/10.1017/S0004972700016178 Published online by Cambridge University Press

[15] A.P. Robertson and W. Robertson, *Topological vector spaces* (Cambridge University Press, 1973).

[16] S.M. Robinson, 'Stability theory for systems of inequalities, Part II: Differentiable nonlinear systems', *SIAM J. Numer. Anal.* **13** (1976), 497–513.

[17] L. Thibault, 'On generalized differentials and subdifferentials of Lipschitz vector-valued functions', *Nonlinear Analisis Theory, Methods and Applications* **6** (1982), 1037–1053.

Department of Mathematics
University of Melbourne
Parkville, Vic. 3052
Australia

https://doi.org/10.1017/S0004972700016178 Published online by Cambridge University Press