SIAM J. OPTIM.
Vol. 31, No. 3, pp. 2080–2083


# THE STRUCTURE OF CONSERVATIVE GRADIENT FIELDS*

ADRIAN S. LEWIS† AND TONGHUA TIAN†

**Abstract.** The classical Clarke subdifferential alone is inadequate for understanding automatic differentiation in nonsmooth contexts. Instead, we can sometimes rely on enlarged generalized gradients called "conservative fields," defined through the natural pathwise chain rule: one application is the convergence analysis of gradient-based deep learning algorithms. In the semialgebraic case, we show that all conservative fields are in fact just Clarke subdifferentials plus normals of manifolds in underlying Whitney stratifications.

**Key words.** variational analysis, Clarke subdifferential, automatic differentiation, deep learning, subgradient descent, conservative field, stratification, semialgebraic

**AMS subject classifications.** 49J53, 90C56, 65K10, 68T07, 14P10

**DOI.** 10.1137/21M1393637

**1. Introduction.** Popular deep learning solvers like PyTorch [11] and TensorFlow [6] increasingly rely on automatic differentiation for gradient-based optimization algorithms. Given an input point $x \in \mathbf{R}^n$, the solver returns a gradient-like vector $g \in \mathbf{R}^n$ that depends not just on the objective function $f$ itself but rather on its algorithmic representation. At least when $f$ is smooth, we might hope that $g$ is the gradient $\nabla f(x)$, but even then, nonsmooth algorithmic ingredients may produce surprises. For example, the formula

$$f(s) \;=\; \big((-s)^+ + s\big) - s^+ \qquad (s \in \mathbf{R})$$

(where $s^+ = \max\{0, s\}$) always outputs the value zero, and yet one implementation [3, Appendix A.2] of automatic differentiation in TensorFlow outputs the derivative

$$g(s) = \begin{cases} 0 & (s \neq 0), \\ 1 & (s = 0). \end{cases} \tag{1.1}$$

For other discussions of the same issue, see [7, Chapter 14] and [9, 10].

Despite this disconcerting behavior, practitioners widely apply automatic differentiation to nonsmooth objective functions $f\colon \mathbf{R}^n \to \mathbf{R}$, as discussed in [8]. For the particular case of stochastic subgradient descent algorithms, see [5]. Fortunately, as demonstrated by [2], automatic differentiation at points $x \in \mathbf{R}^n$ typically does produce outputs $g(x) \in \mathbf{R}^n$ with gradient-like properties: time-dependent absolutely continuous trajectories $x(\cdot)$ satisfy the *chain rule*

$$\tfrac{d}{dt} f(x) \;=\; \langle \tfrac{d}{dt} x, g(x) \rangle \quad \text{for almost all } t, \tag{1.2}$$

thereby justifying the convergence of the stochastic subgradient descent method [5].

Thus motivated, Bolte and Pauwels [2] develop a novel and elegant notion of generalized derivative for locally Lipschitz objectives $f$ precisely around the chain

*Received by the editors January 22, 2021; accepted for publication March 24, 2021; published electronically August 20, 2021.
https://doi.org/10.1137/21M1393637
**Funding:** The first author's research was supported in part by National Science Foundation grant DMS-2006990.
†School of Operations Research and Information Engineering, Cornell University, Ithaca, NY 14853 USA (adrian.lewis@cornell.edu, tt543@cornell.edu).

rule. They consider *conservative fields*: closed set-valued mappings $G\colon \mathbf{R}^n \rightrightarrows \mathbf{R}^n$ with the property that the chain rule (1.2) holds providing that we always select $g(x) \in G(x)$. For example, backpropagation for deep neural networks in practice produces conservative fields [2, Corollary 6].

Being locally Lipschitz, $f$ is differentiable on a full-measure set $\Omega \subset \mathbf{R}^n$. As observed in [2], the value $G(x)$ for any conservative field $G$ must contain the set

$$\{\lim_r \nabla f(x_r) : x_1, x_2, \ldots \in \Omega,\ x_r \to x\}$$

and hence, if convex, also its convex hull $\partial f(x)$, the *Clarke subdifferential* [4].

An objective function with a conservative field is called *path differentiable*. The theoretical existence question has a long history, surveyed in [2] and dating back to [12], but in practice, objectives are always path differentiable: examples include smooth and convex functions and their sums and differences, as well as semialgebraic (or, more generally, tame [13]) functions. The Clarke subdifferential is the minimal convex-valued conservative field for any path differentiable objective. However, as the example (1.1) makes clear for automatic differentiation, we are forced to consider conservative fields larger than the Clarke subdifferential. What do they look like in general?

In this work we focus on the most concrete case, where objective functions and conservative fields are semialgebraic. (The tame generalization is immediate, but we do not pursue it here.) We prove an intuitive structural result, characterizing the conservative fields of an objective function as modest modifications of its Clarke subdifferential, arising simply by including normals to manifolds comprising Whitney stratifications of $\mathbf{R}^n$. For example, the conservative field (1.1) can arise from the stratification $\mathbf{R} = (-\infty, 0) \cup \{0\} \cup (0, +\infty)$. Thus, while the important idea of a conservative field is arrived at very differently from the notion of the Clarke subdifferential, in practice the two ideas are very close.

**2. Characterizing conservative fields.** Turning to the formal development, we consider set-valued operators on $\mathbf{R}^n$, by which we mean set-valued mappings $G\colon \mathbf{R}^n \rightrightarrows \mathbf{R}^n$. The *sum* of two operators $G$ and $H$ maps points $x \in \mathbf{R}^n$ to the sum $G(x) + H(x)$. We call $G$ *closed* if its graph $\{(x, y) \in \mathbf{R}^n \times \mathbf{R}^n : y \in G(x)\}$ is closed and *locally bounded* if every point in $\mathbf{R}^n$ has a neighborhood $\Omega \subset \mathbf{R}^n$ whose image $G(\Omega)$ is bounded. A *selection* of $G$ is an operator whose graph is contained in the graph of $G$. The following definition is from [2, Lemma 2].

DEFINITION 2.1. *A* conservative field *for a locally Lipschitz function $f\colon \mathbf{R}^n \to \mathbf{R}$ is a closed, locally bounded, nonempty-set-valued operator $G$ on $\mathbf{R}^n$ such that all absolutely continuous curves $x\colon [0,1] \to \mathbf{R}^n$ satisfy the following chain rule: for almost all $t \in [0,1]$,*

$$\tfrac{d}{dt} f\big(x(t)\big) \ = \ \langle \tfrac{d}{dt} x(t), g\rangle \quad \textit{for all } g \in G\big(x(t)\big).$$

We next consider smooth stratifications of sets in a Euclidean space $\mathbf{E}$; in all of our discussions of functions and manifolds, "smooth" simply means continuously differentiable. For any smooth manifold $\mathcal{M} \subset \mathbf{E}$, we denote the tangent and normal spaces to $\mathcal{M}$ at any point $x \in \mathcal{M}$ by $T_{\mathcal{M}}(x)$ and $N_{\mathcal{M}}(x)$. A finite collection $\mathcal{W}$ of disjoint smooth manifolds in $\mathbf{E}$ comprise a *Whitney stratification* (of their union) if the following condition holds for all manifolds $\mathcal{M}$ and $\mathcal{M}'$ in $\mathcal{W}$:

$$\left.\begin{array}{ll} x_1, x_2, \ldots \in \mathcal{M}, & y_r \in N_{\mathcal{M}}(x_r) \\ x_r \to x \in \mathcal{M}', & y_r \to y \end{array}\right\} \quad \Rightarrow \quad \left\{\begin{array}{l} \mathcal{M}' \subset \operatorname{cl} \mathcal{M}, \\ y \in N_{\mathcal{M}'}(x). \end{array}\right.$$

In particular, we can associate any Whitney stratification $\mathcal{W}$ of $\mathbf{R}^n$ with a closed *normal operator* $\Phi_{\mathcal{W}}$ on $\mathbf{R}^n$ by setting $\Phi_{\mathcal{W}} = N_{\mathcal{M}}$ on each manifold $\mathcal{M}$ in $\mathcal{W}$. We call $\mathcal{W}$ *semialgebraic* if each $\mathcal{M}$ in $\mathcal{W}$ is semialgebraic.

We can now state our result.

THEOREM 2.2 (semialgebraic conservative fields). *Given a semialgebraic locally Lipschitz function $f\colon \mathbf{R}^n \to \mathbf{R}$, a semialgebraic set-valued operator is a conservative field for $f$ if and only if it is a closed, locally bounded, nonempty-valued selection of the sum of the Clarke subdifferential $\partial f$ and the normal operator for a semialgebraic Whitney stratification of $\mathbf{R}^n$.*

*Proof.* Consider any semialgebraic conservative field $G$ for $f$. By [2, Theorem 4], there exists a Whitney stratification $\mathcal{W}$ of $\mathbf{R}^n$ such that $f$ is smooth on each manifold $\mathcal{M}$ in $\mathcal{W}$, and at each point $x \in \mathcal{M}$, the Riemannian gradient $\nabla_{\mathcal{M}} f(x) \in T_{\mathcal{M}}(x)$ satisfies

$$G(x) \ \subset \ \nabla_{\mathcal{M}} f(x) + N_{\mathcal{M}}(x). \tag{2.1}$$

The proof in [2], using stratification techniques from [13], makes clear that we can assume $\mathcal{W}$ to be semialgebraic. By [1, Lemma 8], there exists a semialgebraic Whitney stratification of the graph of $f$ that maps via the canonical projection $\mathbf{R}^n \times \mathbf{R} \to \mathbf{R}^n$ onto a semialgebraic Whitney stratification $\mathcal{W}'$ of $\mathbf{R}^n$ that is "compatible" with $\mathcal{W}$: in other words, each manifold in $\mathcal{W}$ is a union of manifolds in $\mathcal{W}'$.

Now consider any point $x \in \mathbf{R}^n$. There exist unique manifolds $\mathcal{M}$ in $\mathcal{W}$ and $\mathcal{M}'$ in $\mathcal{W}'$ containing $x$, and $\mathcal{M}'$ must be a submanifold of $\mathcal{M}$, so

$$N_{\mathcal{M}}(x) \ \subset \ N_{\mathcal{M}'}(x). \tag{2.2}$$

By the definition of the Riemannian gradient, we have

$$\nabla_{\mathcal{M}} f(x) \ \in \ \nabla_{\mathcal{M}'} f(x) + N_{\mathcal{M}'}(x). \tag{2.3}$$

On the other hand, by [1, Proposition 4], we have

$$\partial f(x) \ \subset \ \nabla_{\mathcal{M}'} f(x) + N_{\mathcal{M}'}(x).$$

Since $f$ is locally Lipschitz, there exists a vector $g \in \partial f(x)$. We deduce

$$\nabla_{\mathcal{M}'} f(x) + N_{\mathcal{M}'}(x) \ = \ g + N_{\mathcal{M}'}(x) \ \subset \ \partial f(x) + N_{\mathcal{M}'}(x). \tag{2.4}$$

Combining the inclusions (2.1), (2.2), (2.3), and (2.4), we deduce

$$G(x) \ \subset \ \partial f(x) + N_{\mathcal{M}'}(x),$$

so $G$ is a selection of the sum $\partial f + \Phi_{\mathcal{W}'}$, as required.

Conversely, consider a semialgebraic Whitney stratification $\mathcal{W}$ of $\mathbf{R}^n$. By [2, Corollary 2 and Proposition 2], the subdifferential $\partial f$ is a conservative field for $f$. On the other hand, for any radius $r > 0$, the truncated normal operator defined by

$$\Phi^r_{\mathcal{W}}(x) \ = \ \Phi_{\mathcal{W}}(x) \cap rB \qquad (x \in \mathbf{R}^n),$$

where $B \subset \mathbf{R}^n$ is the closed unit ball, is a conservative field for the identically zero function, by [2, Theorem 3]. Hence $\partial f + \Phi^r_{\mathcal{W}}$ is a conservative field for $f$, by [2, Corollary 4].

Now consider any closed, locally bounded, nonempty-valued selection $G$ of the operator $\partial f + \Phi_{\mathcal{W}}$. If $G$ is not a conservative field, then Definition 2.1 fails for some absolutely continuous curve $x\colon [0,1] \to \mathbf{R}^n$. The image $C = x([0,1])$ is compact, so since $G$ is locally bounded, the image $G(C)$ is bounded. Since $f$ is locally Lipschitz, the image $\partial f(C)$ is also bounded. We deduce $G(C) - \partial f(C) \subset rB$ for some radius $r > 0$. All points $x \in C$ therefore satisfy $G(x) \subset \partial f(x) + \Phi^r_{\mathcal{W}}(x)$, since, by assumption, for any vector $g \in G(x)$ there exists a subgradient $y \in \partial f(x)$ such that $g \in y + \Phi_{\mathcal{W}}(x)$, so in fact $g \in y + \Phi^r_{\mathcal{W}}(x)$. But $\partial f + \Phi^r_{\mathcal{W}}$ is conservative, contradicting the failure of Definition 2.1. ∎

## REFERENCES


[1] J. Bolte, A. Daniilidis, A.S. Lewis, and M. Shiota, *Clarke subgradients of stratifiable functions*, SIAM J. Optim., 18 (2007), pp. 556–572.

[2] J. Bolte and E. Pauwels, *Conservative set valued fields, automatic differentiation, stochastic gradient methods and deep learning*, Math. Program. (2020).

[3] J. Bolte and E. Pauwels, *A mathematical model for automatic differentiation in machine learning*, in Proceedings of NeurIPS, 2020.

[4] F.H. Clarke, *Optimization and Nonsmooth Analysis*, Wiley Interscience, New York, 1983.

[5] D. Davis, D. Drusvyatskiy, S. Kakade, and J.D. Lee, *Stochastic subgradient method converges on tame functions*, Found. Comput. Math., 20 (2020), pp. 119–154.

[6] M. Abadi et al., *TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems*, tensorflow.org, 2015.

[7] A. Griewank and A. Walther, *Evaluating Derivatives: Principles and Techniques of Algorithmic Differentiation*, SIAM, Philadelphia, 2008.

[8] A. Griewank, A. Walther, S. Fiege, and T. Bosse, *On Lipschitz optimization based on gray-box piecewise linearization*, Math. Program., 158 (2016), pp. 383–415.

[9] S.M. Kakade and J.D. Lee, *Provably correct automatic subdifferentiation for qualified programs*, in Proceedings of NeurIPS, 2018, pp. 7125–7135.

[10] W. Lee, H. Yu, X. Rival, and H. Yang, *On correctness of automatic differentiation for non-differentiable functions*, in Proceedings of NeurIPS, 2020.

[11] A Paszke et al., *Automatic differentiation in PyTorch*, in Proceedings of NIPS, 2017.

[12] M. Valadier, *Entraînement unilatéral, lignes de descente, fonctions lipschitziennes non pathologiques*, C. R. Acad. Sci. Paris Sér. I Math., 308 (1989), pp. 241–244.

[13] L. van den Dries and C. Miller, *Geometric categories and o-minimal structures*, Duke Math. J., 84 (1996), pp. 497–540.