A variational model for nonsmooth automatic differentiation

Jérôme Bolte,
Joint work with Edouard Pauwels

Toulouse School of Economics
Université Toulouse Capitole & ANITI,
France

Pibrac, July 6th, 2020

Three parts

Our question somehow concerns *formal Clarke subdifferentiation:*

What does the chain rule output out of its validity domain?
Do we obtain a Jacobian of some sort?

I) Observational informal part (model/motivational case: training feedforward neural networks).

II) Theoretical answers: the zero circulation idea and conservative fields

III) Asymptotics & vanishing stepsizes algorithms

Observational part

A model for compositional calculus: conservative fields ($q = 1$)

Asymptotics and algorithms

Our starting point: neural nets training

$$\text{Minimize} \quad \frac{1}{N}\sum_{i=1}^{N} \underbrace{\| \overbrace{\sigma_l(\mathbf{W}_l(\ldots(\sigma(\mathbf{W}_2\sigma(\mathbf{W}_1 x_i + \mathbf{b}_1) + \mathbf{b}_2))\ldots) + \mathbf{b}_l)}^{\text{Prediction function}} - y_i \|^2}_{f_i(\mathbf{W})}$$

with

- $\mathbf{W_1}, \mathbf{b_1}, \ldots, \mathbf{W_l}, \mathbf{b_l}$ variable matrices/vectors, aggregated into $\mathbf{W}$
- $(x_i, y_i)_{i \in N}$ (training) data,
- $\sigma_i : \mathbb{R} \to \mathbb{R}$, acts entrywise on vectors $\sigma(V) = \left[\sigma(V_j)\right]_j$
- Ex. $\boxed{\sigma(t) = \max(0, t) := \mathrm{relu}(t)}$

Write $\min_W \frac{1}{N}\sum_{i=1}^N f_i(W)$. Use stochastic "gradient" descent

$$W^{k+1} = W^k - \frac{\gamma_k}{b}\left[\text{gradient } f_{i_1}(W^k) + \ldots + \text{gradient } f_{i_b}(W^k)\right]$$

where

$$\begin{cases} \{i_1, \ldots, i_b\} \text{ is drawn uniformly at random within } \{1, \ldots, N\} \\ \gamma_k \to 0 \end{cases}$$

Our starting point: neural nets training

$$\text{Minimize} \quad \frac{1}{N}\sum_{i=1}^{N} \underbrace{\| \overbrace{\sigma_l(\mathbf{W}_l(\ldots(\sigma(\mathbf{W}_2\sigma(\mathbf{W}_1 x_i + \mathbf{b}_1) + \mathbf{b}_2))\ldots) + \mathbf{b}_l)}^{\text{Prediction function}} - y_i\|^2}_{f_i(\mathbf{W})}$$

with

- $\mathbf{W_1}, \mathbf{b_1}, \ldots, \mathbf{W_l}, \mathbf{b_l}$ variable matrices/vectors, aggregated into $\mathbf{W}$
- $(x_i, y_i)_{i\in N}$ (training) data,
- $\sigma_i : \mathbb{R} \to \mathbb{R}$, acts entrywise on vectors $\sigma(V) = \left[\sigma(V_j)\right]_j$
- Ex. $\boxed{\sigma(t) = \max(0, t) := \mathrm{relu}(t)}$

Write $\min_W \frac{1}{N}\sum_{i=1}^N f_i(W)$. Use stochastic "gradient" descent

$$W^{k+1} = W^k - \frac{\gamma_k}{b}\left[\mathrm{backprop}\, f_{i_1}(W^k) + \ldots + \mathrm{backprop}\, f_{i_b}(W^k)\right]$$

where

$$\begin{cases} \{i_1, \ldots, i_b\} \text{ is drawn uniformly at random within } \{1, \ldots, N\} \\ \gamma_k \to 0 \end{cases}$$

What is backprop as a mathematical object?

- backprop (Rumelhart et al.) is obtained by "using formal differentiation":

  1. Apply the chain rule
  2. Use (Clarke) subgradients when you hit a nonsmooth part

  In practice, TensorFlow, PyTorch etc... use this principle.

- Fast and efficient way to obtain very sharp numerical derivatives: an instance *automatic/algorithmic differentiation*:
- But we only focus on the theoretical premises: 1. and 2.

Ingredient 1: Clarke Jacobians

Functions are (locally) Lipschitz continuous.

Notation: $f'(x) \simeq \mathrm{Jac}\, f(x)$ when f is differentiable.

- $f : \mathbb{R}^p \to \mathbb{R}^q$ loc. Lipschitz. Rademacher theorem: "f is differentiable almost everywhere"

$$\begin{aligned}&\mathrm{Jac}^{\,c} f(x)\\ &= \mathrm{conv}\left\{M \in \mathbb{R}^{p\times q} : x^k \to x, f \text{ differentiable at } x_k, \mathrm{Jac}\, f(x^k) \to M\right\}\end{aligned}$$

$$q = 1, \text{ then } \mathrm{Jac}^{\,c} f = \partial^c f$$

- So

$$\mathrm{Jac}^{\,c} f = \mathrm{Jac}\, f \text{ a.e.}$$

Ingredient 2: chain rule

- Consider f with a **compositional representation**

$$f = g_1 \circ \ldots \circ g_m$$

(recall $\|\sigma_l(\mathbf{W}_l(\ldots(\sigma(\mathbf{W}_2\sigma(\mathbf{W}_1 x_i + \mathbf{b}_1) + \mathbf{b}_2))\ldots) + \mathbf{b}_l) - y_i\|^2$)

- For each i, x, choose $D_{g_i}(x) \in \mathrm{Jac}^c g_i(x)$
- Example in Deep Learning:

$$D_{\mathrm{relu}}(s) = \begin{cases} 1 & \text{if } s > 0 \\ 0 & \text{if } s \leq 0 \end{cases}$$

In short $\mathrm{relu}'(0) = 0$ (TensorFlow, PyTorch).

- **Chain-rule the D_{g_i}'s**

$$\begin{aligned} & D_f(x) \\ :=& D_{g_1}(g_2(\ldots(g_m(x))\ldots)) \times D_{g_2}(g_3(\ldots(g_m(x))\ldots)) \ldots \times D_{g_m}(x) \end{aligned}$$

When the g_i are differentiable

$$D_f = \mathrm{Jac}\, f. \text{ Otherwise ?}$$

Exploitation of ingredients 1 and 2

- Exploit $D_f(x)$ to devise algorithms, as

$$x^{k+1} = x^k - \gamma_k D_f(x^k) \text{ with } \gamma_k \to 0.$$

- Example:

$$W^{k+1} = W^k - \frac{\gamma_k}{q}\left[\mathrm{backprop}\, f_{i_1}(W^k) + \ldots + \mathrm{backprop}\, f_{i_q}(W^k)\right]$$

Automatic differentiation

- A long history, numerous results, many implementations (our focus was on TensorFlow).
  Many application domains: design optimization, computational fluid dynamics, physical modeling, optimal control, structural mechanics, atmospheric sciences, and computational finance
- The algorithmic and numerical aspects are delicate:
  Griewank and Walther (2008), Evaluating Derivatives.
- **Our focus: understand the practice of using chain rule out of his obvious validity domain.**

Meaning of D_f? The result of a dangerous cocktail...

1. Start with $f = g_1 \circ \dots g_m$
2. Build $D_f(x) := D_{g_1}(g_2(\dots(g_m(x))\dots)) \times \dots \times D_{g_m}(x)$

- **Non uniqueness.** Compositional representation

$$f = g_1 \circ \dots \circ g_m \text{ is NOT UNIQUE}$$

- **Absence of qualification conditions.** In general

$$D_{g_1}(g_2 \circ \dots \circ g_m(x)) \circ D_{g_2}(g_3 \circ \dots \circ g_m(x)) \dots \circ D_{g_m}(x) \notin \mathrm{Jac}^c f(x)$$

unless "transversality conditions/QC" are present

Let's stick to practice → accept the two above imperfections and investigate the consequences

All remarks we make are observable using TensorFlow.

Issue I: outputs are partly unpredictible

- $\mathrm{relu}(t) = \max\{0, t\}$, with $\mathrm{relu}'(0) = 0$ (implemented on TensorFlow or PyTorch)

$$\mathrm{relu}_2 \colon t \mapsto \mathrm{relu}(-t) + t, \qquad \mathrm{relu}_3 \colon t \mapsto \frac{1}{2}(\mathrm{relu}(t) + \mathrm{relu}_2(t)).$$

$$\mathrm{relu} = \mathrm{relu}_2 = \mathrm{relu}_3$$

- Formal differentiation gives

$$\mathrm{relu}_2'(0) = 1 \text{ and } \mathrm{relu}_3'(0) = 1/2.$$

The absurd behavior results both from non uniqueness and the abscence of QC

Issue II: artificial critical points

- $\mathrm{zero} = \mathrm{relu}_2 - \mathrm{relu}$ is the null function but

$$\mathrm{zero}'(0) = 1$$

- $\boxed{x - \mathrm{zero}(x) = x \text{ has a zero derivative at 0 (!?)}}$

- Unexpected derivatives and artificial critical points

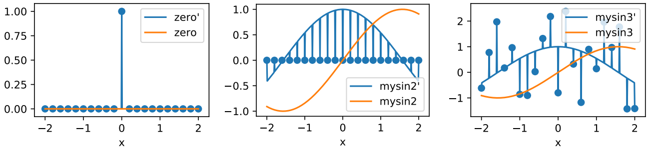

Figure: At the center : artificial critical points

Issue III: non-differentiability zones are not generally activated

- Belief: *"When we compute* $\operatorname{Jac}^c g_1((g_2 \circ \ldots \circ g_m)(x)) \circ \ldots \circ \operatorname{Jac}^c g_m(x)$ *we do not see the singularities of the* g_i *in general"*

  Wrong: $g_1(x) = |x|$, $g_2 : \mathbb{R}^p \to \mathbb{R}$, $g_1 \circ g_2 = |g_2|$ the non differentiability zone is $g_2^{-1}(0)$
- Nonsmooth zones of neural net can be significantly activated

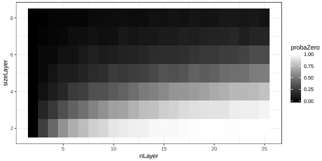

Figure: Estimation of the probability of applying relu to 0 in a feedforward network the weights of the linear term are sampled uniformly at random between -1 and 1. Variations in size and number of layers are also considered.

- A question is do we even have $D_f = \nabla f$ almost everywhere?
  Works in these directions: Griewank, Nesterov, Kakade-Lee...

Issue IV: Impossibility "theorem"

Can we build a larger "Jacobian operator " Jac^A on Lipschitz functions satisfying

(a) $\mathrm{Jac}^A f \supset \mathrm{Jac}^c f$ for all f Lipschitz from $\mathbb{R}^p$ to $\mathbb{R}^q$, $p, q \geq 0$

(b) the chain rule

Theorem (Automatic differentiation does not induce an operator on functions)

There is no nontrivial operator on functions satisfying (*a*) *and* (*b*).

What does formal subdifferentiation compute?

Observations

- Spurious outputs and artificial critical points
- Nonsmooth parts are significantly activated
- Formal subdifferentiation/automatic differentiation does not yield a differential operator

Questions

- Variational meaning of the D_f's without using operators?
- Impact of artificial values?
- Behavior of first order methods

An "operator-free" approach?

- $V : \mathbb{R}^p \to \mathbb{R}^p$ a continuous vector field.
  Circulation along a differentiable loop $\gamma : [0, 1] \to \mathbb{R}^n$ ($\gamma(0) = \gamma(1)$):

$$\int_0^1 \langle V(\gamma(t)), \dot{\gamma}(t)\rangle dt$$

- If $V = \nabla f$ the circulation is always 0

Lemma (Poincaré)

$$\int_0^1 \langle V(\gamma(t)), \dot{\gamma}(t)\rangle = 0 \ \forall \ loop \ \gamma \iff \exists f : \mathbb{R}^n \to \mathbb{R} \ C^1 \ such \ that \ V = \nabla f$$

An "operator-free" approach: The zero circulation idea

Assumptions $D : \mathbb{R}^p \rightrightarrows \mathbb{R}^p$ **nonempty compact values, closed graph,**
i.e., $D(x) \neq \emptyset$ is compact and $\{(x, y) : y \in D(x)\}$ is closed.

- Zero circulation à la Poincaré:

$$\int_0^1 \langle D(\gamma(t)), \dot{\gamma}(t)\rangle \, dt = \{0\},$$

for all loop absolutely continuous $\gamma : [0, 1] \to \mathbb{R}^p$.

Meaning. For any measurable selection $v \colon [0, 1] \to \mathbb{R}^p$, $v(t) \in D(\gamma(t))$ for all t, we have $\int_0^1 \langle v(t), \dot{\gamma}(t)\rangle \, dt = 0$.

- D is called a *conservative set-valued field.*

Similar def for the Jacobian situation.

Potential functions of conservative fields

- $D\colon \mathbb{R}^p \rightrightarrows \mathbb{R}^p$ a conservative field.
  It corresponds to a "unique" potential function f:

$$\begin{aligned} f(x) &= f(0) + \int_0^1 \langle \dot{\gamma}(t), D(\gamma(t)) \rangle \, dt && (1) \\ &= f(0) + \int_0^1 \max_{v \in D(\gamma(t))} \langle \dot{\gamma}(t), v \rangle \, dt && (2) \\ &= f(0) + \int_0^1 \min_{v \in D(\gamma(t))} \langle \dot{\gamma}(t), v \rangle \, dt && (3) \end{aligned}$$

  with γ AC with $\gamma(0) = 0$ and $\gamma(1) = x$.
- f is a *potential function for D* or *D admits f as a potential*, or *D is a conservative field for f*.

Fundamental properties

Theorem (Conservative fields and gradients)

If $f : \mathbb{R}^n \to \mathbb{R}$ is locally Lipschitz and D_f is conservative for f then

$$D_f(x) = \{\nabla f(x)\} \text{ a.e.}$$

Corollary (The Clarke subdifferential as a minimal conservative field)

If D_f is a conservative field for f, then

$$\mathrm{conv} D_f(x) \supset \partial^c f(x), \ \ \forall x \in \mathbb{R}^p$$

and $\partial^c f$ is conservative.

Fundamental examples with the Clarke subdifferential

If f is locally Lipschitz

(i) f is regular: semi-convex (or semi-concave), i.e., for all compact set $f + \alpha\|x\|^2$ is convex, prox regular etc...

(ii) f semi-algebraic (or definable)

then $\partial^c f$ is conservative (for f)

Actually

$$\partial^c f \text{ conservative} \iff f \text{ has a chain rule for the Clarke}$$

Proof The first case is classical. The last one uses stratification theory B-Daniilidis-Lewis-Shiota and Davis-Drusvyatskiy-Kakade-Lee

An operatorless calculus

We do not have an operator, but we have a convenient calculus!

Proposition

The linear combination of conservative fields is a conservative field.

If D_f and D_g have the zero circulation property then $\lambda D_f + \mu D_g$ has the zero circulation property and it is attached to $\lambda f + \mu g$ whenever $\lambda, \mu \in \mathbb{R}$.

Proposition

The composition of conservative Jacobians is a conservative Jacobian

The semi-algebraic/definable case

$$f = g_1 \circ \ldots \circ g_m \text{ with all the } g_i \text{ SA}$$

Theorem (The meaning of chain-ruled operators)

- *For each g_i the "user" provides a semi-algebraic selection $D_{g_i} \in \mathrm{Jac}^c g_i$*
- *Set*

$$D_f(x) = D_{g_1}(g_2(\ldots(g_m(x))\ldots)) \times \ldots \times D_{g_m}(x)$$

Then D_f is a conservative field for f, thus

$$\frac{d}{dt} f(\gamma(t)) = \langle \dot{\gamma}(t), D_f(\gamma(t)) \rangle$$

for all AC curve γ.

Proof. Relies on Whitney stratifications and a projection formula

We answered our initial question with backprop!!!

Conservative fields in a nutshell: zero circulation set-valued maps

- **Conservative=gradient a.e.**
- **Major examples.** The Clarke subdifferential of
  - semi-convex or other regular classes
  - semi-algebraic
- **The formal derivation principle**

  $$D_f := D_{g_1} \circ \ldots \circ D_{g_m}$$

  is conservative whenever the D_{g_i} are conservative
- **Backpropagation in deep learning**: backprop is a conservative field (generated, by e.g. TensorFlow), thus the first-order mapping

  $$W \to \sum_{i=1}^{N} \mathrm{backprop}\, f_i(W) \text{ is a conservative field.}$$

  More generally nonsmooth automatic differentiation process

Observational part

A model for compositional calculus: conservative fields ($q = 1$)

Asymptotics and algorithms

Questionning: asymptotic and algorithms with conservative fields

New model "conservative set-valued fields" (applies to backprop)

Major questions

- Optimizing dynamics
- Impact of spurious points and artificial points

Artificial critical points & asymptotics

Given $f : \mathbb{R}^p \to \mathbb{R}$ and D_f, with D_f conservative for f

- D-critical points $D_f - \mathrm{crit} = \{x \in \mathbb{R}^p : D_f(x) \ni 0\} \subset \mathbb{R}^p$
- D critical values $f(D_f - \mathrm{crit}) \subset \mathbb{R}$
- Artificial critical points: $\mathrm{art}\, D_f = \{x \in \mathbb{R}^p : 0 \in D_f(x) \text{ and } 0 \notin \partial^c f(x)\}$

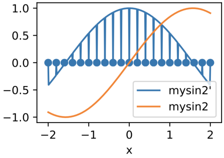

Figure: $f = \sin$. The chosen conservative field in blue $D_{\sin}$ yields many artificial critical points

- In DL $\mathrm{backprop}\, f_1(W) + \ldots + \mathrm{backprop}\, f_N(W) \notin \partial^c(f_1 + \ldots + f_N)(W)$ in general

Artificial critical points & asymptotics

Assume D_f has convex values

- Model dynamics "conservative gradient descent"

$$\dot{x}(t) + D_f(x(t)) \ni 0 \text{ a.e. on } [0, +\infty)$$

  where $x : [0, +\infty) \to \mathbb{R}^p$ is AC is such that $x(0) = x_0$.

- D_f-critical points are stationary

- Theorem (B-Pauwels)

If (f, D_f) are SA, bounded trajectories converges to D_f critical points.

Proof: "Conservative versions" of the projection formula, Sard's theorem, KL inequalities, as in Bolte-Daniilidis-Lewis and Bolte-Daniilidis-Lewis-Shiota.

Stochastic gradient with mini-batch

- Nonsmooth nonconvex: Davies-Drusvyatskiy-Kakade-Lee, Majewski-Miasojedow-Moulines, Adil's PhD thesis, Bianchi-Hachem-Schechtman, Chizat-Bach...
- Consider

$$\min_{x \in \mathbb{R}^p} f(x) = \frac{1}{N} \sum_{i=1}^{N} f_i(x),$$

with conservative fields $D_{f_i} : \mathbb{R}^p \mapsto \mathbb{R}$, $i = 1, \ldots, N$.

- $x_0 \in \mathbb{R}^p$, step sizes $\gamma_k > 0$ and a sequence of *iid* indices $(I_k)_{k \in \mathbb{N}}$ taken uniformly in the nonempty subsets of $\{0, \ldots, N\}$,

$$x_{k+1} = x_k - \gamma_k \frac{1}{|I_k|} \sum_{i \in I_k} D_{f_i}(x_k),$$
$$I_k \subset \{1, \ldots, N\}.$$

Stochastic gradient with mini-batch II

$$x_{k+1} = x_k - \gamma_k \frac{1}{|I_k|} \sum_{i \in I_k} D_{f_i}(x_k), \; I \subset \{1, \dots, n\}. \tag{4}$$

$$\text{Set } \; D_f = \frac{1}{N} \mathrm{conv} \sum_{i=1}^{N} D_{f_i}$$

Theorem (Convergence)

Assume $\gamma_k = o(1/\log k)$ and f semi-algebraic. For all x_0 such that x_k is almost surely bounded, then almost surely,

- *$f(x_k)$ converges as k tends to infinity to a D_f critical value.*
- *all accumulation points, $\bar{x}$, of $(x_k)_{k \in \mathbb{N}}$ are D_f-critical points: $0 \in D_f(\bar{x})$.*

Proof. Use theory of Benaim-Hofbauer-Sorin on differential inclusions and ideas from Davies et al. which proved a similar result with $\partial^c f$

Artificial critical points are never seen

- Deep learning problem

$$\min_W J(W) := \frac{1}{N}\sum_{i=1}^{N} \underbrace{\|\sigma_l(\mathbf{W}_l(\ldots(\sigma(\mathbf{W}_2\sigma(\mathbf{W}_1 x_i + \mathbf{b}_1) + \mathbf{b}_2))\ldots) + \mathbf{b}_l) - y_i\|^2}_{f_i(\mathbf{W})}$$

with e.g.,

$$(\star) \qquad \forall i,\ \sigma_i = \mathrm{relu},\ D_\sigma(s) = \begin{cases} 1 & \text{if } s > 0 \\ 0 & \text{if } s \leq 0 \end{cases}$$

Many other choices are possible .

- Optimization phase

$$W^{k+1} = W^k - \frac{\gamma_k}{b}\left[D_{f_{i_1}}(W^k) + \ldots + D_{f_{i_b}}(W^k)\right]$$

where

  - $i_1, \ldots, i_b$ is drawn uniformly at random within $\{1, \ldots, N\}$
  - D_{f_i} comes from the choice $(\star)$ and chain rule

Artificial critical points are never seen

Theorem (B-Pauwels)

There exist

- *a finite subset of steps* $F \subset (0, +\infty)$ *& zero measure, meager* $N \subset \mathbb{R}^p$

such that for any

- *positive sequence* $\gamma_k = o(1/\log k)$ *avoiding values in* F
- *initialization* $x_0 \in \mathbb{R}^p \setminus N$,

we have

- $J(W^k)$ *converges towards a* **Clarke critical value** *almost surely,*
- *the cluster points of* W^k *are* **Clarke critical point** *almost surely,*

whenever the sequence is almost surely bounded.

More precise results: B-Pauwels-Rios-Zertuche oscillation analysis. *Long term dynamics of the subgradient method for Lipschitz path differentiable functions*

References

For this work

- Bolte, J., Pauwels, E. (2020). Conservative set valued fields, automatic differentiation, stochastic gradient methods and deep learning. Math. Prog.
- Bolte, J., Pauwels, E. (2020). A mathematical model for automatic differentiation in machine learning. arXiv:2006.02080.
- Castera C., Bolte J., Févotte C., Pauwels E., An Inertial Newton Algorithm for Deep Learning

On automatic differentiation

- Griewank, A., Walther, A. (2008). Evaluating derivatives: principles and techniques of algorithmic differentiation. SIAM
- Griewank, A. (2013). On stable piecewise linearization and generalized algorithmic differentiation. Optim. Meth. and Software, 28(6).

Vanishing stepsizes method in nonsmooth analysis

- Benaim, M., Hofbauer, J., Sorin, S. (2005). Stochastic approximations and differential inclusions. SIAM J. on Control and Optimization, 44(1).
- Davis, D., Drusvyatskiy, D., Kakade, S., Lee, J. (2020). Stochastic subgradient method converges on tame functions. Found. of comput. math., 20(1).
- Bianchi, P., Hachem, W., Schechtman, S. (2020). Convergence of constant step stochastic gradient descent for non-smooth non-convex functions. arXiv:2005.08513.