The Mathematics of Automatic Differentiation

Coordinators: J. Bolte and E. Pauwels

March 24, 2022

ANITI

Université Fédérale Toulouse Midi-Pyrénées

Nonsmooth automatic differentiation

What is automatic differentiation?

- Why learning and derivation are connected?

  learning in AI = ”weight” tuning = differentiation of some loss

- Autodiff is a fast algorithm for computing derivatives using chain rule: *a fast learning mechanism*.

Nonsmooth automatic differentiation

What is automatic differentiation?

- Why learning and derivation are connected?

$$\boxed{\textbf{learning in AI} = \textbf{"weight" tuning} = \textbf{differentiation of some loss}}$$

- Autodiff is a fast algorithm for computing derivatives using chain rule: *a fast learning mechanism*.

The power and the versatility of Autodiff

- **Autodiff acts on all *numerical programs***, i.e., programs whose inputs-outputs are vectors.

$$\text{Program examples: } \begin{cases} \text{Neural-Networks} \\ \text{Solvers in Optimization, Robotics, Mechanics, PDE, Control.} \\ \text{Numerical algorithms} \end{cases}$$

- Autodiff is the cornerstone of the "TensorFlow revolution" and **key to the "AI revolution"**

Modern programs are non differentiable by nature

Non-smoothness of numerical programs principally arises from

- **Conditional statements** (if, then, else)
- **Solution maps** (Solvers in applied fields: Physics, Robotics...)
- **Regularization** (Statistics or inverse problems)

Modern programs are non differentiable by nature

Non-smoothness of numerical programs principally arises from

— **Conditional statements** (if, then, else)

— **Solution maps** (Solvers in applied fields: Physics, Robotics...)

— **Regularization** (Statistics or inverse problems)

```python
def myRelu(x):
    if x<=0:
        return 0
    else:
        return x
```

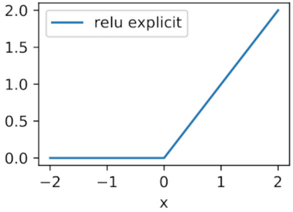


Figure: Conditionals produce nonsmoothness

Examples from Computer Science and Mathematics

- ReLU (see left)
- Sorting functions (e.g. Ranking)
- Max-Pooling (e.g. Imaging)
- Unilateral constraints (e.g. Robotics)
- Bang-bangs & shocks (e.g. Control, PDE)
- ℓ^1 norm (e.g. sparsity, statistical regularization)

A glimpse at the nonsmoothness issue

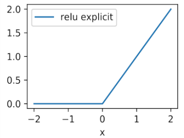
- On some problems, Autodiff is just **branch differentiation**:

$$F = \begin{cases} F_1 \text{ on } S_1 \\ F_2 \text{ on } S_2 \\ \dots \\ F_m \text{ on } S_m \end{cases} \longrightarrow \boxed{\text{Apply TensorFlow/PyTorch autodiff gives}} \longrightarrow F' = \begin{cases} F'_1 \text{ on } S_1 \\ F'_2 \text{ on } S_2 \\ \dots \\ F'_m \text{ on } S_m \end{cases}$$

- **Illustration on the famous ReLU**

Positive part: $\mathrm{relu}(t) = \max\{0, t\} \in \{0, Id\}$,

```
def myRelu(x):
    if x<=0:
        return 0
    else:
        return x
```



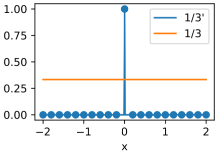
$\Rightarrow \mathrm{relu}'(x) = 0$ if $x \leq 0$, else $\mathrm{relu}'(x) = 1$

Autodiff is not so intuitive

- **The outputs of Autodiff are not everywhere meaningful.**

The derivative of a constant function may not be 0!

```
def oneThird(x):
    return relu(-x) - relu(x) + x + 1/3
```

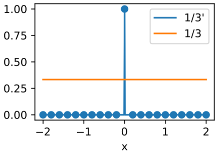


Autodiff is not so intuitive

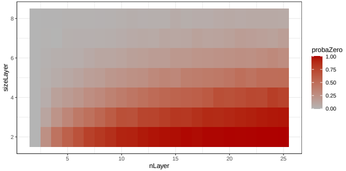
- **The outputs of Autodiff are not everywhere meaningful.**

  The derivative of a constant function may not be 0!

```
def oneThird(x):
    return relu(-x) - relu(x) + x + 1/3
```



- **When Autodiff runs nonsmoothness can significantly be activated**
  - ReLU feed-forward neural networks with increasing size and layers.
  - Redness = high proba. of activation of relu nonsmoothness during training



Contributions: a global view

Nonsmooth autodiff "works well", but why?

- **AD has a tremendous success for training**: ReLU or implicit networks, Neural ODEs, algorithm unrolling…
- Autodiff is an **uncharacterized object**
- **Lack of guarantees** for most concrete problems (partial results by Griewank-Walther 08, Kakade-Lee 18)

Contributions: a global view

Nonsmooth autodiff "works well", but why?

- **AD has a tremendous success for training**: ReLU or implicit networks, Neural ODEs, algorithm unrolling…
- Autodiff is an **uncharacterized object**
- **Lack of guarantees** for most concrete problems (partial results by Griewank-Walther 08, Kakade-Lee 18)

ANITI's contributions

- Conservative gradients: **a mathematical model and calculus for Autodiff.**
- Key prediction of the theory: **spurious outputs are extremely rare events**
- **Training guarantees for Machine Learning libraries** (`Tensorflow, Keras, Pytorch, Jax`)

An Autodiff theory: Conservative Calculus

Conservative calculus: a new differential calculus and its properties

Computer world	Mathematical world	Research papers
TensorFlow's Autodiff	Notion of conservative gradients	6 research articles
Autodiff-friendly programs	Path-differentiable functions	2 articles + 1 in progress
Nonsmoothness in programs	Whitney stratification	**NeurIPS Spotlight**
Implicit Layers	New implicit function theorem	1 article + 1 in progress
Neural ODEs / PDEs	Conservative gradient for ODEs / PDEs	1 article + 1 in progress

An Autodiff theory: Conservative Calculus

Conservative calculus: a new differential calculus and its properties

Computer world	Mathematical world	Research papers
TensorFlow's Autodiff	Notion of conservative gradients	6 research articles
Autodiff-friendly programs	Path-differentiable functions	2 articles + 1 in progress
Nonsmoothness in programs	Whitney stratification	NeurIPS Spotlight
Implicit Layers	New implicit function theorem	1 article + 1 in progress
Neural ODEs / PDEs	Conservative gradient for ODEs / PDEs	1 article + 1 in progress

An Autodiff theory: Conservative Calculus

Conservative calculus: a new differential calculus and its properties

Computer world	Mathematical world	Research papers
TensorFlow's Autodiff	Notion of conservative gradients	6 research articles
Autodiff-friendly programs	Path-differentiable functions	2 articles + 1 in progress
Nonsmoothness in programs	Whitney stratification	**NeurIPS Spotlight**
Implicit Layers	New implicit function theorem	1 article + 1 in progress
Neural ODEs / PDEs	Conservative gradient for ODEs / PDEs	1 article + 1 in progress

An Autodiff theory: Conservative Calculus

Conservative calculus: a new differential calculus and its properties

Computer world	Mathematical world	Research papers
TensorFlow's Autodiff	Notion of conservative gradients	6 research articles
Autodiff-friendly programs	Path-differentiable functions	2 articles + 1 in progress
Nonsmoothness in programs	Whitney stratification	**NeurIPS Spotlight**
Implicit Layers	New implicit function theorem	1 article + 1 in progress
Neural ODEs / PDEs	Conservative gradient for ODEs / PDEs	1 article + 1 in progress

An Autodiff theory: Conservative Calculus

Conservative calculus: a new differential calculus and its properties

Computer world	Mathematical world	Research papers
TensorFlow's Autodiff	Notion of conservative gradients	6 research articles
Autodiff-friendly programs	Path-differentiable functions	2 articles + 1 in progress
Nonsmoothness in programs	Whitney stratification	**NeurIPS Spotlight**
Implicit Layers	New implicit function theorem	1 article + 1 in progress
Neural ODEs / PDEs	Conservative gradient for ODEs / PDEs	1 article + 1 in progress

An Autodiff theory: Conservative Calculus

Conservative calculus: a new differential calculus and its properties

Computer world	Mathematical world	Research papers
TensorFlow's Autodiff	Notion of conservative gradients	6 research articles
Autodiff-friendly programs	Path-differentiable functions	2 articles + 1 in progress
Nonsmoothness in programs	Whitney stratification	**NeurIPS Spotlight**
Implicit Layers	New implicit function theorem	1 article + 1 in progress
Neural ODEs / PDEs	Conservative gradient for ODEs / PDEs	1 article + 1 in progress

Results and applications in ML

- **Training guarantees:**
  - First theoretical guarantees for training with nonsmoothness (e.g. SGD for ReLU networks).
  - Innocuousness of spurious values (e.g. spurious stationary points).
  - Risk analysis of general ML minimization problems (e.g. online Deep Learning)

Results and applications in ML

- **Training guarantees:**
  - First theoretical guarantees for training with nonsmoothness (e.g. SGD for ReLU networks).
  - Innocuousness of spurious values (e.g. spurious stationary points).
  - Risk analysis of general ML minimization problems (e.g. online Deep Learning)
- First rigorous analysis of implicit neural networks (e.g. Deep Equilibrium Networks)
- Algorithms for hyper-parameter tuning (e.g. differentiate LASSO solutions)
- New optimization algorithms: Newton like method, ridge method for adversarial learning (e.g. GANs).

Applications, Metrics, Perspectives

Results and applications in ML

- **Training guarantees:**
  - First theoretical guarantees for training with nonsmoothness (e.g. SGD for ReLU networks).
  - Innocuousness of spurious values (e.g. spurious stationary points).
  - Risk analysis of general ML minimization problems (e.g. online Deep Learning)
- First rigorous analysis of implicit neural networks (e.g. Deep Equilibrium Networks)
- Algorithms for hyper-parameter tuning (e.g. differentiate LASSO solutions)
- New optimization algorithms: Newton like method, ridge method for adversarial learning (e.g. GANs).

Metrics (for nonsmooth AD only)

10 articles (3 NeurIPS, 2 Math. Prog., JMLR), around 100 Google Scholar citations, **CNRS Bronze Medal for Pauwels**, NeurIPS spotlight, USAF grant award on the topic.

Perspectives

Arithmetic complexity, parameter optimization, robustness/sensitivity, second-order optimization

Some references

Seminal papers

[a] A mathematical model for automatic differentiation, Bolte-Pauwels, NeurIPS 2020
[b] Conservative set valued fields, automatic differentiation, stochastic gradient methods and deep learning, Bolte-Pauwels in Math. Prog. 2020
[c] An inertial Newton algorithm for deep learning, Castera, Bolte, Févotte, Pauwels, in J. of Machine Learning Research 2021

Some follow-up papers

1. Nonsmooth Implicit Differentiation for Machine Learning and Optimization, Bolte, Le, Pauwels, Silveti-Falls, NeurIPS 2021
2. Incremental Without Replacement Sampling in Nonconvex Optimization, Pauwels, in JOTA 2021
3. Numerical influence of ReLU'(0) on backpropagation, Bertoin, Bolte, Gerchinovitz, Pauwels, in NeurIPS 2021
4. The structure of conservative gradient fields, A. Lewis, T. Tian, in SIAM Opt., 2021
5. Conservative and semismooth derivatives are equivalent for semialgebraic maps D. Davis, D. Drusvyatskiy, in Set valued Analysis, 2021