AFRL-AFOSR-UK-TR-2023-0062

Landscapes of large scale problems with applications to machine learning

Bolte, Jerome
Fondat J J Laffont Tlse Sciences Eco
21, Allee De Brienne
TOULOUSE, , 31000
FR

04/17/2023
Final Technical Report

Air Force Research Laboratory
Air Force Office of Scientific Research
European Office of Aerospace Research and Development
Unit 4515 Box 14, APO AE 09421

REPORT DOCUMENTATION PAGE

PLEASE DO NOT RETURN YOUR FORM TO THE ABOVE ORGANIZATION.

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED	
20230417	Final	START DATE 20190701	END DATE 20220630

4. TITLE AND SUBTITLE
Landscapes of large scale problems with applications to machine learning

5a. CONTRACT NUMBER	5b. GRANT NUMBER	5c. PROGRAM ELEMENT NUMBER
	FA9550-19-1-7026	
5d. PROJECT NUMBER	**5e. TASK NUMBER**	**5f. WORK UNIT NUMBER**

6. AUTHOR(S)
Jerome Bolte

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Fondat J J Laffont Tlse Sciences Eco
21, Allee De Brienne
TOULOUSE 31000
FR

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)	11. SPONSOR/MONITOR'S REPORT NUMBER(S)
EOARD UNIT 4515 APO AE 09421-4515	AFRL/AFOSR IOE	AFRL-AFOSR-UK-TR-2023-0062

12. DISTRIBUTION/AVAILABILITY STATEMENT
A Distribution Unlimited: PB Public Release

13. SUPPLEMENTARY NOTES

14. ABSTRACT
Abstract. The proposal was originally centered on landscapes in deep learning, but after our first investigations, we discovered a "hole" in the variational theory of modern machine learning. There was indeed no satisfying mathematical model of autodiff –which is massively used by the AI community and even part of the applied mathematics community. The project was then naturally redirected toward the building of a proper theory of nonsmooth automatic differentiation. Great emphasis was put to have a model corresponding to real-world implementation through TensorFlow, PyTorch, or Jax. The first stones of the theory have been set with two papers at MPA and NeurIPS (SpotLight 20), and then we have studied various aspects such as nonsmooth implicit differentiation (with application to DEQs, differentiable programming), the differentiation of algorithms, and various regularity approaches. At the same time, other investigations have been led in ML and optimization, such as convex optimization, global optimization, and SOS methods.

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES
a. REPORT U	b. ABSTRACT U	c. THIS PAGE U	SAR	10

19a. NAME OF RESPONSIBLE PERSON	19b. PHONE NUMBER *(Include area code)*
CHARLTON LEWIS	3142356045

Standard Form 298 (Rev.5/2020)
Prescribed by ANSI Std. Z39.18

Final report for grant FA9550-19-1-7026, 2019-2023
"Landscapes of large scale problems with applications to machine learning"

PI : Jérôme Bolte, TSE and Edouard Pauwels, IRIT,
Université de Toulouse

Abstract. The proposal was originally centered on landscapes in deep learning, but after our first investigations, we discovered a "hole" in the variational theory of modern machine learning. There was indeed no satisfying mathematical model of autodiff –which is massively used by the AI community and even part of the applied mathematics community. The project was then naturally redirected toward the building of a proper theory of nonsmooth automatic differentiation. Great emphasis was put to have a model corresponding to real-world implementation through TensorFlow, PyTorch, or Jax. The first stones of the theory have been set with two papers at MPA and NeurIPS (SpotLight 20), and then we have studied various aspects such as nonsmooth implicit differentiation (with application to DEQs, differentiable programming), the differentiation of algorithms, and various regularity approaches. At the same time, other investigations have been led in ML and optimization, such as convex optimization, global optimization, and SOS methods.

1 Overview

The COVID-19 crisis considerably compromised the workflow for this proposal. A no-cost extension was awarded.

Research Lines Our work is mainly structured along three lines with a strong emphasis on “autodiff"

- **A nonsmooth theory of automatic differentiation**: automatic differentiation applies the chain rule algorithmically to large problems for fast computation. In deep learning, this algorithm is applied, although the cost is not smooth. To model this practice and the resulting objects, which are not gradients nor subgradients, we introduced new objects called *conservative gradients and conservative Jacobians*. The strength of our approach is to model faithfully backpropagation as implemented in TensorFlow or PyTorch.

  This approach led us to provide numerous practices such as implicit differentiation, hyperparameter differentiation, differentiation of algorithms, neural ODEs and many other aspects of differentiable programming.

- **Algorithms in deep learning and optimization**: We developed several optimization methods for the training of Deep Learning problems, as a new efficient second-order-like method for DL and a Holderian backtrack method for GANs.

- **Convergence issues in optimization**: A major problem related to DL and their losses is the behavior of the vanishing step subgradient method. We have addressed this topic through several angles: understanding of oscillations, the role of artificial critical points, and in an optimization line, the convergence/optimal complexity in the mirror descent case.

Impact: The project in a few salient facts

- Around 25 articles were produced.
- Participation in conferences, organization of seminars and research visits.
- SpotLight Prize at NeurIPS 2020 with [6].
- SpotLight Prize at ICLR 2023 with [1].
- U. Rothblum OR prize was awarded to J. Bolte, S. Sabach, M. Teboulle, Y. Vaisbourd for “First Order Methods beyond Convexity and Lipschitz Gradient Continuity with Applications to Quadratic Inverse Problems" published in SIAM
- J. Bolte was awarded a chair in AI within the Artificial and Natural Intelligence Toulouse Institute (ANITI) within Villani’s national AI plan
- Growing synergy with TSE Digital center
  `https://www.tse-fr.eu/digital`

– Growing synergy with ANITI
https://aniti.univ-toulouse.fr/

– E. Pauwels was awarded the prestigious “Médaille de Bronze" du CNRS.
https://youtu.be/I4Ui2OKBsBE

– J. Bolte’s talk at One World Optimization Seminar
https://www.youtube.com/watch?v=JNRWYS5GGZc

– E. Pauwel’s talk at One World Optimization Seminar
https://youtu.be/g1BZ47pLeGc

– Recruitments using USAF Grant. We recruited two post-doctoral fellows, C. Combettes (Georgia Tech, PhD with Pokutta), A. Silveti-Falls (ENSICAEN PhD with J. Fadili)

Involved researchers

– Jérôme Bolte, Full Professor, PI, TSE, ANITI, Université Toulouse Capitole.

– Edouard Pauwels, Assistant Professor, co-PI, IRIT, ANITI and University Toulouse 3.

– Radu Dragomir, University Toulouse 3 & ENS Paris, now EPFL.

– Rodolfo Rios-Zertuche, ANITI and LAAS.

– Antonio Silveti Falls, USAF Post-Doc now associate Prof. at Centrale Supelec.

– Cyrille Combettes, USAF Post-Doc, now working at CFM.

– Tam Le, ANITI Ph.D. student.

– Bruce Suter, USAF.

– Erin Tripp, USAF.

2 Research report

The initial goal of this proposal was to study landscapes of deep learning problems and to understand/design new efficient algorithms. As often in science, some objectives may hide other delicate issues that can be fundamental. In the course of our research, we discovered that the variational models used to study “stochastic gradient methods" (SGD, ADAGRAD, ADAM, RMS Prop, etc...) *do not reflect the numerical/theoretical reality*. In particular, the models for the backpropagation algorithm that computes “gradients" in deep learning were restrictively considered as a subgradient or even gradients. This widespread approach does not reflect the computational reality of modern libraries such as TensorFlow or PyTorch. This is why we put a lot of energy into providing a proper

model for studying the training of deep learning losses. We called the objects we introduced *conservative fields*; they allowed us to provide the first rigorous convergence proof of SGD with backpropagation and mini-batches in deep learning.

It is also of importance to mention that our results are, to our knowledge, the first ones to prove that *backpropagation/automatic differentiation outputs a gradient almost everywhere* (see [3]) – which is not a trivial result as it is related to Whitney stratifiability of semialgebraic functions.

We now provide an overview of the results that were obtained for this proposal.

2.1 Automatic differentiation and conservative fields

Conservative fields [Math. Prog. 21] Modern problems in AI or numerical analysis require nonsmooth approaches with a flexible calculus. We introduced generalized derivatives called conservative fields, for which we developed calculus and provided representation formulas. We studied locally Lipschitz continuous functions having a conservative field and called them *path differentiable*. A remarkable fact is that most common functions are path differentiable, as, for instance, convex, concave, Clarke regular and any semialgebraic Lipschitz continuous functions. Our model provides variational formulas for nonsmooth automatic differentiation oracles; for instance, the famous backpropagation algorithm in deep learning produces naturally conservative gradients. With this model, we proved for the first time the convergence of SGD with backpropagation and mini-batches for nonsmooth nonconvex problems.

A simple model for automatic differentiation [NeurIPS SpotLight 20] In this work, we have specifically articulated the relationships between the differentiation of *programs* as implemented in practice and the differentiation of nonsmooth functions. To this end, we have provided a simple but very wide class of functions for which differentiation along branches produces conservative gradients. In that work, we also evidenced the issue of artificial critical points created by algorithmic differentiation and showed how usual methods avoid these points with probability one.

Numerical influence of ReLU'(0) on backpropagation [NeurIPS 21] As alluded above, the choice of ReLU'(0) in $[0, 1]$ for a neural network has, *in theory*, a negligible influence both on backpropagation and training. Yet, in the real world, 32 bits default precision combined with the size of deep learning problems makes it a hyperparameter of training methods. We investigated the importance of the value of ReLU'(0) for several precision levels (16, 32, 64 bits) on various networks (fully connected, VGG, ResNet) and datasets (MNIST, CIFAR10, SVHN, ImageNet). We observed considerable variations of backpropagation outputs, which occur around half of the time in 32 bits precision. The effect disappears with double precision, while it is systematic at 16 bits. For vanilla SGD training, the choice ReLU'$(0) = 0$ seems to be the most efficient. We also evidenced that reconditioning approaches as batch-norm or ADAM tend to buffer the influence of ReLU'(0)'s value. One of the conclusions of this work is that algorithmic differentiation of nonsmooth prob-

lems potentially hides parameters that could be tuned advantageously. Research along this line is being pursued.

Nonsmooth Implicit Differentiation for Machine Learning and Optimization [NeurIPS 21 & submission SIAM] This work shows that one of the strengths of conservative calculus is to provide a nonsmooth world in which formal differentiation is possible. Specifically, we focused on implicit differentiation, which is essential to train complex learning architectures (displaying implicit layers, and bilevel programs). Our result applies to most practical problems as definable problems, provided that a nonsmooth form of the classical invertibility condition is fulfilled. This approach allows for formal subdifferentiation as in the smooth case: it suffices to replace derivatives by Clarke Jacobians in the usual differentiation formulas. This calculus is entirely compatible with algorithmic differentiation (e.g., backpropagation). We provided several applications, such as training deep equilibrium networks, training neural nets with conic optimization layers, or hyperparameter-tuning.

Differentiation of algorithms [NeurIPS 22] Along a related line, we studied the connection between nonsmooth implicit differentiation and forward differentiation of fixed point algorithms and show linear convergence of piggyback derivatives of several well-known iterative optimization algorithms. This is illustrated and commented in Figure 1.

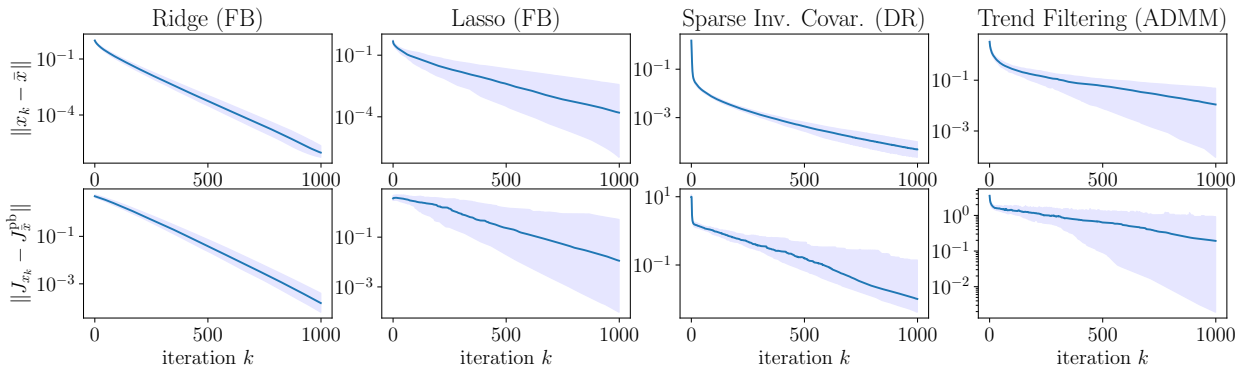

Figure 1: Illustration of the linear convergence of proximal splitting methods on four different convex optimization problems. (First line) Distance of the iterates to the fixed point. (Second line) Distance of the piggyback Jacobians to the Jacobian of the fixed point. The acronyms are FB for Forward-Backward, DR for Douglas-Rachford and ADMM for Alternating Direction Method of Multipliers. In all cases, despite nonsmoothness, piggyback Jacobians converge, illustrating Corollary 2. Blue lines represent the median of 100 repetitions with random data, and the blue shaded area represents the first and last deciles.

On the complexity of nonsmooth automatic differentiation [ICLR 2022 Spotlight] Since the famous Baur-Strassen theorem backprop is known to be very fast for rational func-

tions or smooth functions –this is known under the term of “the cheap gradient principle". But what happens for “real-world autodiff", which applies to many nonsmooth objects as relu networks, algorithms, or even solvers? Using the notion of conservative gradient, we proved that the complexity overhead of the backward mode turns out to be independent of the dimension when using programs with locally Lipschitz semi-algebraic or definable elementary functions. This extends considerably the Baur-Strassen’s smooth cheap gradient principle. We illustrate our results by establishing fast backpropagation results of conservative gradients through feedforward neural networks with standard activation and loss functions. Nonsmooth backpropagation’s cheapness contrasts with concurrent forward approaches, which have, to this day, dimensional-dependent worst-case overhead estimates. We also provided further results suggesting the superiority of backward propagation of conservative gradients. For instance, we showed that finding two subgradients in the Clarke subdifferential of a function is a NP-hard problem.

2.2 Optimization methods for learning

This part is related to the initial objectives of this proposal before we realized the importance of a nonsmooth calculus for autodiff.

Our goal was here to understand training methods in deep learning (as nonsmooth SGD) and to design new methods.

The vanishing subgradient step method [JEMS 22, Math. OR 22] We developed a new approach for analyzing the dynamics of first-order algorithms with vanishing step size. No restrictions whatsoever are made on the steps apart from the fact that they tend to zero (which is necessary to approach minimizing points). Addressing some long-standing open questions in the area, we apply our methodology to obtain oscillation compensation, slow-down, and criticality results of several kinds for the model case of the small-step subgradient method. We also evidenced the natural role of the essential accumulation set that highlights the points near which the process spends most of its time recurrently while ignoring sporadic escapades.

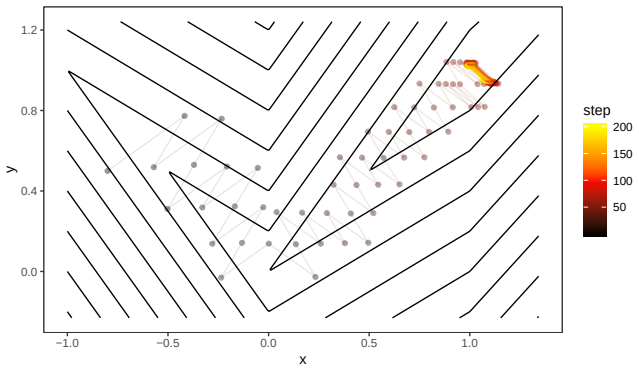

Figure 2: Contour plot of a Lipschitz function with a vanishing subgradient sequence. The color reflects the iteration count. The sequence converges to the unique global minimum but is constantly oscillating.

We also gave two examples that show that the previous results on the long-term dynamics of the subgradient method are very sharp, hence showing the power of the methodology developed in that paper. Our examples show that in the context of path differentiable Lipschitz objective functions, it would be impossible to get better convergence results, and the oscillation compensation and slowdown results, in general, are only true near the essential accumulation set, see [23].

INNA a new algorithm for Deep Learning [JMLR 20] We devised a learning algorithm for possibly nonsmooth deep neural networks featuring inertia and Newtonian directional intelligence only using a back-propagation oracle. Our algorithm, called INNA, has an appealing mechanical interpretation modeled on a second-order equation of the form

$$\theta(t) + \alpha\theta(t)' + \beta\nabla^2 J(\theta(t))\theta(t)' + \nabla J(\theta(t)) = 0, \quad t \geq 0$$

making the role of its two hyperparameters transparent. An elementary phase space lifting allows both for its implementation and its theoretical study under very general assumptions. We handled, in particular, stochastic versions of our method (which encompasses usual mini-batch approaches) for nonsmooth activation functions (such as ReLU). Our algorithm shows high efficiency and reaches the state of the art on image classification problems.

Second-order step-size tuning of SGD for non-convex optimization [Neural Proc. Lett. 21] In view of a direct and simple improvement of vanilla SGD, this paper presents a fine-tuning of its step sizes in the mini-batch case. To do so, one estimates curvature based on a local quadratic model and using only noisy gradient approximations. One obtains a new stochastic first-order method (Step-Tuned SGD), enhanced by second-order information, which can be seen as a stochastic version of the classical Barzilai-Borwein method. Our theoretical results ensure almost sure convergence to the critical set and we provide convergence rates. Experiments on deep residual network training illustrate the favorable properties of our approach. For such networks, we observe, during training, both a sudden drop of the loss and an improvement of test accuracy at medium stages, yielding better results than SGD, RMSprop, or ADAM

Without replacement sampling in stochastic optimization [JOTA 2021] In a large sum context, the classical analysis of SGD relies on "with replacement" sampling. However in practice, "with replacement" sampling is barely used and machine learning users prefer to use random reshufling (evaluating each input data exactly once at each epoch). We describe a general algorithmic recipe which allows to model "without replacement" sampling, which is much closer to strategies used in practice. In the smooth setting, we analyse its complexity and provide a convergence rate of order $O(k^{-1/3})$ which is asymptotically better than best known rates for SGD. In the nonsmooth setting we prove that this without replacement sampling satisfies the same qualitative properties as with replacement SGD in the context of conservative gradients.

2.3 Miscellaneous

We also published papers or preliminary reports on several topics in optimization:

- Convex results
  - A paper with C^k smooth convex counterexamples to old standing convergence questions: we showed that block-coordinate, steepest descent with exact search or Bregman descent methods do not generally converge. Other failures of various desirable features can be established: directional convergence of Cauchy's gradient curves, convergence of Newton's flow, finite length of Tikhonov path, convergence of central paths, or smooth Kurdyka-Łojasiewicz inequality.
  - On mirror descent: a result on the optimal complexity of mirror descent within the class of relatively smooth functions with an arbitrary kernel.
- Robustness in DL: through semi-algebraic optimization, we leveraged the semialgebraic representation of the ReLU function to describe the input-output relation of monotone Deep Equilibrium networks (MONDEQ) using a polynomial system. This constitutes the first general approach for the certification of implicit networks of this type.
- Approximation: The Christoffel Darboux kernel approach was developped in the context of support estimation from a statistical sample, with high probability guaranties, and in the context of the moment relaxation to estimate solutions of partial differential equations with approximation guaranties.
- Global optimization: Using jointly geometric and stochastic reformulations of nonconvex problems and exploiting a Monge-Kantorovich gradient system formulation with vanishing forces, we formally extend the simulated annealing method to a wide class of global optimization methods. Due to an inbuilt combination of a gradient-like strategy and particle interactions, we called them *swarm gradient dynamics*.

[1] J. Bolte, R. Boustany, E., Pauwels, B. Pesquet-Popescu, (2023, May). On the complexity of nonsmooth automatic differentiation. In International Conference on Learning Representations.

[2] J. Bolte, L. Glaudin, E. Pauwels, M. Serrurier. A Holderian backtracking method for min-max and min-min problems. To appear in OJMO.

[3] J. Bolte, T. Le, E. Pauwels, A. Silveti-Falls, Nonsmooth Implicit Differentiation for Machine Learning and Optimization, to appear in SIAM Optimization.

[4] J. Bolte, E. Pauwels, *Conservative set-valued fields, automatic differentiation, stochastic gradient method and deep learning*. Mathematical Programming 188 (2021): 19-51.

[5] J. Bolte, E. Pauwels, *Curiosities and counterexamples in smooth convex optimization*, Math. Prog.

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

[7] J. Bolte, E. Pauwels, R. Rios-Zertuche, Long-term dynamics of the subgradient method for Lipschitz path differentiable functions. J. European Mathematical Society, 2022

[8] Bolte, Jérôme, Edouard Pauwels, and Samuel Vaiter. Automatic differentiation of nonsmooth iterative algorithms. NeurIPS (2022).

[9] C. Castera, J. Bolte, E. Pauwels, C. Févotte, *An Inertial Newton Algorithm for Deep Learning*, The Journal of Machine Learning Research, 22(1), 5977-6007.

[10] C. Castera, J. Bolte, C. Févotte, E.Pauwels. Second-order step-size tuning of SGD for non-convex optimization, Neural Processing Letters, 2022, vol. 54, no 3, p. 1727–1752.

[11] D. Bertoin, J. Bolte, S. Gerchinovitz, E. Pauwels, Numerical influence of ReLU'(0) on backpropagation, NeurIPS 21

[12] T. Chen, J.B. Lasserre, V. Magron, E. Pauwels. Semialgebraic optimization for Lipschitz constants of ReLU networks. NeurIPS 2020.

[13] T. Chen, J.B. Lasserre, V. Magron, E. Pauwels. A sublevel moment-sos hierarchy for polynomial optimization, Computational Optimization and Applications, 2022, vol. 81, no. 1: 31-66.

[14] R. Dragomir, A d'Aspremont, J Bolte, *Quartic First-Order Methods for Low Rank Minimization*, Journal of Optimization Theory and Applications, 2021, vol. 189, p. 341-363.

[15] RA. Dragomir, M. Even, H. Hendrikx, *Fast Stochastic Bregman Gradient Methods: Sharp Analysis and Variance Reduction*, Proceedings of the 38th International Conference on Machine Learning, PMLR 139, pp. 2815-2825, 2021.

[16] R. Dragomir, A. Taylor, A. d'Aspremont, J. Bolte, *Optimal Complexity and Certification of Bregman First-Order Methods*, Mathematical Programming, 2021, p. 1-43.

[17] Marx, Swann, et al. Semi-algebraic approximation using Christoffel–Darboux kernel. Constructive Approximation (2021): 1-39.

[18] Marx, Swann, and Edouard Pauwels. Path differentiability of ODE flows. Journal of Differential Equations 338 (2022): 321-351.

[19] Fabian, M., Hiriart-Urruty, J. B., Pauwels, E. (2022). On the Generalized Jacobian of the Inverse of a Lipschitzian Mapping. Set-Valued and Variational Analysis, 1-9.

[20] E. Pauwels. Incremental without replacement sampling in nonconvex optimization. Journal of Optimization Theory and Applications 190(1) pp 274-299, 2021.

[21] E. Pauwels, M. Putinar, J.B. Lasserre. Data analysis from empirical moments and the Christoffel function. FoCM (2020)

[22] Pauwels, Edouard, and Samuel Vaiter. The derivatives of Sinkhorn-Knopp converge. To appear in SIAM journal on optimization (2023).

[23] R. Rios-Zertuche, Examples of pathological dynamics of the subgradient method for Lipschitz path-differentiable functions. Mathematics of Operations Research, 2022, vol. 47, no 4, p. 3184-3206.

[24] C. Traoré, E. Pauwels, Sequential convergence of AdaGrad algorithm for smooth convex optimization, Operations Research Letters 49 (4), 452-458 2021

[25] M.T. Vu, F. Bachoc, E. Pauwels. Rate of convergence for geometric inference based on the empirical Christoffel function. ESAIM: Probability and Statistics, 26, 171-207 (2019)