AAPG2024	SIMPLES		JCJC
Coordinated by :	Antonio SILVETI-FALLS CENTRE FOR VISUAL COMPUTING (CVN) CENTRALESUPÉLEC	48 months	262.5 k€
Domaine: 2, Sciences du Numériques - Axe: E.2 Intelligence artificielle et science des données			

New Methods for Nonsmooth Stochastic Bilevel Optimization

Solutions Implicites et Méthodes non-lisses Pour L'Optimisation En Bilevel Stochastique

1 Pre-proposal's context, positioning and objectives

1.1 Project's objectives, research hypothesis, state of the art

Stochastic bilevel optimization is a field of research within optimization in which we have one stochastic optimization problem (the inner problem) nested inside of another one (the outer problem). This special structure makes it particularly useful for modeling hierarchical problems with uncertainty, where the solution at one level directly influences the parameters or constraints of the solution at the subsequent level. Such hierarchical problems are prevalent in modern machine learning; without being exhaustive these include hyperparameter optimization [17], implicit layers in neural networks [4, 2], adversarial examples [35], and dictionary learning [18]. In all of these examples, we want to select the best model according to some stochastic criteria (the outer problem) given that the models to choose from are themselves solutions to some stochastic optimization problem (the inner problem). More succinctly, a bilevel optimization problem aims to answer the question: of all the solutions to the inner problem, which is the best according to the outer problem?

Many algorithms and analyses have been developed in the name of solving bilevel formulations of machine learning problems [25, 16, 7, 22]. However, the majority of the algorithms currently used to solve these types of problems are bottlenecked by the fact that they are derived from smooth optimization techniques, which require continuous differentiability assumptions, if not explicit C^p smoothness assumptions for $p \geq 2$, that are unrealistic. For instance, in the overwhelming majority of problems involving deep neural networks (e.g., implicit layers, adversarial examples, meta-learning, neural architecture search, etc), the ReLU activation function is used despite the fact that it is not differentiable (it contains a kink at 0), which poses a major hurdle to theoretical analysis. Similarly, in hyperparameter optimization, the use of nonsmooth regularizers like the ℓ^1 norm, the group-LASSO norm, the nuclear norm, or many others to induce sparsity, low-rank, or other inductive biases yield nonsmooth solution mappings that cannot be differentiated.

In spite of this, the methods inspired by smooth analyses tend to work surprisingly well in practice, even on problems that fall completely outside of the theory employed to develop them. This points to a significant gap in the analysis and modeling used to create these methods. Indeed, many successful implementations deployed in practice try to construct quantities to act as "Jacobians" or "Hessians", using so-called automatic differentiation and backpropagation, despite the fact that neither the Jacobian nor the Hessian exist for the problems they are applied to.

The aim of this project is to bridge this gap between the theory and the practice, leveraging a recent construction in nonsmooth analysis called the conservative field [10] that faithfully models what is computed by automatic differentiation and backpropagation, even when they are applied to nonsmooth functions. To accomplish this we will use a recently developed nonsmooth implicit function theorem [9] and a nonsmooth envelope theorem [24], in conjunction with new efforts to expand the theory of conservative calculus. This will result in new theorems that encompass problems that were previously outside the boundaries of the theory, e.g., a constrained envelope theorem and an implicit function theory without an invertibility assumption. With this we can begin to develop methods that not only work well and are implementable in practice but also have rigorous mathematical guarantees backing them. Additional aspects, such as numerical implementation and application of these developments using state of the art deep learning frameworks (PyTorch, TensorFlow, JAX) to problems in data science and computer vision are also within the scope of this project and will be explored.

1.2 Methodology, innovation and originality

The first difficulty is to successfully combine nonsmoothness with stochasticity. Typically, the mathematical theory used to analyze stochasticity for these bilevel optimization problems is too brittle to be applied to nonsmooth problems. In fact many works on this topic resort to assuming a high degree of regularity (C^3 smoothness or more) on the functions involved, contrary to what is observed in practice. More specifically, the current state of the art analyses require that one can interchange the gradient operator with the expectation, which fails for the usual generalized gradients (subdifferentials) available for nonsmooth problems. For instance, the expectation of the Clarke subgradient of a function is not a Clarke subgradient of the expectation of that function.

The second difficulty is ensuring that the methods are implementable and effective on practical problems, without sacrificing the rigorous theoretical guarantees. Although some methods are known to perform well empirically, there is no mathematical theory that guarantees their performance on stochastic, nonsmooth problems. Even in the deterministic setting, it is virtually impossible to compute an element of the generalized derivatives associated to a nonsmooth function when the dimension of the problem is high or when the function is the composition of many simple (although nonsmooth) functions because they do not admit a chain rule. As deep learning is fundamentally based on optimizing functions defined by composing many layers with nonsmooth activation functions, it is immediately outside the realm of these generalized derivatives. Even worse, most problems in modern machine learning take place in high-dimensional spaces and use stochasticity in their modeling, compounding the challenges associated to analyzing them mathematically.

A crucial aspect of the approach we plan to develop is that we are able to tackle both of these difficulties simultaneously. The conservative calculus we will base our efforts on is robust to these difficulties in the following ways:

- it admits a chain rule for nonsmooth functions [10] (even for functions which are implicitly defined [9]) making it applicable to a broad class of functions (including semialgebraic functions or compositions of many functions),
- it is compatible with stochastic expectations in the ways necessary for analyzing stochastic optimization algorithms, i.e., under mild assumptions met in practice we can commute the expectation with generalized differentiation [8, 20],
- it is completely implementable using well-known deep learning libraries (JAX, TensorFlow, PyTorch) and compatible with automatic differentiation (and thus also backpropagation).

◦ **Task 1 (Ill-posed implicit functions)** is to address the case where the implicit function to be differentiated is not well defined. The conservative implicit function proved in [9] requires an invertibility condition on the conservative Jacobian of the defining equation to ensure that there is a (locally) unique solution. However, many problems in theory and practice with unique solutions violate this condition or, in the worst case, have multiple solutions. We will therefore extend the results of [9] and [24] to address the question of how to compute the conservative gradient when there are multiple solutions or when there is a unique solution but the invertibility condition fails. This will require generalizing the conservative envelope theorem of [24] to encompass constrained problems with multiple solutions and further analysis of the implicit function theorem in [9] to the case where the invertibility condition does not hold.

◦ **Task 2 (Algorithm design)** is to integrate the findings of Task 1 in a way that is compatible with stochasticity and to furthermore develop implementable algorithms with convergence guarantees based on the improved analyses. It was shown in [8] that conservative gradients are compatible with expectations in the sense that the expectation of a conservative gradient is the conservative gradient of the expectation. However, further investigations involving optimization algorithms based on this result and in combination with [9, 24] are yet to be explored. Consequently, we plan to initiate this study, building on the work done in [20] to develop fast algorithms applicable to the stochastic problems.

◦ **Task 3 (Applications to machine learning)** is dedicated to the practical validation and implementation of the resulting computational methods on realistic large-scale problems coming from machine learning, deep learning, signal processing, and inverse problems. Because conservative gradients admit a chain rule, they are computable using standard automatic differentiation libraries popular in the deep learning communities. We will therefore be able to apply our results to practical problems in machine learning with mathematical guarantees. We intend to focus on the design of neural networks with implicitly defined layers, stochastic layers (such as those used for Bayesian variational autoencoders), differentiable architecture search, and their application to computer vision. This task can be handled concurrently with Task 1 and Task 2, using the theoretical advances in the other tasks as a guide for how to address practical problems in a principled way with mathematical guarantees.

2 Organization and implementation of the project

2.1 Scientific coordinator and their team

Antonio Silveti-Falls will be the coordinator of the SIMPLES project. He is an assistant professor at the Center for Visual Computing (CVN) lab and the mathematics department at CentraleSupélec, Graduate School of Engineering and Systems of University Paris-Saclay. Besides this he is part of the INRIA team OPIS, and a member of the Fédération de Mathématiques de CentraleSupélec. His research interests, cultivated during his studies and collaborations with esteemed leaders in optimization like Jalal Fadili, Gabriel Peyré, Edouard Pauwels, and Jérôme Bolte, straddle the frontiers of nonsmooth analysis [31, 11], stochastic optimization [33, 32], and the theory of deep learning using conservative calculus [9]. His work has been published in prestigious conferences and journals (NeurIPS, SIOPT) in addition to the reviewing he does (NeurIPS highlighted reviewer, ICLR highlighted reviewer, AISTATS, ICML, BMVC, SIOPT, MathProg, SIMODS, JMLR). He has also been a speaker at over 20 conferences and seminars, winning the best student paper award at SPARS2019.

He plans to initiate a new research regime based on the notion of conservative calculus that will fruitfully marry nonsmooth analysis and algorithms for practical applications. Towards this goal, Antonio Silveti-Falls (with an 80% implication) will create a balanced research team, consisting of expert researchers (described below) from multiple adjacent domains, that is capable of advancing not only the theory of nonsmooth analysis but also novel applications to computer vision, machine learning, and inverse problems. This unique combination of international talents will provide the future students and postdoctoral researchers recruited with the necessary foundations and mentorship to perform state of the art research with publications expected in high impact conferences (NeurIPS, ICLR, ICML, CVPR, ICCV, ECCV, SIOPT, SIMODS, JMLR).

Within CVN

Hugues Talbot is full professor at CentraleSupélec. He is an expert in image processing and computer vision [3, 15, 21], regularly producing open source software libraries in this domain [34]. He will play a major role in the application of the theoretical advances made in the project to real-world problems and data. He will invest 20% of his time to this project. Jean-Christophe Pesquet is a distinguished professor at CentraleSupélec and the director of the CVN lab. He is a world-renowned expert in not only optimization [12, 13, 14] and mathematical imaging [1, 6, 30, 19] but also, together with Hugues Talbot, in applications to medical imaging. Together, they routinely collaborate with several hospitals and medical professionals in the greater Paris region. He agrees to invest 20% of his time in this project.

Non CVN

Peter Ochs is a full professor at the University of Saarland. He is an expert in bilevel optimization [28, 29, 23] and nonsmooth optimization [26, 27, 5]. He has contributed numerous algorithms and open source software for practical applications in data science, inverse problems, and computer vision. He plans to invest 30% of his time in this project.

2.2 ORGANIZATION TO REACH THE OBJECTIVES

Combining the skills and expertise of the members of this team will provide a balance between nonsmooth analysis and generalized differentiation (Silveti-Falls and Pesquet), continuous bilevel optimization (Silveti-Falls and Ochs), and implementation/programming/applications to data science and computer vision (Talbot, Pesquet, and Ochs).

The three senior members of the project already supervise PhD students and, with the funding of the project, future recruiting efforts will be facilitated.

◦ **Task 1 (Ill-posed implicit functions)** is the at the core of the theoretical novelty of this proposal. Silveti-Falls will lead the parts related to extending the conservative calculus to account for ill-posed implicit functions, e.g., relaxing the hypotheses of the implicit function theorem he proved in [9] and extending the conservative envelope theorem of [24] to account for linear constraints. Combining these expected results with those of [8] to allow for stochasticity in the bilevel optimization problems treated will also fall under the purview of Silveti-Falls.

◦ **Task 2 (Algorithm design)** is devoted to constructing fast, implementable algorithms for stochastic bilevel optimization problems without smoothness assumptions. This will aggregate the backgrounds of all members of the team and build on the fundamental advances of Task 1. More precisely, the expertise of Pesquet and Ochs, having already designed and analyzed multiple state of the art first-order nonsmooth optimization algorithms, will be leveraged in combination with the expertise of Silveti-Falls in stochastic optimization and Talbot in global optimization to create novel bilevel optimization algorithms with rigorous convergence guarantees.

◦ **Task 3 (Applications to machine learning)** represents the practical ambitions of the project and will benefit primarily from the experience and domain expertise of Talbot and Pesquet. We envision applying the algorithms we develop and analyze to the following concrete applications: image segmentation for medical imaging, Bayesian variational autoencoders, sparse dictionary learning, and plug and play neural networks for computer vision. Computational resources of the CVN lab will be utilized to train and test the algorithms and models on these applications. Moreover, we plan to release all code and software developed as open source, available through GitHub, to allow it to be used throughout the wider machine learning and computer vision communities.

2.3 DEVELOPMENT OF RESEARCH TEAM - NEEDED RESOURCES FOR OBJECTIVES

Total 262.5 k€

Personnel 187 200 €
Travel 20 000€
Equipment 5 000€
Fees 33 242€

The requested funding is for 48 months. It will cover the costs of developing a small team consisting of a PhD student, several masters students, and a post-doctoral researcher. It will also encompass travel expenses for collaborative efforts within the team and for international conferences.

The bulk of the funding will go to hiring a PhD student (120k€) and a 1-year postdoctoral researcher (60k€) at CentraleSupélec. Additionally, two master student internships in CentraleSupélec laboratories (7.2k€ $\sim$ 6 months each at 600€) will be hired. It also includes travel costs for the members corresponding to travels between Paris and Saarland, team meetings, international conferences and missions of the students and postdoctoral researcher. The total resources include equipment of researchers in the form of laptops and monitors for the PhD student and postdoctoral researcher.

REFERENCES

[1] F. Abboud, M. Stamm, E. Chouzenoux, **J.-C. Pesquet**, and **H. Talbot**. Distributed algorithms for scalable proximity operator computation and application to video denoising. *Digital Signal Processing*, 2022.

[2] B. Amos and J.Z. Kolter. Optnet: Differentiable optimization as a layer in neural networks. In *ICML*. PMLR, 2017.

[3] B. Appleton and **Talbot, H.** Globally minimal surfaces by continuous maximal flows. *IEEE TPAMI*, 2005.

[4] S. Bai, J.Z. Kolter, and V. Koltun. Deep equilibrium models. *NeurIPS*, 2019.

[5] S. Becker, J. Fadili, and **P. Ochs**. On quasi-Newton forward-backward splitting: Proximal calculus and convergence. *SIOPT*, 2019.

[6] C. Bertocchi, E. Chouzenoux, M.-C. Corbineau, **J.-C. Pesquet**, and M. Prato. Deep unfolding of a proximal interior point method for image restoration. *Inverse Problems*, 2020.

[7] Q. Bertrand, Q. Klopfenstein, M. Massias, M. Blondel, S. Vaiter, A. Gramfort, and J. Salmon. Implicit differentiation for fast hyperparameter selection in non-smooth convex learning. *JMLR*, 2022.

[8] J. Bolte, T. Le, and E. Pauwels. Subgradient sampling for nonsmooth nonconvex minimization. *arXiv preprint arXiv:2202.13744*, 2022.

[9] J. Bolte, T. Le, E. Pauwels, and **A. Silveti-Falls**. Nonsmooth implicit differentiation for machine-learning and optimization. *NeurIPS*, 2021.

[10] J. Bolte and E. Pauwels. Conservative set valued fields, automatic differentiation, stochastic gradient methods and deep learning. *Mathematical Programming*, 2021.

[11] J. Bolte, E. Pauwels, and **A. Silveti-Falls**. Differentiating nonsmooth solutions to parametric monotone inclusion problems. *To appear in SIOPT*, 2023.

[12] E. Chouzenoux, **J.-C. Pesquet**, and A. Repetti. Variable metric forward–backward algorithm for minimizing the sum of a differentiable function and a convex function. *JOTA*, 2014.

[13] P.L. Combettes and **J.-C. Pesquet**. Primal-dual splitting algorithm for solving inclusions with mixtures of composite, Lipschitzian, and parallel-sum type monotone operators. *Set-Valued and Variational Analysis*, 2012.

[14] P.L. Combettes and **J.-C. Pesquet**. Stochastic quasi-Fejér block-coordinate fixed point iterations with random sweeping. *SIOPT*, 2015.

[15] Camille Couprie, Leo Grady, Laurent Najman, and **Talbot, H.** Power watershed: A unifying graph-based optimization framework. *IEEE TPAMI*, 2010.

[16] M. Dagréou, P. Ablin, S. Vaiter, and T. Moreau. A framework for bilevel optimization that enables stochastic and global variance reduction algorithms. *NeurIPS*, 2022.

[17] L. Franceschi, P. Frasconi, S. Salzo, R. Grazzi, and M. Pontil. Bilevel programming for hyperparameter optimization and meta-learning. In *ICML*. PMLR, 2018.

[18] H. Ghanem, N. Keriven, J. Salmon, and S. Vaiter. Supervised graph learning with bilevel optimization. In *18th Intl. Workshop on Mining and Learning with Graphs*, 2022.

[19] N. Lassau, S. Ammari, E. Chouzenoux, H. Gortais, P. Herent, M. Devilder, S. Soliman, O. Meyrignac, M.-P. Talabard, J.-P. Lamarque, et al. Integrating deep learning CT-scan model, biological and clinical variables to predict severity of COVID-19 patients. *Nature communications*, 2021.

[20] T. Le. Nonsmooth nonconvex stochastic heavy ball. *arXiv preprint arXiv:2304.13328*, 2023.

[21] L. Le Bescond, M. Lerousseau, I. Garberis, F. André, S. Christodoulidis, M. Vakalopoulou, and **Talbot, H.** Unsupervised Nuclei Segmentation using Spatial Organization Priors. In *MICCAI*, 2022.

[22] B. Liu, M. Ye, S. Wright, P. Stone, and Q. Liu. Bome! bilevel optimization made easy: A simple first-order approach. *NeurIPS*, 2022.

[23] S. Mehmood and **P. Ochs**. Automatic differentiation of some first-order methods in parametric optimization. In *AISTATS*. PMLR, 2020.

[24] E. Pauwels. Conservative parametric optimality and the ridge method for tame min-max problems. *Set-Valued and Variational Analysis*, 2023.

[25] Z. Ramzi, F. Mannel, S. Bai, J.-L. Starck, P. Ciuciu, and T. Moreau. SHINE: SHaring the INverse Estimate from the forward pass for bi-level optimization and implicit models. *ICLR*, 2022.

[26] **P. Ochs**, Y. Chen, T. Brox, and T. Pock. iPiano: Inertial proximal algorithm for nonconvex optimization. *SIAM Imaging*, 2014.

[27] **P. Ochs**, J. Fadili, and T. Brox. Non-smooth non-convex Bregman minimization: Unification and new algorithms. *JOTA*, 2019.

[28] **P. Ochs**, R. Ranftl, T. Brox, and T. Pock. Bilevel optimization with nonsmooth lower level problems. In *SSVM*. Springer, 2015.

[29] **P. Ochs**, R. Ranftl, T. Brox, and T. Pock. Techniques for gradient-based bilevel optimization with non-smooth lower level problems. *Journal of Mathematical Imaging and Vision*, 2016.

[30] **J.-C. Pesquet**, A. Repetti, M. Terris, and Y. Wiaux. Learning maximally monotone operators for image recovery. *SIAM Imaging*, 2021.

[31] **A. Silveti-Falls**, C. Molinari, and J. Fadili. Generalized conditional gradient with augmented lagrangian for composite minimization. *SIOPT*, 2020.

[32] **A. Silveti-Falls**, C. Molinari, and J. Fadili. A stochastic bregman primal-dual splitting algorithm for composite optimization. *arXiv preprint arXiv:2112.11928*, 2021.

[33] **A. Silveti-Falls**, C. Molinari, and J. Fadili. Inexact and stochastic generalized conditional gradient with augmented lagrangian and proximal step. *Journal of Nonsmooth Analysis and Optimization*, 2021.

[34] **Talbot, H.** Open-source software for image analysis. `https://hugues-talbot.github.io/portfolio/`, 2023.

[35] Y. Zhang, G. Zhang, P. Khanduri, M. Hong, S. Chang, and S. Liu. Revisiting and advancing fast adversarial training through the lens of bi-level optimization. In *ICML*. PMLR, 2022.