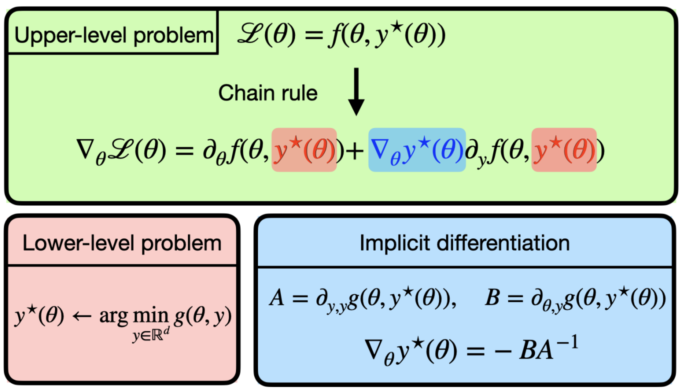
A learning theory for over-parametrized bilevel optimization

Supervision:

- Michael N. Arbel, Thoth, Inria (michael.arbel@inria.fr)
- Edouard Pauwels, Toulouse School of Economics (edouard.pauwels@tse-fr.eu)
- Samuel Vaiter, CNRS, Université Côte d'Azur (samuel.vaiter@cnrs.fr)

Context Bilevel optimization is a class of methods for solving optimization problems that have a hierarchical structure. These problems typically require optimizing two interdependent objectives: a *lower-level* objective g whose optimal solution is provided to an *upper-level* objective f. The hierarchical structure arises by taking into account the dependence of the *lower-level* solution on the *upper-level* variable (see figure below). These methods are increasingly recognized as a promising approach for solving a multitude of machine learning problems such as hyper-parameter optimization, meta-learning, meta-reinforcement learning [1] and metric learning [2]. Consequently, there has been an increased interest in developing scalable and reliable bilevel optimization methods for machine learning [3].



Existing algorithms for bilevel optimization, such as those based on implicit differentiation, are restricted to the setting where **the lower-level objective is smooth and strongly convex**. This setting ensures the uniqueness of the lower-level solution $y^\star(\theta)$ thus making the bilevel problem well-defined. Additionally, it allows deriving an expression for the Jacobian $\nabla_\theta y^\star(\theta)$ by application of the implicit function theorem thus providing an essential tool for gradient-based optimization.

Challenges Despite recent progress in bilevel optimization, the hierarchical structure of bilevel problems raises many challenges when applied to machine learning problems involving large over-parametrized neural networks. While the use of such networks is ubiquitous and offers high modelling flexibility, it often results in non-convex bilevel problems with multiple solutions for which the generalization properties are poorly understood.

The main attempts to understand generalization of the solutions to bilevel optimization problems are limited as they require strongly-convex lower-level objectives which is rarely the case in the context of deep-learning [4, 5]. Recent work [6] analyzes the generalization of a class of bilevel algorithms (unrolled optimization) for performing hyper-parameter optimization. While the analysis provides some first insights, the approach does not take into account the **implicit bias** of the lower-level optimization procedure which was shown to heavily impact generalization [7]. This is particularly relevant when the lower-level possess multiple global solutions with different generalization properties thus making the quality of the selected solution highly dependent on the optimization procedure.

When the lower-level variable y of a bilevel problem represents the parameters of an over-parameterized neural networks, the lower-level objective becomes non-convex and often possesses **multiple degenerate solutions**. This degeneracy makes the problem **ambiguous** and methods based on implicit differentiation **ill-justified**. Additionally, the generalization properties of the obtained solutions become highly dependent on the optimization procedure [7].

This internship is intended as a preliminary study leading to a PhD project aiming at developing a learning theory for predictive models resulting from a bilevel optimization procedure. As a starting point, the project will consider bilevel problems involving over-parameterized linear models (kernel methods) for which closed form solutions can be easily characterized. This will allow to precisely characterize the implicit bias of the bilevel optimization procedure while taking advantage of recent work studying the statistical properties of over-parameterized linear models in regression problems [8]. The approach will then be extended to over-parameterized networks in the Neural Tangent Kernel (NTK) regime for which convergence and generalization have also been studied [9]. Such NTK regime constitutes an intermediate setting of interest between kernel methods and deep learning that can still provide valuable practical insights. In particular, it is expected that the resulting theory will shed light on regularization techniques for controlling the bias-variance trade-off arising from finite data, therefore, allowing the development of more reliable methods.

Environment The internship will take place at Thoth team at Inria Grenoble and is part of a joint collaboration with Edouard Pauwels and Samuel Vaiter. A fully funded PhD position will be the natural continuation of the internship in case there is a good match between the student, topic and supervisors.

Requirements/Skills

- Strong mathematical background, specially advanced knowledge in probability and statistics.
- Good working knowledge in scientific programming; experience with a deep learning library (e.g. `PyTorch` will be a plus). The applicant should show a strong interest in conducting rigorous machine learning experiments using these libraries.
- A successful candidate should have or be willing to develop a rigorous approach in their research, be receptive to feedback while also having critical thinking and willing to collaborate with researchers from different backgrounds and disciplines.
- Language: excellent English skills both written and spoken.

How to apply Please send an email to the supervisors along with a CV and grades from both Bachelor and Master.

References

[1] Franceschi *et al.*, "Bilevel Programming for Hyperparameter Optimization and Meta-Learning," in *ICML*, 2018.

[2] D. Bansal, R. T. Chen, M. Mukadam, and B. Amos, "Taskmet: Task-driven metric learning for model learning," *arXiv preprint arXiv:2312.05250*, 2023.

[3] Lorraine *et al.*, "Optimizing millions of hyperparameters by implicit differentiation," in *AISTATS*, 2020.

[4] Ghadimi and Wang, "Approximation methods for bilevel programming," *arXiv*, 2018.

[5] Ji *et al.*, "Bilevel Optimization: Convergence Analysis and Enhanced Design," in *ICML*, 2021.

[6] F. Bao *et al.*, "Stability and Generalization of Bilevel Programming in Hyperparameter Optimization," *NeurIPS*, 2021.

[7] P. Vicol *et al.*, "Implicit Regularization in Overparameterized Bilevel Optimization," in *ICML*, 2022.

[8] T. Hastie *et al.*, "Surprises in high-dimensional ridgeless least squares interpolation," *The Annals of Statistics*, vol. 50, 2022.

[9] S. Arora *et al.*, "Fine-grained analysis of optimization and generalization for overparameterized two-layer neural networks," in *ICML*, 2019.