

A Study of the Implicit Bias in Deep Learning through a Higher-Order Analysis of the Loss

- **Keywords:** Deep Learning, Optimization, Gradient Flow, Implicit Bias.
- **Supervisors:**
Pierre Wolinski: pierre.wolinski@dauphine.psl.eu
Clément Royer: clement.royer@dauphine.psl.eu
- **Affiliation and locations:**
MILES Team, LAMSADE Laboratory, Paris Dauphine University–PSL.

1 Context

Deep learning theory is an active field of research, which occupies a central place between the outstanding empirical results of deep learning (e.g., computer vision, natural language processing...) and the well-studied traditional machine learning methods. In this field, studying the training trajectory of a neural network is a key to better understand how it converges quickly to a well-performing model on so many tasks, both from an optimization and a generalization point of view. Leveraging theoretical results in neural network optimization is fundamental to improve our models and training algorithms with a reasonable computational cost, especially at a time when the issue of AI energy consumption is becoming increasingly important.

In particular, over the last years, a promising research direction is being studied in the deep learning theory community: the impressive results of deep learning emerge from a subtle interaction between the training algorithm and the *overparameterization* of the models (i.e., models with many redundant parameters, which include most of neural networks). Thus, contrary to the classical point of view in statistical learning, the ability of neural networks to generalize is not an intrinsic property of the optimal model, independent of the optimization process. This *implicit bias* of the training process is an emerging property of the model and training algorithm, which is not enforced *explicitly*. The implicit bias has been studied and observed for several classes of models and algorithms, but a rigorous theoretical result is still missing in the case of generic neural networks.

Improving our understanding of this phenomenon is then crucial to discover new optimization algorithms adapted to smaller neural network architectures, that require much less computational power to train than the models that are currently used.

2 Project overview

In deep learning, we consider a model parameterized by $\theta \in \mathbb{R}^p$ and a dataset \mathcal{D} , and we aim to minimize the *loss* function $\mathcal{L}_{\mathcal{D}} : \mathbb{R}^p \rightarrow \mathbb{R}$ with respect to θ . In short, the loss $\mathcal{L}_{\mathcal{D}}(\theta)$ measures the

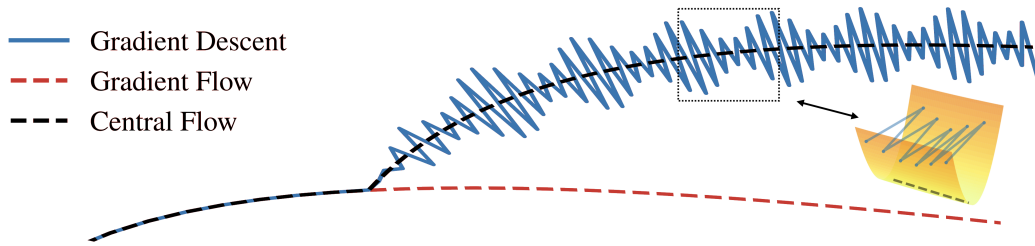


Figure 1: Training trajectory when entering the Edge of Stability phase: gradient flow does not predict any instability, while the gradient descent starts to oscillate; these oscillations are explained by the “central flows” dynamics [Cohen et al., 2025].

quality of the model with parameter θ on the dataset \mathcal{D} (the lower the better). In the following, we may use \mathcal{L} instead of $\mathcal{L}_{\mathcal{D}}$ when there is no ambiguity on the dataset we use.

Studying the implicit bias involves generally a refined analysis of the training trajectory of a model under a specific training algorithm. In the case of neural networks, such a study is very challenging, not only because of their size and complexity, but also because the training algorithm is usually stochastic and discrete. For these reasons, it is common to restrict the study to a small class of models, to make some regularity assumptions about the loss to minimize, or to consider an approximation of the actual training algorithm. For example, the Stochastic Gradient Descent (SGD) algorithm could be modeled with a deterministic, continuous version of it, the *gradient flow*:

$$\text{SGD: } \theta_{t+1} = \theta_t - \eta \nabla \mathcal{L}_{\mathcal{B}_t}(\theta_t), \quad \text{gradient flow: } \frac{d\theta}{dt}(t) = -\nabla \mathcal{L}_{\mathcal{D}}(\theta(t))$$

where $\eta > 0$ is the learning rate and \mathcal{B}_t is some random subset of \mathcal{D} , called *minibatch*.

Edge of Stability. In recent years, approaches based on regularity assumptions on the loss or the gradient flow approximation of the SGD have been challenged by the *Edge of Stability* (EoS) phenomenon [Cohen et al., 2021, Damian et al., 2023]. This phenomenon occurs consistently when training neural networks using Stochastic Gradient Descent (SGD): for an arbitrary small learning rate $\eta > 0$, the largest eigenvalue λ_1 of the Hessian of the loss tends to grow until reaching the boundary between the “stable” phase $\lambda_1 < 2/\eta$ and the “unstable” phase $\lambda_1 > 2/\eta$, and oscillates around this boundary for the rest of the training [Cohen et al., 2021, Damian et al., 2023]. The EoS is partly explained by the *central flows* heuristic [Cohen et al., 2025] (see Figure 1), which exhibits an implicit bias of gradient descent towards regions where the maximal eigenvalue of the Hessian of the loss $\nabla^2 \mathcal{L}$ is not too large.

However, the EoS remains to be studied theoretically with rigorous results. A good starting point is to construct simple loss functions or models. These minimal examples should demonstrate the validity of the existing heuristics, at least in specific cases. On the practical side, the first step is to reproduce existing practical results and isolate the EoS phenomenon to determine its role in optimization and generalization. The aim of this first phase of research is to acquire a good theoretical understanding of the phenomenon when using small models, and to build an experimental framework for empirically evaluating theoretical hypotheses. It is then crucial that, in the process, the higher-order information on the loss involved in the theoretical results remains computable [Pearlmutter, 1994, Wolinski, 2025], since they appear explicitly in the central flows.

Generalization and higher-order derivatives of the loss. Despite their heuristic nature, the work on the central flows shows the importance of the higher-order derivatives of the loss in a specific aspect of the implicit bias: limiting the sharpness of the loss surface, i.e., the maximum eigenvalue of the Hessian. At this point, the study of the implicit bias focuses on the trajectory of the *training*

loss, and not on the *test loss*. So, the curvature of the loss can be explained, but only on the training set.

However, the curvature of the training loss is known to be related to generalization properties of neural networks [Foret et al., 2021]. The results obtained in the first part can be used to draw a link between the EoS, the higher-order derivatives, and generalization. The goal of this second part is thus to obtain theoretical results on generalization, even in small models, that involve the training dynamics via the Hessian and order-3 derivatives of the loss. Again, these results will be tested empirically by using the generalization of the Hessian-vector product and projections of the higher-order derivatives on relevant subspaces [Pearlmutter, 1994, Wolinski, 2025].

Improving optimization and generalization. Ultimately, we will use the preceding theoretical results to propose new training algorithms, to improve optimization and generalization. On the optimization part, a better understanding of the EoS through the higher-order derivatives will indicate if the EoS is desirable at the beginning of training, at the end, or not at all. With this better understanding, we will be able to use the higher-order derivatives to propose new training algorithms, e.g., Newton-based algorithms [Arbel et al., 2023, Royer et al., 2020], possibly with cubic regularization [Nesterov and Polyak, 2006, Wolinski, 2025].

On the generalization part, our main goal is to simulate the implicit regularization on small neural networks. That is, once we are able to relate the higher-order derivatives of the loss to generalization, it will be possible to provide an *explicit* penalty we have to add to the loss to mimic the influence of the implicit bias. Expressing such a penalty starting from an implicit constraint can be done with various approaches, including a Bayesian one [Wolinski et al., 2020].

References

- Jeremy Cohen, Simran Kaur, Yuanzhi Li, J Zico Kolter, and Ameet Talwalkar. Gradient descent on neural networks typically occurs at the edge of stability. In *International Conference on Learning Representations*, 2021.
- Alex Damian, Eshaan Nichani, and Jason D Lee. Self-stabilization: The implicit bias of gradient descent at the edge of stability. In *International Conference on Learning Representations*, 2023.
- Jeremy Cohen, Alex Damian, Ameet Talwalkar, J Zico Kolter, and Jason D Lee. Understanding optimization in deep learning with central flows. In *International Conference on Learning Representations*, 2025.
- Barak A Pearlmutter. Fast exact multiplication by the Hessian. *Neural computation*, 6(1):147–160, 1994.
- Pierre Wolinski. Gathering and exploiting higher-order information when training large structured models. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, 2025.
- Pierre Foret, Ariel Kleiner, Hossein Mobahi, and Behnam Neyshabur. Sharpness-aware minimization for efficiently improving generalization. In *International Conference on Learning Representations*, 2021.
- Michael Arbel, Romain Menegaux, and Pierre Wolinski. Rethinking Gauss-Newton for learning over-parameterized models. In *Advances in Neural Information Processing Systems*, volume 36, pages 33379–33402, 2023.

- Clément W Royer, Michael O’Neill, and Stephen J Wright. A Newton-CG algorithm with complexity guarantees for smooth unconstrained optimization. *Mathematical Programming*, 180(1):451–488, 2020.
- Yurii Nesterov and Boris T Polyak. Cubic regularization of Newton method and its global performance. *Mathematical Programming*, 108(1):177–205, 2006.
- Pierre Wolinski, Guillaume Charpiat, and Yann Ollivier. An equivalence between bayesian priors and penalties in variational inference. *arXiv preprint arXiv:2002.00178*, 2020.