Pierre Wolinski

Curriculum Vitæ

Themes: theory of deep learning, variational inference, initialization, pruning, Bayes.

Professional Experience

2021– **Post-doc**, *Statify Team, LJK, UGA, Inria Grenoble-Alpes*, Grenoble, France.
Themes: theory of deep learning, variational inference, optimization and generalization.
Supervisor: Julyan Arbel.

2020–2021 **Post-doc**, *Department of statistics, University of Oxford*, Oxford, UK.
Themes: Bayesian deep learning, variational inference, theory of deep learning.
Supervisor: Judith Rousseau.

Study

2016–2020 **PhD in Computer Science**, *TAO/Tau Team, LRI, Inria Saclay, Université Paris-Saclay*.
Title: *Structural Learning of Neural Networks*.
Supervisor: Guillaume Charpiat, Yann Ollivier.

2011–2016 **École Normale Supérieure (Mathematics)**, Paris, France.
2016: Graduate, Math with option Physics.
2015: Master in Math (Probability and Statistics), Université Paris-Sud, Orsay, France.
2015: Master Thesis: *Consistency of RKHS Methods in the Case of Minimization of a Convex Risk*, supervised by Éric Moulines, Florence d'Alché-Buc and François Roueff, Télécom Paris, France.

2008–2011 **Classe Préparatoire aux Grandes Écoles (Physics and Chemistry)**, *Lycée Fénelon*, Paris, France.

2008 **Baccalauréat (S)**, *Lycée Marie-Curie*, Sceaux, France.

Teaching

2016–2020 **Lecturer in Math and Computer Science**, *IUT d'informatique*, Orsay, France.
Courses: Algebra; Probability and Stat.; Java et OOP; Graphs, Languages and Finite-state machines.

2012–2013 **Lecturer in CPGE (Math)**, *Lycée Saint-Louis*, Paris, France.

Works

- *Rethinking Gauss-Newton for learning over-parameterized models* (2023, submitted), M. Arbel, R. Ménégaux*, **P. Wolinski***.
- *Gaussian Pre-Activations in Neural Networks: Myth or Reality?* (2022, submitted), **P. Wolinski**, J. Arbel.
- *Interpreting a Penalty as the Influence of a Bayesian Prior* (2020, preprint), **P. Wolinski**, G. Charpiat, Y. Ollivier.
- *Learning with Random Learning Rates* (2019, ECML PKDD), L. Blier*, **P. Wolinski***, Y. Ollivier.
- *Asymmetrical Scaling Layers for Stable Network Pruning* (2019, preprint), **P. Wolinski**, G. Charpiat, Y. Ollivier.

* Equal contribution.

68, avenue du maréchal Foch – 92 260 Fontenay-aux-Roses – France
06 48 42 31 17 • *pierre.wolinski@normalesup.org*
http://pierre-wolinski.fr/ • *p-wol*

Conferences

2022 **ISBA** – *An Equivalence between Bayesian Priors and Penalties in Variational Inference*
(oral presentation)

2022 **JdS** – *How to Impose Gaussian Pre-Activations in a Neural Network?*
(oral presentation)

2020 **CMStatistics** – *Interpreting a Penalty as the Influence of a Bayesian Prior*
(oral presentation)

2019 **ECML PKDD** – *Learning with Random Learning Rates*
(oral presentation + poster)

Skills

Languages

French, English (+ German).

Computer Science

- Languages: Python, C++ (+ Java, matlab).
- Libraries: PyTorch, matplotlib (+ pandas, Hydra).
- Software: git.
- Cluster: GPU, job scheduling (Slurm, OAR), environment management (conda, docker).
- Experience on the computation servers of the Idris (Jean Zay).

Code

- `https://github.com/p-wol/gaussian-preact`:
reproducibility of *Gaussian Pre-Activations in Neural Networks: Myth or Reality?*
- `https://github.com/leonardblier/alrao`:
implementation of the technique proposed in *Learning with Random Learning Rates*.

Experiences

- Paper reviewing for: NeurIPS, ICML, ICLR, AISTATS, JMLR.
- Support and writing of a wiki for the use of clusters and GPUs.

Recommendations

- Julyan Arbel: supervisor, Inria Grenoble-Alpes — `julyan.arbel@inria.fr`
- Judith Rousseau: supervisor, Université d'Oxford — `rousseau@ceremade.dauphine.fr`
- Florence Forbes: head of team, Inria Grenoble-Alpes — `florence.forbes@inria.fr`
- Guillaume Charpiat: PhD advisor, Inria Saclay — `guillaume.charpiat@inria.fr`
- Yann Ollivier: PhD co-advisor, CNRS/Meta — `yann@yann-ollivier.org`
- Michèle Sebag: head of team, Inria Saclay — `michele.sebag@inria.fr`

Hobbies

- Activities: theater, dance (rock, waltz, tango).
- History/philosophy of science.
- Participation to the French Cup of Robotics (2012, 2013, 2015).

68, avenue du maréchal Foch – 92 260 Fontenay-aux-Roses – France
06 48 42 31 17 • *pierre.wolinski@normalesup.org*
http://pierre-wolinski.fr/ • *p-wol*

Outline of Papers

Gaussian Pre-Activations in Neural Networks: Myth or Reality?

Studying the propagation of the pre-activations of a neural network at initialization is the starting point of many weight and biases initialization techniques. In this field, it is very common to assume that that the pre-activations are Gaussian. Although this very practical assumption is justified when the number of neurons per layer tends to infinity, it has been called into question by theoretical and experimental work on finite-width neural networks. Our main contribution is to construct a family of pairs of activation functions and initialization distributions that guarantee that pre-activations remain Gaussian throughout their propagation in the network, even with few neurons per layer. In the course of our analysis, we discover a set of constraints that a neural network must fulfill in order to have Gaussian pre-activations. In addition, we make a critical review of the results obtained in the "Edge of Chaos" line of works, testing several hypotheses and results related to the propagation of the distribution of the pre-activations. We also propose a unified representation of the propagation of the pre-activations, encompassing several well-known initialization procedures. Finally, our study provides a basis for answering the question: is it desirable to initialize a neural network so as to have Gaussian pre-activations?

Interpreting a Penalty as the Influence of a Bayesian Prior

In machine learning, it is common practice to optimize the parameters of a model with an *ad hoc* penalty term, which pushes the parameters in directions chosen at the outset (e.g., towards zero). The penalty term naturally appears in variational inference, a technique for approximating the Bayesian posterior when it is too difficult to calculate exactly. In this context, the penalty is proportional to a Kullback-Leibler (KL) divergence between the approximation of the posterior distribution and the prior distribution. We characterize which penalties can take the form of a KL, and propose a formula for calculating the prior distribution corresponding to a given penalty. Among other things, this approach makes it possible to estimate the penalty factor, which is usually a hyperparameter to be optimized, in the case of neural networks.

Learning with Random Learning Rates

When training a neural network, the choice of the learning rate η for gradient descent has a major influence on its final performance. This phenomenon prevents us from training a model without first searching for the optimal η. We propose the *All Learning Rates At Once* (Alrao) algorithm for neural networks: each neuron is trained with its own learning rate, randomly drawn before training according to a probability distribution covering several orders of magnitude. In short, we replace the *optimality* of the learning rate by the *diversity* of learning rates in the network. It is therefore not necessary to carry out several runs to optimize η. Our experimental results show that Alrao performs almost as well as conventional gradient descent with an optimized learning rate. Moreover, *all* runs performed with Alrao resulted in a network with decent performance, whereas some training sessions with Adam failed.

This work was carried out in collaboration with Léonard Blier and Yann Ollivier (facebook AI Research). I contributed to the implementation of the main part of Alrao (python + PyTorch), to the construction of an easy-to-use user interface, and to experiments in the context of natural language processing (NLP). I have also carried out additional experiments to calculate the order of magnitude of the optimal learning rate for each layer.

68, avenue du maréchal Foch – 92 260 Fontenay-aux-Roses – France
06 48 42 31 17 • *pierre.wolinski@normalesup.org*
http://pierre-wolinski.fr/ • *p-wol*

Research Axes

My research, past and present, is based on a constant back-and-forth between, on the one hand, empirical and theoretical results, sometimes obtained at the cost of simplifying hypotheses, and, on the other hand, the applications and testing of these hypotheses in practical cases. For example:

- *Learning with Random Learning Rates*: we have proposed a practical application of the study [1], according to which it is possible to train a network where certain neurons remain fixed;
- *Interpreting a Penalty as the Influence of a Bayesian Prior*: we have built a theoretical relation between the penalty and the Bayesian prior, and we have deduced a heuristic on the penalty factor;
- *Gaussian Pre-Activations in Neural Networks: Myth or Reality?*: we have tested and rejected the hypothesis of Gaussian pre-activations, derived from the study of infinitely large neural networks, and proposed a method for making it valid in practice, on networks of finite width.

Optimization trajectory of a neural network and generalization

In the field of neural networks, generalization is still poorly understood: although capable of learning entire data sets by heart [5], they generalize well in practice. Resolving this paradox is crucial on both theoretical and practical levels. On the theoretical side, several lines of research are investigating the relationship between the training trajectory and the properties of the final neural network [3, 2]. On the practical side, predicting a model's ability to generalize could speed up the search for optimal architecture and hyperparameters, or dispense with a validation set.

In this line of research, my aim is to evaluate the assumptions commonly made in work on training trajectories. A first approach is to check whether the approximation of the discrete gradient descent by a continuous gradient descent gives a faithful representation of the training trajectory. To check this, it is possible to use measures of the regularity of the loss surface along the trajectory, such as the higher-order derivatives of the loss with respect to the parameters. In the process, these measures could be used to predict generalization performance.

Variational inference: interactions between the training set and the loss

Research on the "cold posterior effect" [4] has shown that, to set a good penalty factor, we need to take into account the increase in data and the structure of the dataset itself. Indeed, theory predicts that this factor should be $1/N$, where N is the number of data items. It is therefore normal that transformations in the dataset influence this factor. However, there is no practical relationship between the choice of this factor, the topology of the dataset and the transformations performed beforehand. In this axis, my aim is to find such a relationship, in order to provide a heuristic for setting the penalty factor (or penalty), depending on the variety of data in the training set.

References

[1] A. M. Saxe, P. W. Koh, Z. Chen, M. Bhand, B. Suresh, and A. Y. Ng. On random weights and unsupervised feature learning. *ICML*, 2(3):6, 2011.

[2] U. Simsekli, O. Sener, G. Deligiannidis, and M. A. Erdogdu. Hausdorff dimension, heavy tails, and generalization in neural networks. *NeurIPS*, 33:5138–5151, 2020.

[3] D. Soudry, E. Hoffer, M. S. Nacson, S. Gunasekar, and N. Srebro. The implicit bias of gradient descent on separable data. *JMLR*, 19(1):2822–2878, 2018.

[4] F. Wenzel, K. Roth, B. S. Veeling, J. Świątkowski, L. Tran, S. Mandt, J. Snoek, T. Salimans, R. Jenatton, and S. Nowozin. How good is the bayes posterior in deep neural networks really? *ICML*, pages 10248–10259, 2020.

[5] C. Zhang, S. Bengio, M. Hardt, B. Recht, and O. Vinyals. Understanding deep learning (still) requires rethinking generalization. *Communications of the ACM*, 64(3):107–115, 2021.

68, avenue du maréchal Foch – 92 260 Fontenay-aux-Roses – France

06 48 42 31 17 • *pierre.wolinski@normalesup.org*

http://pierre-wolinski.fr/ • *p-wol*