Pierre Wolinski

Curriculum Vitæ

Thèmes : théorie des réseaux de neurones, inférence variationnelle, initialisation, élagage, Bayes.

Parcours professionnel

2021– **Post-doc**, *Équipe Statify, LJK, UGA, Inria Grenoble-Alpes*, Grenoble.
Thèmes : théorie des réseaux de neurones, inférence variationnelle, optimisation et généralisation.
Encadrant : Julyan Arbel.

2020–2021 **Post-doc**, *Department of statistics, University of Oxford*, Oxford.
Thèmes : réseaux de neurones bayésiens, inférence variationnelle, théorie des réseaux de neurones.
Encadrante : Judith Rousseau.

Formation

2016–2020 **Doctorat en Informatique**, *Équipe TAO/Tau, LRI, Inria Saclay, Université Paris-Saclay*.
Titre : *Apprentissage de structure pour les réseaux de neurones*.
Encadrants : Guillaume Charpiat, Yann Ollivier.

2011–2016 **École Normale Supérieure (Mathématiques)**, Paris.
2016 : Diplôme de l'ENS, Mathématiques option Physique.
2015 : Master de Mathématiques (Probabilités et Statistiques), Université Paris-Sud, Orsay.
2015 : Mémoire de Master : *Consistance des méthodes RKHS dans le cadre de la minimisation d'un risque convexe*, encadré par Éric Moulines, Florence d'Alché-Buc et François Roueff, Télécom Paris.

2008–2011 **Classe Préparatoire aux Grandes Écoles (Physique et Chimie)**, *Lycée Fénelon*, Paris.

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

Enseignement

2016–2020 **Chargé de TD/TP (Mathématiques et Informatique)**, *IUT d'informatique*, Orsay.
Matières : algèbre ; probabilités et statistiques ; Java et POO ; graphes, langages et automates finis.

2012–2013 **Interrogateur en CPGE (Mathématiques)**, *Lycée Saint-Louis*, Paris.

Travaux

- *Rethinking Gauss-Newton for learning over-parameterized models* (2023, soumis), M. Arbel, R. Ménégaux*, **P. Wolinski***.
- *Gaussian Pre-Activations in Neural Networks: Myth or Reality?* (2022, soumis), **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.

* Contribution égale.

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*

Conférences

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

2022 **JdS** – *Comment imposer des pré-activations gaussiennes dans un réseau de neurones ?* (présentation orale)

2020 **CMStatistics** – *Interpreting a Penalty as the Influence of a Bayesian Prior* (présentation orale)

2019 **ECML PKDD** – *Learning with Random Learning Rates* (présentation orale + poster)

Compétences

Langues

Français, anglais (+ allemand).

Informatique

- Langages : Python, C++ (+ Java, matlab).
- Bibliothèques : PyTorch, matplotlib (+ pandas, Hydra).
- Logiciels : git.
- Utilisation de clusters : GPU, job scheduling (Slurm, OAR), gestion d'environnement (conda, docker).
- Expérience sur les serveurs de calcul de l'Idris (Jean Zay).

Code

- `https://github.com/p-wol/gaussian-preact` : reproductibilité de *Gaussian Pre-Activations in Neural Networks: Myth or Reality?*
- `https://github.com/leonardblier/alrao` : implémentation de la technique proposée dans *Learning with Random Learning Rates*.

Expériences

- Relecture d'articles pour : NeurIPS, ICML, ICLR, AISTATS, JMLR.

Recommandations

- Julyan Arbel : encadrant, Inria Grenoble-Alpes.
- Florence Forbes : cheffe d'équipe, Inria Grenoble-Alpes.

Centres d'intérêt

- Activités : théâtre, danse (rock, valse, tango).
- Histoire/philosophie des sciences.
- Participation à la Coupe de France de robotique (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*

Présentation d'articles

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

L'étude de la propagation des pré-activations lors de l'initialisation des réseaux de neurones est à l'origine de nombreuses techniques d'initialisation des poids et des biais. Dans ce domaine, il est très courant de supposer que les pré-activations sont gaussiennes. Bien que cette hypothèse, très pratique, soit justifiée lorsque le nombre de neurones par couche tend vers l'infini, elle est remise en question par les travaux théoriques et expérimentaux sur les réseaux de neurones à largeur finie. Notre principale contribution consiste à construire une famille de paires de fonctions d'activation et de distributions d'initialisation qui garantissent que les pré-activations restent gaussiennes tout au long de leur propagation dans le réseau, même avec peu de neurones par couche. Au cours de notre analyse, nous découvrons un ensemble de contraintes qu'un réseau de neurones doit respecter pour avoir des pré-activations gaussiennes. De plus, nous faisons une revue critique des résultats obtenus dans la lignée de l'« *Edge of Chaos* », en testant plusieurs hypothèses et résultats liés à la propagation de la distribution des pré-activations. Nous proposons également une représentation unifiée de la propagation des pré-activations, englobant plusieurs procédures d'initialisation bien connues. Enfin, notre étude fournit une base de travail pour répondre à la question : est-il souhaitable d'initialiser un réseau de neurones de façon à avoir des pré-activations gaussiennes ?

Interpreting a Penalty as the Influence of a Bayesian Prior

En apprentissage automatique, il est courant d'optimiser les paramètres d'un modèle avec un terme de pénalité *ad hoc*, qui pousse les paramètres dans des directions choisies au départ (par exemple vers zéro). Le terme de pénalité apparaît naturellement en inférence variationnelle, technique permettant d'approximer l'*a posteriori* bayésien dans des contextes où il est trop difficile à calculer exactement. Dans ce cadre, la pénalité est proportionnelle à une divergence de Kullback–Leibler (KL) entre l'approximation de la loi *a posteriori* et la loi *a priori*. Nous caractérisons quelles pénalités peuvent prendre la forme d'une KL, et proposons une formule pour calculer la loi *a priori* correspondant à une pénalité donnée. Entre autres, ce point de vue permet d'estimer le facteur de pénalité, qui est usuellement un hyperparamètre à optimiser, dans le cas des réseaux de neurones.

Learning with Random Learning Rates

Lors de l'entraînement d'un réseau de neurones, le choix du taux d'apprentissage η de la descente de gradient influence grandement sa performance finale. Ce phénomène nous empêche d'entraîner un modèle sans rechercher préalablement le η optimal. Nous proposons l'algorithme All Learning Rates At Once (Alrao, « tous les taux d'apprentissage à la fois ») pour les réseaux de neurones : chaque neurone est entraîné avec son propre taux d'apprentissage, tiré aléatoirement avant l'entraînement selon une loi de probabilité recouvrant plusieurs ordres de grandeur. En somme, nous remplaçons l'*optimalité* du taux d'apprentissage par la *diversité* des taux d'apprentissage dans le réseau. Il n'est donc pas nécessaire de procéder à plusieurs entraînements pour optimiser η. Nos résultats expérimentaux montrent qu'Alrao est presque aussi performant que la descente de gradient usuelle avec un taux d'apprentissage optimisé. De plus, tous les entraînements avec Alrao ont abouti à un réseau aux performances convenables, alors que certains entraînements avec Adam ont échoué.

Ce travail a été réalisé en collaboration avec Léonard Blier et Yann Ollivier (facebook AI Research). J'ai contribué à l'implémentation de la partie principale d'Alrao (python + PyTorch), à la construction d'une interface utilisateur facile à prendre en main, et à des expériences dans le contexte du traitement du langage naturel (NLP). J'ai également réalisé des expériences qui permettent de calculer l'ordre de grandeur du taux d'apprentissage optimal pour chaque couche.

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*

Axes de recherche

Mes axes de recherches, passés et actuels, sont fondés sur un aller-retour permanent entre, d'un côté, les résultats empiriques et théoriques, parfois obtenus au prix d'hypothèses simplificatrices, et, de l'autre côté, les applications et le test de ces hypothèses dans des cas pratiques. Par exemple :

- *Learning with Random Learning Rates* : nous avons proposé une application pratique du résultat [1], selon lequel il est possible d'entraîner un réseau où certains neurones restent figés ;
- *Interpreting a Penalty as the Influence of a Bayesian Prior* : nous avons construit une relation théorique entre la pénalité et l'*a priori* bayésien, et en avons déduit une heuristique sur le facteur de pénalité ;
- *Gaussian Pre-Activations in Neural Networks: Myth or Reality?* : nous avons testé et invalidé l'hypothèse des pré-activations gaussiennes, issue de l'étude des réseaux de neurones infiniment larges, et nous avons proposé une méthode pour la rendre valide en pratique, sur des réseaux de largeur finie.

Trajectoire d'optimisation d'un réseau de neurones et performance en généralisation

Dans le domaine des réseaux de neurones, la généralisation est encore mal comprise : bien que capables d'apprendre par cœur des jeux de données entiers [5], ils généralisent très bien en pratique. La résolution de ce paradoxe est cruciale à la fois sur les plans théorique et pratique. Du côté théorique, plusieurs axes de recherche étudient la relation entre la trajectoire d'entraînement et les propriétés du réseau de neurones final [3, 2]. Du côté pratique, prédire la capacité d'un modèle à généraliser permettrait d'accélérer la recherche d'une architecture et d'hyperparamètres optimaux, ou de se passer d'un ensemble de validation.

Dans cet axe de recherche, mon objectif est d'évaluer les hypothèses couramment effectuées dans les travaux sur la trajectoire d'entraînement. Une première approche consiste à vérifier si l'approximation de la descente de gradient discrète par une descente de gradient continue rend compte fidèlement de la trajectoire d'entraînement. Pour vérifier cela, il est possible d'utiliser des mesures de régularité de la surface de perte le long de la trajectoire, telles que les dérivées d'ordre supérieur de la perte par rapport aux paramètres. Au passage, ces mesures pourraient être utilisées pour prédire la performance en généralisation.

Inférence variationnelle : interactions entre l'ensemble d'entraînement et la perte

Des recherches sur le « *cold posterior effect* » [4] ont montré que, pour fixer un bon facteur de pénalité, il faut prendre en compte l'augmentation de données et la structure du jeu de données lui-même. En effet, la théorie prévoit que ce facteur doit être de $1/N$, où N est le nombre de données. Il est donc normal que des transformations du jeu de données influent sur ce facteur. Mais il n'existe pas de relation utilisable en pratique entre le choix de ce facteur, la topologie du jeu de données, et les transformations qu'on lui fait subir. Dans cet axe, mon but est de trouver une telle relation, afin de fournir une heuristique pour régler le facteur de pénalité (ou la pénalité), en fonction de la variété des données de l'ensemble d'entraînement.

Références

[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*