Modèles de Réseaux de neurones avec poids dépendants: Limite, parcimonie et compressibilité

Hoil Lee [1], Fadhel Ayed [2], Paul Jung [3], Juho Lee [1], Hongseok Yang [1], François Caron [4]

[1] *KAIST, South Korea, hoil.lee@kaist.ac.kr, juholee@kaist.ac.kr, hongseok.yang@kaist.ac.kr*
[2] *Huawei Technologies, France, fadhel.ayed@gmail.com*
[3] *Fordham University, USA, pjung3@fordham.edu*
[4] *University of Oxford, UK, caron@stats.ox.ac.uk*

Résumé. Ce travail étudie la limite des réseaux neuronaux profonds dont les poids sont dépendants et modélisés via un mélange de distributions gaussiennes. Sous ce modèle, nous montrons que chaque couche du réseau neuronal, losque la largeur tend vers l'infini, peut être caractérisée par deux quantités simples : un paramètre scalaire non-négatif et une mesure de Lévy sur les réels positifs. Si les paramètres scalaires sont strictement positifs et les mesures de Lévy sont triviales alors on retrouve la limite classique du processus gaussien (PG), obtenue avec des poids gaussiens iid. De façon plus intéressante, si la mesure de Lévy d'au moins une couche n'est pas triviale, nous obtenons un mélange de processus gaussiens (MdPG) dans la limite de grande largeur. Le comportement du réseau neuronal dans ce régime est très différent du régime PG. On obtient en effet des sorties corrélées, avec des distributions non gaussiennes, possiblement à queues lourdes. Nous illustrons certains des avantages du régime MdGP sur le régime PG en termes d'apprentissage de représentation et de compressibilité sur des ensembles de données simulées, MNIST et Fashion MNIST.

Mots-clés. Réseaux de neurones, processus gaussien, processus stochastique, mesure de Lévy

Abstract. This work studies the infinite-width limit of deep feedforward neural networks whose weights are dependent, and modelled via a mixture of Gaussian distributions. Under this model, we show that each layer of the infinite-width neural network can be characterised by two simple quantities: a non-negative scalar parameter and a Lévy measure on the positive reals. If the scalar parameters are strictly positive and the Lévy measures are trivial at all hidden layers, then one recovers the classical Gaussian process (GP) limit, obtained with iid Gaussian weights. More interestingly, if the L'evy measure of at least one layer is non-trivial, we obtain a mixture of Gaussian processes (MoGP) in the large-width limit. The behaviour of the neural network in this regime is very different from the GP regime. One obtains correlated outputs, with non-Gaussian distributions, possibly with heavy tails. We illustrate some of the benefits of the MoGP regime over the GP regime in terms of representation learning and compressibility on simulated, MNIST and Fashion MNIST datasets.

Keywords. Neural networks, Gaussian process, stochastic processes, Lévy measure

1 Introduction

Deux décennies après le travail fondateur de Radford Neal [1996], ces dernières années ont vu un intérêt renouvelé et croissant pour l'analyse des réseaux neuronaux (profonds) avec des poids aléatoires, dans la limite de largeur infinie. Lorsque les poids sont indépendants et identiquement distribués (iid) selon une loi gaussienne, la fonction aléatoire correspondant à ce réseau neuronal aléatoire converge vers un processus gaussien [Neal, 1996, Lee et al., 2018, Matthews et al., 2018]. La connexion avec les processus gaussiens a approfondi notre compréhension des réseaux neuronaux de grande largeur et a motivé à la fois l'utilisation de méthodes d'inférence de régression bayésienne ou par noyau [Lee et al., 2018] ainsi que le développement de méthodes de noyau pour l'entraînement par descente de gradient dans la limite de largeur infinie [Jacot et al., 2018].

Bien qu'instructive, la connexion avec les processus gaussiens souligne également certaines des limitations des réseaux neuronaux de grande largeur avec des poids gaussiens iid. Comme déjà noté par Neal, " avec des priors gaussiens, les contributions des unités cachées individuelles sont toutes négligeables, et par conséquent, ces unités ne représentent pas des 'caractéristiques cachées' qui capturent des aspects importants des données." De plus, les différentes dimensions de la sortie du réseau neuronal deviennent indépendantes dans la limite de largeur infinie, ce qui est généralement indésirable. Enfin, d'un point de vue bayésien, l'hypothèse d'indépendance gaussienne sur les poids est souvent considérée comme irréaliste : les poids estimés des réseaux neuronaux profonds montrent généralement des dépendances et des queues lourdes, et donc une famille de distributions a priori permettant des queues lourdes est souhaitable. Pour atténuer certaines de ces limitations, des poids aléatoires non gaussiens iid ont été considérés. Cependant, en raison de la même hypothèse iid, certaines des limitations mentionnées persistent, telles que l'indépendance des dimensions de la sortie.

Nous considérons ici une distribution plus structurée sur les poids du réseau neuronal. Nous supposons que les poids émanant d'un noeud donné sont dépendants, et cette dépendance est capturée via un mélange de gaussiennes. Nous nous intéressons à la limite, lorsque la largeur tend vers l'infini, de ce réseau de neurones, et montrons que la limite est un mélange de processus gaussiens. Cette limite est simplement caractérisée, pour chaque couche du réseau, par un paramètre scalaire et une mesure de Lévy. Cette limite a également un certain nombre de propriétés intéressantes par rapport au régime asymptotique standard du processus gaussien: les sorties sont dépendantes, potentiellement avec des queues lourdes, et le réseau de neurone associé est parcimonieux et compressible. Les détails peuvent être trouvés dans l'article [Lee, 2023].

Bibliographie

A. Jacot, F. Gabriel, and C. Hongler. Neural tangent kernel: Convergence and generalization in neural networks. In Proceedings of the 32nd Conference on Neural Information Processing Systems (NeurIPS'18), pages 8571–8580, 2018.

H. Lee, F. Ayed, P. Jung, J. Lee, H. Yang, F. Caron. Deep Neural Networks with Dependent Weights: Gaussian Process Mixture Limit, Heavy Tails, Sparsity and Compressibility. Journal of Machine Learning Research, 2023.

J. Lee, Y. Bahri, R. Novak, S. S. Schoenholz, J. Pennington, and J. Sohl-Dickstein. Deep neural networks as Gaussian processes. In Proceedings of the 6th International Conference on Learning Representations (ICLR'18), 2018.

A. G de G Matthews, J. Hron, M. Rowland, R. E. Turner, and Z. Ghahramani. Gaussian process behaviour in wide deep neural networks. In Proceedings of the 6th International Conference on Learning Representations (ICLR'18), 2018.

R. M. Neal. Priors for infinite networks. In Bayesian Learning for Neural Networks, pages 29–53. Springer New York, 1996.