Régularisation implicite des réseaux de neurones profonds vers des EDO neuronales

Pierre Marion [1] & Yu-Han Wu [2] & Michael E. Sander [3] & Gérard Biau [4]

[1] *Institut de mathématiques, EPFL, Suisse,* pierre.marion@epfl.ch
[2] *Sorbonne Université, CNRS, Laboratoire de Probabilités, Statistique et Modélisation, LPSM, F-75005 Paris, France,* yu-han.wu@ens.psl.eu
[3] *École Normale Supérieure, CNRS, Département de Mathématiques et Applications, F-75005 Paris, France,* michael.sander@ens.fr
[4] *Sorbonne Université, CNRS, Laboratoire de Probabilités, Statistique et Modélisation, LPSM, F-75005 Paris, France,* gerard.biau@sorbonne-universite.fr

Résumé. Les réseaux neuronaux résiduels sont des modèles de pointe en apprentissage profond. Leur analogue à profondeur continue, les équations différentielles ordinaires (EDO) neuronales, sont également largement utilisées. Malgré leur succès, le lien entre les modèles discrets et continus manque encore d'une base mathématique solide. Dans cette contribution, nous faisons un pas dans cette direction en établissant une régularisation implicite des réseaux neuronaux résiduels profonds vers les EDO neuronales, pour des réseaux non linéaires entraînés avec un flot de gradient. Nous démontrons que si le réseau est initialisé comme une discrétisation d'une EDO neuronale, alors cette propriété est maintenue tout au long de l'entraînement. Nos résultats sont valides pour un temps d'entraînement fini, et également lorsque le temps d'entraînement tend vers l'infini à condition que le réseau satisfasse une condition de Polyak-Łojasiewicz. De plus, cette condition est vérifiée pour une famille de réseaux résiduels où les résidus sont des perceptrons à deux couches avec une surparamétrisation en largeur qui est seulement linéaire. Dans ce cas, nous montrons la convergence du flot de gradient vers un minimum global. Des expériences numériques illustrent nos résultats.

Mots-clés. Réseaux de neurones, apprentissage, flot de gradient, régularisation implicite

Abstract. Residual neural networks are state-of-the-art deep learning models. Their continuous-depth analog, neural ordinary differential equations (ODEs), are also widely used. Despite their success, the link between the discrete and continuous models still lacks a solid mathematical foundation. In this contribution, we take a step in this direction by establishing an implicit regularization of deep residual networks towards neural ODEs, for nonlinear networks trained with gradient flow. We prove that if the network is initialized as a discretization of a neural ODE, then such a discretization holds throughout training. Our results are valid for a finite training time, and also as the training time tends to infinity provided that the network satisfies a Polyak-Łojasiewicz condition. Importantly, this condition holds for a family of residual networks where the residuals are two-layer perceptrons with an overparameterization in width that is only linear, and implies the convergence of gradient flow to a global minimum. Numerical experiments illustrate our results.

Keywords. Neural networks, learning, gradient flow, implicit regularization

Nous nous intéressons aux propriétés des réseaux de neurones résiduels qui s'écrivent

$$\begin{aligned} h_0 &= Ax, \\ h_{k+1} &= h_k + \frac{1}{L} V_{k+1} \sigma(W_{k+1} h_k), \quad 0 \leqslant k \leqslant L-1, \\ F(x) &= B h_L, \end{aligned}$$

où la donnée est $x \in \mathbb{R}^d$, la matrice A appartient à $\mathbb{R}^{q \times d}$, les états cachés h_k sont dans $\mathbb{R}^q$, les matrices V_{k+1}, W_{k+1} appartiennent respectivement à $\mathbb{R}^{q \times m}$ et $\mathbb{R}^{m \times q}$, et $B \in \mathbb{R}^{d' \times q}$. Nous considérons une initialisation dite régulière des paramètres V_k et W_k, ce qui correspond à prendre les V_k et W_k comme des discrétisations de fonctions régulières (potentiellement aléatoires) $\mathcal{V} : [0,1] \to \mathbb{R}^{q \times m}$ et $\mathcal{W} : [0,1] \to \mathbb{R}^{m \times q}$, soit $V_k = \mathcal{V}(k/L)$ and $W_k = \mathcal{W}(k/L)$ pour $k \in \{1, \dots, L\}$. Cela inclut en particulier le cas où les V_k et W_k sont initialisées égales aux mêmes matrices V et W indépendamment de k. Dans ce cas, le réseau de neurones réalise à l'initialisation une discrétisation d'Euler de l'EDO

$$\begin{aligned} H(0) &= Ax, \\ \frac{dH}{ds}(s) &= \mathcal{V}(s) \sigma(\mathcal{W}(s) H(s)), \quad s \in [0,1], \\ F(x) &= B H(1). \end{aligned} \tag{1}$$

D'autres limites possibles à l'initialisation sont étudiées dans Marion *et al.* (2022).

Notre objectif dans ce travail (Marion *et al.*, 2024) est d'étudier le comportement du modèle dans le cas discuté ci-dessus, après entraînement. Nous montrons que les poids du réseau *entraîné* présentent toujours une structure de type EDO. Cette propriété était connue dans le cas des activations linéaires et dans un cadre plus restrictif (Sander *et al.*, 2022). Nous étendons ces résultats à un réseau résiduel non-linéaire assez général, qui se rapproche des réseaux utilisés en pratique. À cette fin, nous faisons l'hypothèse que le réseau est entraîné par flot de gradient, selon les équations d'évolution

$$\frac{\partial V_k}{\partial t}(t) = -L \frac{\partial \hat{\mathcal{R}}_n}{\partial V_k}(t), \quad \frac{\partial W_k}{\partial t}(t) = -L \frac{\partial \hat{\mathcal{R}}_n}{\partial W_k}(t), \quad t \geqslant 0,$$

où $\hat{\mathcal{R}}_n$ désigne un risque empirique. Notons en particulier que la variable temporelle t de l'EDO qui décrit l'évolution des poids n'est pas la même que la variable s de l'EDO neuronale (1) qui décrit la limite en large profondeur.

Notre première contribution est de prouver que la convergence (lorsque L tend vers l'infini) du réseau résiduel vers une EDO neuronale est également valide après entraînement. Cette convergence est valide pour tout temps d'entraînement fini $t \in [0,T]$. Cette propriété est en fait valide dans un cadre beaucoup plus général, dès lors que le réseau de neurones s'écrit comme

$$h_{k+1}^L = h_k^L + \frac{1}{L} f(h_k^L, Z_{k+1}^L), \quad k \in \{0, \dots, L-1\},$$

où $f : \mathbb{R}^q \times \mathbb{R}^p \to \mathbb{R}^q$ est une fonction $\mathcal{C}^2$ telle que $f(0, \cdot) \equiv 0$ and $f(\cdot, z)$ est uniformément Lipschitz pour z dans tout compact.

Néanmoins, la convergence de l'algorithme d'optimisation lorsque T tend vers l'infini n'est pas garantie sans hypothèse supplémentaire, du fait de la non-convexité du problème d'optimisation. Nous prouvons cette convergence grâce à une condition de type Polyak-Łojasiewicz (PL), un outil majeur dans l'analyse des algorithmes d'optimisation pour les réseaux de neurones (Liu *et al.*, 2022). La condition PL implique la convergence du flot de gradient vers un minimum global. Notre seconde contribution est de prouver que cette condition est vérifiée lorsque la largeur q des couches cachées est plus grande qu'une constante fois la taille de l'échantillon n. Cette condition de surparamétrisation est meilleure que les résultats de la littérature, qui requiert soit une surparamétrisation polynomiale, soit des conditions plus restrictives sur les données. D'autres hypothèses légères sont nécessaires, en particulier sur la forme exacte de l'initialisation. Nous obtenons alors la convergence en grande profondeur et en grand temps d'entraînement, c'est-à-dire l'existence de fonctions Lipschitz $\mathcal{V}_\infty$ et $\mathcal{W}_\infty$ telles que le réseau de neurones entraîné converge lorsque L *et* T tendent vers l'infini vers l'EDO

$$\frac{dH}{ds}(s) = \mathcal{V}_\infty(s)\sigma(\mathcal{W}_\infty(s)H(s)), \quad s \in [0,1]. \tag{2}$$

De plus, l'erreur d'entraînement de l'EDO limite est égale à zéro. Cette analyse représente une première étape dans la compréhension de la régularisation implicite du flot de gradient pour les réseaux résiduels, c'est-à-dire la caractérisation des propriétés du réseau entraîné parmi tous les minimiseurs du risque empirique.

Nos résultats théoriques sont complétés par des illustrations expérimentales, qui montrent en particulier qu'il est possible d'apprendre avec un réseau initialisé comme décrit au début de ce document. Nous obtenons ainsi une performance de l'ordre de 80% sur CIFAR-10, et les poids après entraînement réalisent bien une discrétisation d'une fonction lisse (Figure 1 à gauche), c'est-à-dire que le réseau entraîné discrétise bien une EDO.

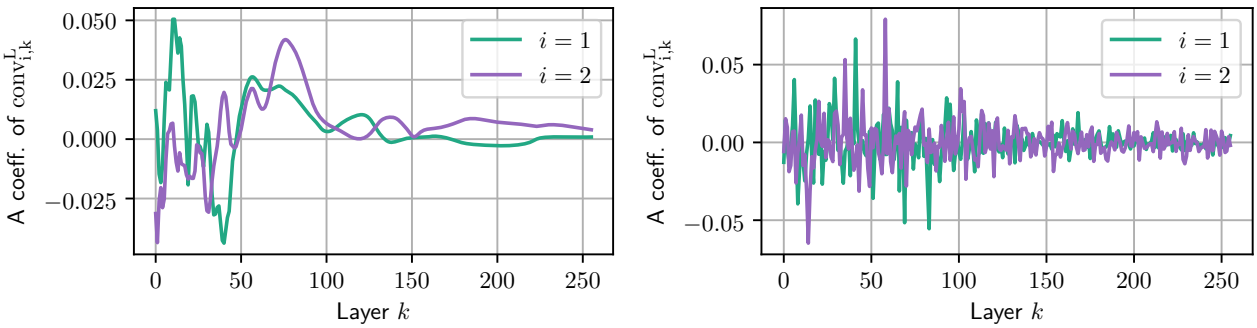

FIGURE 1 – Un coefficient aléatoire des matrices de poids suivi le long de la profondeur du réseau, après entraînement. **Gauche :** L'initialisation lisse des poids conduit à des poids qui discrétisent une fonction lisse après entraînement. **Droite :** En initialisant les poids de manière i.i.d., nous obtenons des poids non lisses après entraînement.

Références

C. LIU, L. ZHU et M. BELKIN : Loss landscapes and optimization in over-parameterized non-linear systems and neural networks. *Applied and Computational Harmonic Analysis*,

59:85–116, 2022. Special Issue on Harmonic Analysis and Machine Learning.

P. Marion, A. Fermanian, G. Biau et J.-P. Vert : Scaling ResNets in the large-depth regime. *arXiv :2206.06929*, 2022.

P. Marion, Y.-H. Wu, M.E. Sander et G. Biau : Implicit regularization of deep residual networks towards neural ODEs. *In International Conference on Learning Representations*, 2024.

M.E. Sander, P. Ablin et G. Peyré : Do residual neural networks discretize neural ordinary differential equations ? *In* I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan et R. Garnett, éditeurs : *Advances in Neural Information Processing Systems*, volume 35, pages 36520–36532. Curran Associates, Inc., 2022.