Projet de thèse
Convergence et biais implicites des réseaux de neurones résiduels

Raphaël Barboni, encadré par G.Peyré et F-X.Vialard

Résumé

Ce sujet de thèse porte sur les propriétés des réseaux de neurones profonds en apprentissage automatique. L'objectif du projet de recherche est de développer des outils à la fois théoriques et numériques pour comprendre les réseaux de neurones résiduels et certifier leurs performances.

1 Contexte : Apprentissage automatique et réseaux résiduels

Apprentissage automatique et réseaux de neurones En mathématique le domaine de l'apprentissage automatique concerne l'étude des procédures automatiques qui permettent à des algorithmes de régression ou de classification d'apprendre des fonctions de prédictions à partir d'exemples annotés (apprentissage supervisé) ou non-annotés (apprentissage non-supervisé). Les réseaux de neurones font par exemple partie des modèles prédictifs les plus couramment utilisés. Ils consistent en la succession de transformations linéaires et non-linéaires simples assemblées en couches successives et dont les paramètres sont "appris" par un algorithme d'optimisation visant à la minimisation d'une fonction de risque empirique associée aux données. Récemment, des modèles de réseaux de neurones "profonds", avec un grand nombre de couches, se sont distingués en battant le niveau de l'état de l'art dans un certain nombre de tâches de classification ou de régression. Ceux-ci ont en particulier un espace de paramètres de dimension importante, typiquement bien plus grande que la dimension du jeu de données d'entraînement. On parle également de modèles "sur-parametrés". L'entraînement des réseaux profonds pose d'abord un problème numérique car, du fait du grand nombre de paramètres à optimiser, la recherche d'un minimiseur de la fonction de risque est une tâche algorithmiquement difficile. Un premier enjeu est donc de concevoir des stratégies d'entraînement efficaces de ces modèles. Les performances des réseaux profonds posent ensuite de nombreuses questions théoriques en remettant en cause le principe statistique, jusque-là bien établi, de compromis "biais-variance" : de telles architectures arrivent à la fois à atteindre un risque empirique faible, souvent presque nul, sur les données d'entraînement et à généraliser sur les données test [17]. Une littérature importante s'est développée autour de la compréhension de ce phénomène [3] en suggérant par exemple la présence de "biais" dans les méthodes d'apprentissage.

Réseaux de neurones résiduels Nous nous intéressons plus particulièrement aux réseaux de neurones dits "résiduels" (*Residual Networks* (ResNets)), dont les couches sont constituées de transformations proches de l'identité [10]. Le choix des ResNets est d'abord motivé par le fait qu'en facilitant l'entraînement des réseaux profonds, ces architectures ont joué un rôle important dans l'émergence des modèles sur-paramétrés et dans l'évolution récente des performances des réseaux de neurones. Les ResNets ont également l'avantage d'offrir un certains nombres de connexions intéressantes avec des outils mathématiques déjà largement étudiés. Ils peuvent par exemple être vu comme l'intégration numérique d'une Équation Différentielle Ordinaire (EDO) [5]. Cette formulation suggère en particulier de s'intéresser au modèle théorique directement donné par l'intégration d'une EDO ("Neural-ODE"), correspondant à un réseau infiniment profond. Le problème d'apprentissage se traduit alors par un problème de contrôle optimal [8] dont la solution dépend de l'espace fonctionnelle des champs de vecteurs admissibles. Concrètement, cet

espace est défini à la fois explicitement par la paramétrisation du modèle ainsi qu'implicitement par la structure métrique considérée sur l'espace des paramètres.

Le projet de thèse Une première partie de ce projet de thèse consiste donc à identifier les espaces fonctionnels pertinents ainsi que les propriétés mathématiques associées aux différentes paramétrisations des ResNets. Ce cadre théorique nous permettra ensuite de mieux comprendre les conditions de convergence ainsi que les biais implicites des algorithmes d'apprentissage lors de l'entraînement des réseaux profonds. Ce travail sera enfin l'occasion de développer des outils numériques avec par exemple pour objectif la création d'une librairie *open-source* pour l'entraînement des ResNets.

2 Organisation de la thèse

Notre projet de recherche a pour but de développer une compréhension mathématique claire des propriétés d'apprentissage et de généralisation des ResNets. Ce programme s'articule autour de trois aspects.

2.1 Partie 1 – Paramétrisation des ResNets

Paramétrisation linéaire : Espaces à Noyaux Reproduisant et flots de difféomorphismes Récemment, un lien a été établi entre de nombreuses instances de ResNets et des algorithmes de recalage d'images [15]. Dans le cadre d'une paramétrisation linéaire, ce travail suggère l'identification de l'espace des paramètres à un espace à un Espace de Hilbert à Noyaux Reproduisant (Reproducing Kernel Hilbert Space (RKHS)) ainsi qu'une formulation du problème d'apprentissage en terme de flot sur un groupe de difféomorphismes munit d'une métrique invariante à droite [16]. Une telle formulation nous a par exemple permis de conclure à un premier résultat de convergence [2]. En plus du fait qu'une importante littérature existe déjà sur les flots de difféomorphismes, cette formulation a l'avantage de proposer une interprétation géométrique claire du problème d'apprentissage. De plus, les RKHS apportent une structure métrique clairement identifiable sur l'espace de champs de vecteurs admissibles et permettent le contrôle explicite de leur régularité. Cette paramétrisation linéaire correspond cependant assez peu aux architectures les plus populaires qui comprennent généralement un nombre important de non-linéarités.

Paramétrisation non-linéaire : Espace de Barron et de Wasserstein D'autres travaux ont par exemple proposé d'utiliser une représentation en termes d'espace de Barron [9], menant à des résultats sur les capacités de généralisation des réseaux. Plus récemment, de nombreux travaux ont également proposé d'exploiter une paramétrisation des réseaux de neurones en termes de mesure sur l'espace des paramètres, associé à une formulation de la phase d'apprentissage en termes flot de gradient Wasserstein sur l'espace des mesures. Si une telle formulation a abouti à de nombreux résultats de convergence pour des réseaux peu profonds [6], elle pourrait également s'avérer intéressante pour l'étude des ResNets. Toutefois, de telles paramétrisations ne permettent pas nécessairement le contrôle explicite de la régularité des paramètres au cours de la phase d'apprentissage.

Création d'une librairie *open-source* Chacune des paramétrisation présentées ci-dessus correspond actuellement en pratique à des implémentations parfois très différentes des ResNets et de leur méthode d'entraînement. Un objectif de notre projet sera donc également de synthétiser ces différentes implémentations avec le développement d'une librairie *open-source* spécialement dédiée à l'entraînement des ResNets.

2.2 Partie 2 – Convergence Globale

Convergence de la descente de gradient pour la minimisation du risque empirique On parle de "convergence globale" lorsque l'algorithme d'apprentissage arrive à atteindre un minimum global de la fonction de risque empirique associée aux données d'entraînement. En pratique, on observe ce phénomène

de convergence pour la descente de gradient (ou descente de gradient stochastique) dans l'entraînement des réseaux de neurones alors que le risque empirique associé est typiquement une fonction fortement non-convexe des paramètres. De nombreux résultats ont récemment réussi à surmonter cette difficulté en montrant que la fonction de risque empirique vérifie certaines inégalités fonctionnelles de type Polyak-Lojasiewicz [12, 4]. La plupart des résultats concernent cependant des réseaux peu profonds à 2 ou 3 couches [7, 1] et peu de résultats existent pour des réseaux profonds. La profondeur apparaît de plus souvent comme un paramètre restrictif qui empêche de généraliser à un nombre arbitraire de couches.

Premier résultat pour des réseaux infiniment profond et extensions Une première contribution de notre travail a consisté à montrer un résultat de convergence pour un modèle de ResNets avec un nombre infini de couches [2]. En s'appuyant sur une paramétrisation linéaire du modèle nous avons pu montrer des inégalités de types Polyak-Lojasiewicz locales pour la fonction de risque empirique associée. Si ces résultats constituent une première étape nous souhaitons poursuivre dans cette direction en les étendant de plusieurs manières : (i) Une première extension serait de considérer une paramétrisation plus générale du modèle, qui ne soit pas restreinte à une paramétrisation linéaire et qui puisse s'appliquer à un grand nombre d'architectures utilisées en pratique. (ii) Une deuxième extension serait de considérer la limite d'un nombre infini de données d'entraînement. L'étude de ce problème permettrait en effet de comprendre la dynamique d'entraînement de modèles génératifs tels que les *Normalizing Flows* [11] et aurait des applications dans le domaine de l'apprentissage non-supervisé. Une difficulté est cependant que ce passage à la limite dégrade la qualité des inégalités fonctionnelles obtenue en appliquant les mêmes méthodes que précédemment. (iii) Enfin les résultats obtenus ne permettent de montrer la convergence que dans un certain voisinage autour du minimum et ne permettent pour l'instant pas d'expliquer le phénomène de convergence globale des réseaux de neurones en toute généralité. Si il est possible de montrer des contre exemple dans des cadre bien précis, nous conjecturons qu'il y a convergence globale pour "presque toutes" les configurations du problème d'apprentissage.

2.3 Partie 3 – Biais Implicite

Régularisation implicite des algorithmes d'apprentissage Un paradoxe des réseaux de neurones profonds est que, du fait de leur grand nombre de paramètres, il existe typiquement de nombreuses configurations minimisant le risque empirique. Ces minimiseurs sont, a priori, susceptibles d'être associés à des performances très variables en termes de généralisation. Une manière de s'assurer que le minimiseur choisi permette au modèle de bien généraliser est par exemple de rajouter explicitement un terme de régularisation de la fonction de risque (e.g. la norme L1 ou L2 des paramètres). Cependant, ces méthodes sont peu utilisées en pratique et les réseaux de neurones sont la plupart du temps entraînés pour la minimisation du risque non-régularisé. Ainsi, les bonnes performances des réseaux entraînés suggèrent l'existence de propriétés de régularisation implicites ou "biais implicites" qui conduisent les méthodes d'optimisation à converger vers des configurations avec de bonnes propriétés de généralisation. Comprendre et identifier ces biais serait donc une contribution conséquente au domaine de l'apprentissage automatique. Si un certain nombre de travaux ont commencé à s'intéresser à cette problématique, les résultats obtenus se limitent pour l'instant à des modèles relativement simples [14].

Régularité induite Dans le cadre des ResNets, une piste que nous souhaitons développer est l'étude de la régularité de la dynamique d'entraînement induite par la régularité initiale de la distribution des données. Par exemple, dans le cas d'une paramétrisation linéaire, une linéarisation du problème autour de l'initialisation suggère de s'intéresser aux comportements de certaines EDPs de diffusion non-locale encore peu étudiées jusqu'à présent. Si certains résultats d'existence et de convergence des solutions existent déjà pour ces équations [13], s'intéresser à la régularité des solutions permettrait une meilleure compréhension de la dynamique d'entraînement et des capacités de généralisation des ResNets.

Références

[1] Z. Allen-Zhu, Y. Li, and Z. Song, *A convergence theory for deep learning via over-parameterization*, in International Conference on Machine Learning, PMLR, 2019, pp. 242–252.

[2] R. Barboni, G. Peyré, and F.-X. Vialard, *Global convergence of resnets : From finite to infinite width using linear parameterization*, arXiv preprint arXiv :2112.05531, (2021).

[3] M. Belkin, D. Hsu, S. Ma, and S. Mandal, *Reconciling modern machine-learning practice and the classical bias–variance trade-off*, Proceedings of the National Academy of Sciences, 116 (2019), pp. 15849–15854.

[4] J. Bolte, A. Daniilidis, O. Ley, and L. Mazet, *Characterizations of Łojasiewicz inequalities : Subgradient flows, talweg, convexity*, Transactions of the American Mathematical Society, 362 (2009), pp. 3319–3363.

[5] R. T. Q. Chen, Y. Rubanova, J. Bettencourt, and D. Duvenaud, *Neural ordinary differential equations*, Advances in Neural Information Processing Systems, (2018).

[6] L. Chizat and F. Bach, *On the global convergence of gradient descent for over-parameterized models using optimal transport*, Advances in Neural Information Processing Systems, 31 (2018), pp. 3036–3046.

[7] S. Du, J. Lee, H. Li, L. Wang, and X. Zhai, *Gradient descent finds global minima of deep neural networks*, in International Conference on Machine Learning, PMLR, 2019, pp. 1675–1685.

[8] W. E, J. Han, and Q. Li, *A mean-field optimal control formulation of deep learning*, Research in the Mathematical Sciences, 6 (2019), p. 10.

[9] W. E, C. Ma, and L. Wu, *The Barron Space and the Flow-Induced Function Spaces for Neural Network Models*, Constructive Approximation, (2021).

[10] K. He, X. Zhang, S. Ren, and J. Sun, *Deep residual learning for image recognition*, in Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 770–778.

[11] I. Kobyzev, S. Prince, and M. Brubaker, *Normalizing Flows : An Introduction and Review of Current Methods*, IEEE Transactions on Pattern Analysis and Machine Intelligence, (2020), pp. 1–1.

[12] S. Lojasiewicz, *Sur les trajectoires du gradient d'une fonction analytique*, Seminari di geometria, 1983 (1982), pp. 115–117.

[13] J. Lu, Y. Lu, and J. Nolen, *Scaling limit of the stein variational gradient descent : The mean field regime*, SIAM Journal on Mathematical Analysis, 51 (2019), pp. 648–671.

[14] B. Neyshabur, *Implicit regularization in deep learning*, arXiv preprint arXiv :1709.01953, (2017).

[15] H. Owhadi, *Do ideas have shape ? Plato's theory of forms as the continuous limit of artificial neural networks*, arXiv :2008.03920 [cs, stat], (2020). arXiv : 2008.03920.

[16] A. Trouvé, *Diffeomorphisms groups and pattern matching in image analysis*, International journal of computer vision, 28 (1998), pp. 213–221.

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