Benign overfitting et régression non-paramétrique adaptative

Julien Chhor[1], Suzanne Sigalla[2] and Alexandre B. Tsybakov[3]

[1] *Toulouse School of Economics, France, julien.chhor@tse-fr.eu*
[2] *CREST/ENSAE, France, suzanne.sigalla@ensae.fr*
[3] *CREST/ENSAE, France, alexandre.tsybakov@ensae.fr*

Résumé. Le benign overfitting est un phénomène contre-intuitif récemment découvert dans le cadre du deep learning. Il a été observé expérimentalement que les réseaux de neurones profonds peuvent dans certains cas parfaitement overfitter des données d'entraînement bruitées, tout en ayant d'excellentes performances de généralisation pour prédire de nouveaux points de données. Cela va à l'encontre du point de vue statistique conventionnel selon lequel il devrait y avoir un compromis nécessaire entre le biais et la variance. Ce papier vise à comprendre le benign overfitting dans le cadre simplifié de la régression non paramétrique. Nous proposons d'utiliser des polynômes locaux pour construire un estimateur de la fonction de régression avec les deux propriétés suivantes. Premièrement, cet estimateur est optimal au sens minimax sur les classes de Hölder. Deuxièmement, il s'agit d'une fonction continue qui interpole l'ensemble des observations avec une grande probabilité. L'élément clé de la construction est l'utilisation de noyaux singuliers. De plus, nous démontrons que l'adaptation à une régularité inconnue est compatible avec le surajustement bénin : nous proposons en effet un autre estimateur interpolant qui atteint l'optimalité minimax de manière adaptative à la régularité de Hölder inconnue. Nos résultats mettent en lumière le fait que, dans le modèle de régression non paramétrique, l'interpolation peut être fondamentalement découplée du compromis biais-variance.

Mots-clés. Régression non-paramétrique, adaptation, benign overfitting, estimateurs par polynômes locaux, agrégation.

Abstract. Benign overfitting is a counter-intuitive phenomenon recently discovered in the context of deep learning. It has been experimentally observed that in certain cases, deep neural networks can perfectly overfit noisy training data, while still achieving excellent generalization performance for predicting new data points. This goes against the conventional statistical viewpoint which posits that there should be a necessary tradeoff between bias and variance. This paper aims to understand benign overfitting in the simplified setting of nonparametric regression. We propose using local polynomials to construct an estimator of the regression function with the following two properties. First, this estimator is minimax-optimal over Hölder classes. Second, it is a continuous function that interpolates the set of observations with high probability. The key element of the construction is the use of singular kernels. Moreover, we demonstrate that adaptation to unknown smoothness is compatible with benign overfitting: indeed, we propose another interpolating estimator that achieves minimax optimality adaptively to the unknown Hölder smoothness. Our results highlight that in the nonparametric regression model, interpolation can be fundamentally decoupled from the bias-variance tradeoff.

Keywords. Nonparametric regression, adaptation, benign overfitting, local polynomial estimators, aggregation.

1 Texte long

Benign overfitting is a counter-intuitive phenomenon recently discovered in the deep-learning community. Empirically, deep neural networks can conciliate two seemingly conflicting properties: 1) perfect *overfitting* (namely, zero error when predicting a data point in the training set) and 2) excellent prediction accuracy outside of the training data Zhang et al. (2021). Overfitting was previously believed to deteriorate statistical performance. Given neural networks' ubiquity, identifying contexts where the phenomenon of benign overfitting can occur is of practical interest. In this paper (Chhor et al., 2024), we study this phenomenon in the simplified setting of non-parametric regression, defined below.

State of the art: Benign overfitting has not yet been explained in the context of deep neural networks, and it was mostly studied in simplified statistical settings: linear regression Bartlett et al. (2020), kernel regression Liang et al. (2020) and more recently, in non-parametric regression Belkin et al. (2019). The only paper examining benign overfitting with non-asymptotic guarantees was Belkin et al. (2019), considering only the case $\beta \in (0, 1]$ and $\beta \in (1, 2]$ under additional smoothness constraints on the density of the design.

Model and contributions: For $n, d \geq 1$, assume that we observe i.i.d. random pairs $(X_i, Y_i) \in \mathbb{R}^d \times \mathbb{R}, i = 1, \dots, n$ such that

$$\forall i \in \{1, \dots, n\} : Y_i = f(X_i) + \xi_i$$

where the unknown link function f to estimate belongs to a suitably defined Hölder class of functions with smoothness $\beta > 0$ and Hölder constant $L > 0$, denoted as $\Sigma(\beta, L)$. The design random variables X_i's are i.i.d. with density p that has a convex compact support and such that $0 < c \leq p(\cdot) \leq C < \infty$ over the support of p for some $c, C > 0$, and the noise random variables $(\xi_i)_i$ are i.i.d. with a finite moment of order $2 + \delta$ for some $\delta > 0$.

The goal is to construct an estimator $\widehat{f}$ of f that interpolates the training dataset $(X_i, Y_i)_i$ with high probability (i.e. satisfying $\forall i \in \{1, \dots, n\} : \hat{f}(X_i) = Y_i$), while being minimax optimal for the quadratic risk over the Hölder class with smoothness $\beta > 0$:

$$\sup_{f \in \Sigma(\beta, L)} \mathbb{E}\left[\|\hat{f}(X) - f(X)\|^2_{L^2(X)}\right] \leq n^{-2\beta/(2\beta+d)}.$$

We propose a local polynomial estimator of order β that achieves the two requirements, thereby showing that benign overfitting is compatible with minimax optimality over the whole scale of Hölder classes. The key element of the construction is the use of a singular kernel.

Then, we use an aggregation technique to obtain a second estimator attaining this optimal rate adaptively to the unknown smoothness $\beta > 0$.

Along the way, we strengthen the theory of local polynomial estimators, by proving that they can produce optimal estimators under much milder assumptions than previously known conditions.

Our results highlight that in nonparametric regression, interpolation can be fundamentally decoupled from the bias-variance tradeoff.

References

Bartlett, P. L., Long, P. M., Lugosi, G., and Tsigler, A. (2020). Benign overfitting in linear regression. *Proceedings of the National Academy of Sciences*, 117(48):30063–30070.

Belkin, M., Rakhlin, A., and Tsybakov, A. B. (2019). Does data interpolation contradict statistical optimality? In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 1611–1619. PMLR.

Chhor, J., Sigalla, S., and Tsybakov, A. B. (2024). Benign overfitting and adaptive nonparametric regression. *Probability Theory and Related Fields (to appear); arXiv preprint arXiv:2206.13347*.

Liang, T., Rakhlin, A., and Zhai, X. (2020). On the multiple descent of minimum-norm interpolants and restricted lower isometry of kernels. In *Conference on Learning Theory*, pages 2683–2711. PMLR.

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