Retraction note for "Nonparametric Regression with Shallow Overparameterized Neural Networks Trained by GD with Early Stopping"

Ilja Kuzborskij ILJAK@DEEPMIND.COM
DeepMind, London

Csaba Szepesvári SZEPI@DEEPMIND.COM
DeepMind, Canada and University of Alberta, Edmonton

Editors: Mikhail Belkin and Samory Kpotufe

Abstract

Theorems 1 and 2 appearing in the paper contain a shared critical issue and this note serves as a retraction note. The issue is traced to the Theorem 4, which concerns the control of the Lipschitzness of a Gradient Descent (GD)-trained overparameterized shallow neural network in the input. In the following we describe the issue in some detail.

Setting. We first briefly recall the setting the paper. The paper considers the problem of nonparametric regression with random design where inputs $\boldsymbol{X}, \boldsymbol{X}_1, \dots, \boldsymbol{X}_n \sim P_X \in \mathcal{M}_1(\mathbb{S}^{d-1})$ are i.i.d. and targets are generated by the Lipschitz (possibly non-differentiable) and bounded regression function $f^\star$ on $\mathbb{S}^{d-1}$, as $Y_i = f^\star(\boldsymbol{X}_i) + \varepsilon_i$, with σ-subgaussian independent noise $\varepsilon_1, \dots, \varepsilon_n$. The paper considers learning of $f^\star$ by the shallow neural network predictor

$$\hat{f}_{\boldsymbol{W}}(\boldsymbol{x}) = \sum_{k=1}^{m} u_k \phi\left(\boldsymbol{w}_k^\top \boldsymbol{x}\right) , \qquad (\boldsymbol{x} \in \mathbb{S}^{d-1},\ \boldsymbol{W} \in \mathbb{R}^{d\times m})$$

defined with respect to a fixed twice differentiable activation function $\phi : \mathbb{R} \to \mathbb{R}$, where ϕ', ϕ'' are bounded on $\mathbb{R}$. The predictor is parameterized by an output layer, a (non-tunable) random weight vector $\boldsymbol{u} \overset{\text{iid}}{\sim} \text{unif}\left(\{\pm 1/\sqrt{m}\}\right)^m$ (sampled once at the beginning of the training), and a tunable hidden layer weight matrix $\boldsymbol{W} \in \mathbb{R}^{d\times m}$, where m is the width of the network. In particular, weights of a hidden layer are obtained by the standard GD procedure by approximately minimizing the empirical risk

$$\hat{L}(\boldsymbol{W}) = \frac{1}{n}\sum_{i=1}^{n} (\hat{f}_{\boldsymbol{W}}(\boldsymbol{X}_i) - Y_i)^2 .$$

Here, $\boldsymbol{W}_0$ is obtained by drawing each entry from $\mathcal{N}(0, \nu_{\text{init}}^2)$ independently from each other, while $\boldsymbol{W}_T$ is obtained by the recursive update rule $\boldsymbol{W}_{t+1} = \boldsymbol{W}_t - \eta \nabla \hat{L}(\boldsymbol{W}_t)$ for $t = 0, \dots, T-1$ and where $\eta > 0$ is a fixed step size.

Claimed result and proof sketch. The aim of the paper is to control the excess risk defined as

$$\mathcal{E}_T = \mathbb{E}\left[(f_{\boldsymbol{W}_T}(\boldsymbol{X}) - f^\star(\boldsymbol{X}))^2 \mid \boldsymbol{u}, \boldsymbol{W}_0\right] .$$

In particular, for $\sigma = 0$[1] the paper claims a minimax optimal rate with high probability over $(\boldsymbol{u}, \boldsymbol{W}_0)$, assuming that the network is overparameterized, $m \geq \text{poly}\left((n/d)(1+\nu_{\text{init}}^2)\right)$:

$$\mathcal{E}_T = \mathcal{O}_{\mathbb{P}}\left((\text{Lip}(f^\star)^2 + (1+d\nu_{\text{init}}^2)^2) n^{-\frac{2}{2+d}}\right) \quad \text{as} \quad n \to \infty ,$$

which is based on the following proof idea. Introduce a nearest-neighbor operator (with ties broken arbitrarily)

$$\pi(\boldsymbol{x}) = \operatorname*{arg\,min}_{i \in [n]} \|\boldsymbol{x} - \boldsymbol{X}_i\|_2 , \qquad (\boldsymbol{x} \in \mathbb{S}^{d-1})$$

and consider a decomposition of the excess risk (here $\mathbb{E}[\cdot] = \mathbb{E}[\cdot \mid \boldsymbol{u}, \boldsymbol{W}_0]$):

$$\underbrace{\mathbb{E}\left(f^\star(\boldsymbol{X}) - f^\star(\boldsymbol{X}_{\pi(\boldsymbol{X})})\right)^2}_{(i)} + \underbrace{\mathbb{E}\left(f^\star(\boldsymbol{X}_{\pi(\boldsymbol{X})}) - \hat{f}_{\boldsymbol{W}_T}(\boldsymbol{X}_{\pi(\boldsymbol{X})})\right)^2}_{(ii)} + \underbrace{\mathbb{E}\left(\hat{f}_{\boldsymbol{W}_T}(\boldsymbol{X}_{\pi(\boldsymbol{X})}) - \hat{f}_{\boldsymbol{W}_T}(\boldsymbol{X})\right)^2}_{(iii)} .$$

Here term (i) is controlled thanks to the Lipschitzness of $f^\star$ and the standard partitioning analysis of the one nearest-neighbor rule (Lemma 7 in the paper) gives

$$\mathbb{E}\,\|\boldsymbol{X} - \boldsymbol{X}_{\pi(\boldsymbol{X})}\|_2^2 \lesssim n^{-\frac{2}{2+d}} .$$

Term (ii) is the expected empirical risk $\mathbb{E}\,\hat{L}(\boldsymbol{W}_T)$, which is controlled by the convergence of optimization error of GD to zero for shallow overparameterized neural networks (Du et al., 2018; Oymak and Soltanolkotabi, 2020).

Issue with term (iii). The issue occurs in the control of term (iii) which captures Lipschitzness of the predictor trained for T steps. Before discussing the issue we need to introduce some technical notions. In particular, for the analysis we require the Neural Tangent Random Feature (NTRF) predictor, which is obtained by linearization of $\boldsymbol{W} \mapsto \hat{f}_{\boldsymbol{W}}(\boldsymbol{x})$ around $\boldsymbol{W}_0$, namely

$$\hat{f}^{\text{rf}}_{\boldsymbol{W}}(\boldsymbol{x}) = \hat{f}_{\boldsymbol{W}_0}(\boldsymbol{x}) + \boldsymbol{\Psi}(\boldsymbol{x})^\top \text{vec}(\boldsymbol{W} - \boldsymbol{W}_0) ,$$

where $\boldsymbol{\Psi}(\cdot)$ is called the NTRF feature map and defined as

$$\boldsymbol{\Psi}(\boldsymbol{x}) = [u_1 \phi'(\boldsymbol{W}_{0,1}^\top \boldsymbol{x})\boldsymbol{x}^\top, \ldots, u_m \phi'(\boldsymbol{W}_{0,m}^\top \boldsymbol{x})\boldsymbol{x}^\top]^\top .$$

In addition, we introduce NTRF iterates $(\boldsymbol{W}_t^{\text{rf}})_{t=0}^T$ obtained by running GD as discussed in the beginning of the note, however with the empirical risk replaced by the least-squares objective $\boldsymbol{W} \mapsto ((\hat{f}^{\text{rf}}_{\boldsymbol{W}}(\boldsymbol{X}_1) - Y_1)^2 + \cdots + (\hat{f}^{\text{rf}}_{\boldsymbol{W}}(\boldsymbol{X}_n) - Y_n)^2)/n$.

The attractive property of the NTRF iterates is that the distance $\|\boldsymbol{W}_T - \boldsymbol{W}_T^{\text{rf}}\|_F$ can be made small by overparameterization of the neural network thanks to the so-called 'coupling' argument, see for instance (Bartlett et al., 2021, Theorem 5.1) and Appendix B in the paper we discuss. Then,

1. In this note we only consider a noise-free setting, but the same issue applies for the case $\sigma > 0$.

assuming that in addition to the boundedness of derivatives, activation function ϕ also satisfies $\phi'(x) \le B_{\phi''}|x|$ with $\sup_{x\in\mathbb{R}}|\phi''(x)| \le B_{\phi''}$ for all $x \in \mathbb{R}$, we attempt to control the Lipschitzness:

$$\begin{aligned}
\mathrm{Lip}(\hat{f}_{\boldsymbol{W}_T}) &= \sup_{\boldsymbol{x}\in\mathbb{S}^{d-1}} \left\| \sum_{k=1}^{m} u_k \phi'(\boldsymbol{W}_{T,k}^\top \boldsymbol{x}) \boldsymbol{W}_{T,k} \right\|_2 \\
&\le \frac{1}{\sqrt{m}} \sup_{\boldsymbol{x}\in\mathbb{S}^{d-1}} \sum_{k=1}^{m} |\phi'(\boldsymbol{W}_{T,k}^\top \boldsymbol{x})| \|\boldsymbol{W}_{T,k}\|_2 \\
&\le \frac{B_{\phi''}}{\sqrt{m}} \sum_{k=1}^{m} \|\boldsymbol{W}_{T,k}\|_2^2 \\
&\le \frac{2B_{\phi''}}{\sqrt{m}} \underbrace{\|\boldsymbol{W}_T - \boldsymbol{W}_T^{\mathrm{rf}}\|_F^2}_{(a)} + \frac{2B_{\phi''}}{\sqrt{m}} \underbrace{\|\boldsymbol{W}_T^{\mathrm{rf}}\|_F^2}_{(b)} .
\end{aligned}$$

While (a) is controlled by the aforementioned 'coupling' argument, it turns out that $\|\boldsymbol{W}_T^{\mathrm{rf}}\|_F^2$ cannot be made small enough. The problem comes in assuming that $\boldsymbol{W}_0^{\mathrm{rf}} = \boldsymbol{0}$ while it must be $\boldsymbol{W}_0^{\mathrm{rf}} = \boldsymbol{W}_0$ for the coupling argument of (Bartlett et al., 2021, Theorem 5.1) to hold. Then, having $\boldsymbol{W}_0^{\mathrm{rf}} = \boldsymbol{W}_0$, it is clear that $\|\boldsymbol{W}_T^{\mathrm{rf}}\|_F^2 \lesssim \|\boldsymbol{W}_0\|_F^2$ while $\|\boldsymbol{W}_0\|_F^2 = \mathcal{O}(m)$ which renders the above bound on the Lipschitz constant of order $\mathcal{O}(\sqrt{m})$.

Prospects on fixing the issue. One potential resolution to the above could be to follow a more careful analysis and replace $\|\boldsymbol{W}_T^{\mathrm{rf}}\|_F^2$ with $\|\boldsymbol{W}_T^{\mathrm{rf}} - \boldsymbol{W}_0\|_F^2$, where the later quantity is definitely no larger since GD takes the shortest ℓ^2 path from the initialization for least-squares problems of the considered kind (Oymak and Soltanolkotabi, 2020).[2] Such an alternative analysis of term (iii) would be based upon a decomposition (for any $\boldsymbol{x}, \tilde{\boldsymbol{x}} \in \mathbb{S}^{d-1}$),

$$\hat{f}_{\boldsymbol{W}_T}(\boldsymbol{x}) - \hat{f}_{\boldsymbol{W}_T}(\tilde{\boldsymbol{x}}) = \underbrace{\hat{f}_{\boldsymbol{W}_T}(\boldsymbol{x}) - \hat{f}_{\boldsymbol{W}_T^{\mathrm{rf}}}(\boldsymbol{x})}_{(iii.a)} + \underbrace{\hat{f}_{\boldsymbol{W}_T^{\mathrm{rf}}}(\boldsymbol{x}) - \hat{f}_{\boldsymbol{W}_T^{\mathrm{rf}}}(\tilde{\boldsymbol{x}})}_{(iii.b)} + \underbrace{\hat{f}_{\boldsymbol{W}_T^{\mathrm{rf}}}(\tilde{\boldsymbol{x}}) - \hat{f}_{\boldsymbol{W}_T}(\tilde{\boldsymbol{x}})}_{(iii.c)} .$$

Here terms $(iii.a)$ and $(iii.c)$ could be controlled thanks to the 'coupling' argument as mentioned above. It is not hard to see that

$$\begin{aligned}
(iii.b) &= (\hat{f}_{\boldsymbol{W}_0}(\boldsymbol{x}) - \hat{f}_{\boldsymbol{W}_0}(\tilde{\boldsymbol{x}})) + \mathrm{vec}(\boldsymbol{W} - \boldsymbol{W}_0)^\top (\boldsymbol{\Psi}(\boldsymbol{x}) - \boldsymbol{\Psi}(\tilde{\boldsymbol{x}})) \\
&\le c\|\boldsymbol{x} - \tilde{\boldsymbol{x}}\|_2 + \|\boldsymbol{W}_T^{\mathrm{rf}} - \boldsymbol{W}_0\|_F \|\boldsymbol{\Psi}(\boldsymbol{x}) - \boldsymbol{\Psi}(\tilde{\boldsymbol{x}})\|_2 \\
&\le c\|\boldsymbol{x} - \tilde{\boldsymbol{x}}\|_2 + c'\|\boldsymbol{W}_T^{\mathrm{rf}} - \boldsymbol{W}_0\|_F \|\boldsymbol{x} - \tilde{\boldsymbol{x}}\|_2 \qquad (\boldsymbol{\Psi}(\cdot) \text{ is Lipschitz})
\end{aligned}$$

with high probability over $(\boldsymbol{u}, \boldsymbol{W}_0)$, where using independence of $(\boldsymbol{u}, \boldsymbol{W}_0)$, $\hat{f}_{\boldsymbol{W}_0}(\cdot)$ is Lipschitz with a universal constant c. Now, it remains to see whether $\|\boldsymbol{W}_T^{\mathrm{rf}} - \boldsymbol{W}_0\|_F$ can be controlled. To this end let $(\lambda_i)_{i=1}^n$ be the eigenvalues of the covariance matrix $\boldsymbol{\Psi}(\boldsymbol{X}_1)\boldsymbol{\Psi}(\boldsymbol{X}_1)^\top + \cdots + \boldsymbol{\Psi}(\boldsymbol{X}_n)\boldsymbol{\Psi}(\boldsymbol{X}_n)^\top$ and let $\boldsymbol{Y}^\star = [f^\star(\boldsymbol{X}_1), \ldots, f^\star(\boldsymbol{X}_n)]^\top$. The standard analysis of GD dynamics (Yao et al., 2007) leads to the identity

$$\|\boldsymbol{W}_T^{\mathrm{rf}} - \boldsymbol{W}_0\|_F^2 = \sum_{i=1}^{n} (1 - (1 - \tfrac{2\eta}{n}\lambda_i)^T)^2 \, \frac{(\boldsymbol{u}_i^\top \boldsymbol{Y}^\star)^2}{\lambda_i} .$$

2. The quantity $\|\boldsymbol{W}_T^{\mathrm{rf}} - \boldsymbol{W}_0\|_F^2$ was considered before in the analysis of GD-trained shallow neural networks as a complexity measure of a GD solution, for instance (Arora et al., 2019).

Now, since we focus on a noise-free setting ($\sigma = 0$), we can consider the limiting solution with $T \to \infty$ (no need to control overfitting). To this end,

$$\|\boldsymbol{W}_T^{\mathrm{rf}} - \boldsymbol{W}_0\|_F^2 \leq \boldsymbol{Y}^{\star\top} \hat{\boldsymbol{K}}^{-1} \boldsymbol{Y}^{\star} \tag{$*$}$$

where $\hat{\boldsymbol{K}} \in \mathbb{R}^{n\times n}$ is a Gram matrix with entries $(\hat{\boldsymbol{K}})_{i,j} = (\boldsymbol{\Psi}(\boldsymbol{X}_i)^\top \boldsymbol{\Psi}(\boldsymbol{X}_j))$. Thus, to have $(iii.b)$ of the right order, we need to have $(*) = \tilde{\mathcal{O}}(1)$ as $n \to \infty$, however, it is unclear how to analyze such a quantity. Arora et al. (2019) used some Reproducing kernel Hilbert space (RKHS) arguments and considered a closely related quantity $\boldsymbol{Y}^{\star\top} \boldsymbol{K}^{-1} \boldsymbol{Y}^{\star}$, where $\boldsymbol{K}$ is a kernel matrix of a reproducing kernel function $\kappa(\boldsymbol{x}, \tilde{\boldsymbol{x}}) = \mathbb{E}[\boldsymbol{\Psi}(\boldsymbol{x})^\top \boldsymbol{\Psi}(\tilde{\boldsymbol{x}})]$.[3] While, they have established some interesting constant-order upper bounds on $\boldsymbol{Y}^{\star\top} \boldsymbol{K}^{-1} \boldsymbol{Y}^{\star}$ by considering specific cases of $f^\star$, all of the considered cases were differentiable and not necessarily Lipschitz, which violates the setting of our paper.

Acknowledgements

We thank Di Wu for drawing our attention to the issue.

References

S. Arora, S. Du, W. Hu, Z. Li, and R. Wang. Fine-grained analysis of optimization and generalization for overparameterized two-layer neural networks. In *International Conference on Machine Learing (ICML)*, 2019.

P. L. Bartlett, A. Montanari, and A. Rakhlin. Deep learning: a statistical viewpoint. *Acta Numerica*, 2021. URL https://arxiv.org/abs/2103.09177. To appear.

S. S. Du, X. Zhai, B. Poczos, and A. Singh. Gradient descent provably optimizes over-parameterized neural networks. In *International Conference on Learning Representations (ICLR)*, 2018.

S. Oymak and M. Soltanolkotabi. Towards moderate overparameterization: global convergence guarantees for training shallow neural networks. *IEEE Journal on Selected Areas in Information Theory*, 2020.

J. A. Tropp. An introduction to matrix concentration inequalities. *Foundations and Trends® in Machine Learning*, 8(1-2):1–230, 2015.

Y. Yao, L. Rosasco, and A. Caponnetto. On early stopping in gradient descent learning. *Constructive Approximation*, 26(2):289–315, 2007.

3. Some standard matrix concentration inequalities (Tropp, 2015) allow to say that $(*)$ is close to $\boldsymbol{Y}^{\star\top} \boldsymbol{K}^{-1} \boldsymbol{Y}^{\star}$ whenever $m \geq \mathrm{poly}(n/\lambda_{\min}(\boldsymbol{K}))$.