Supplementary material to "On the universal consistency of an over-parametrized deep neural network estimate learned by gradient descent"

Selina Drews and Michael Kohler

Auxiliary results for the proof of Theorem 1

Lemma 6 *Let $\sigma : \mathbb{R} \to \mathbb{R}$ be bounded and differentiable, and assume that its derivative is bounded. Let $\alpha_n \geq 1$, $t_n \geq L_n$, $\gamma_n^* \geq 1$, $B_n \geq 1$, $r \geq 2d$,*

$$|w_{1,1,k}^{(L)}| \leq \gamma_n^* \quad (k = 1, \dots, K_n), \tag{64}$$

$$|w_{k,i,j}^{(l)}| \leq B_n \quad \text{for } l = 1, \dots, L-1 \tag{65}$$

and

$$\|\mathbf{w} - \mathbf{v}\|_\infty^2 \leq \frac{2t_n}{L_n} \cdot \max\{F_n(\mathbf{v}), 1\}. \tag{66}$$

Then we have

$$\|(\nabla_{\mathbf{w}} F_n)(\mathbf{w})\| \leq c_{44} \cdot K_n^{3/2} \cdot B_n^{2L} \cdot (\gamma_n^*)^2 \cdot \alpha_n^2 \cdot \sqrt{\frac{t_n}{L_n} \cdot \max\{F_n(\mathbf{v}), 1\}}.$$

Proof. We have

$$\begin{aligned}
&\|\nabla_{\mathbf{w}} F_n(\mathbf{w})\|^2 \\
&= \sum_{k,i,j,l} \left(\frac{2}{n} \sum_{s=1}^n (Y_s - f_{\mathbf{w}}(X_s)) \cdot 1_{[-\alpha_n,\alpha_n]^d}(X_s) \cdot \frac{\partial f_{\mathbf{w}}}{\partial w_{k,i,j}^{(l)}}(X_s) \right. \\
&\qquad \left. + \frac{\partial}{\partial w_{k,i,j}^{(l)}} \left(c_2 \cdot \sum_{r=1}^{K_n} |w_{1,1,r}^{(L)}|^2 \right) \right)^2 \\
&\leq 8 \cdot \sum_{k,i,j,l} \frac{1}{n} \sum_{s=1}^n (Y_s - f_{\mathbf{w}}(X_s))^2 \cdot 1_{[-\alpha_n,\alpha_n]^d}(X_s) \cdot \left(\frac{\partial f_{\mathbf{w}}}{\partial w_{k,i,j}^{(l)}}(X_s) \right)^2 \\
&\qquad + 8 \cdot c_2^2 \cdot K_n \cdot (\gamma_n^*)^2 \\
&\leq c_{45} \cdot K_n \cdot L \cdot r^2 \cdot d \cdot \max_{k,i,j,l,s} \left(\frac{\partial f_{\mathbf{w}}}{\partial w_{k,i,j}^{(l)}}(X_s) \right)^2 \cdot 1_{[-\alpha_n,\alpha_n]^d}(X_s) \\
&\qquad \cdot \frac{1}{n} \sum_{s=1}^n (Y_s - f_{\mathbf{w}}(X_s))^2 \cdot 1_{[-\alpha_n,\alpha_n]^d}(X_s) + 8 \cdot c_2^2 \cdot K_n \cdot (\gamma_n^*)^2.
\end{aligned}$$

The chain rule implies

$$\frac{\partial f_{\mathbf{w}}}{\partial w_{k,i,j}^{(l)}}(x) = \sum_{s_{l+2}=1}^r \cdots \sum_{s_{L-1}=1}^r f_{k,j}^{(l)}(x) \cdot \sigma' \left(\sum_{t=1}^r w_{k,i,t}^{(l)} \cdot f_{k,t}^{(l)}(x) + w_{k,i,0}^{(l)} \right)$$

$$\cdot w^{(l+1)}_{k,s_{l+2},i} \cdot \sigma' \left(\sum_{t=1}^{r} w^{(l+1)}_{k,s_{l+2},t} \cdot f^{(l+1)}_{k,t}(x) + w^{(l+1)}_{k,s_{l+2},0} \right) \cdot w^{(l+2)}_{k,s_{l+3},s_{l+2}}$$
$$\cdot \sigma' \left(\sum_{t=1}^{r} w^{(l+2)}_{k,s_{l+3},t} \cdot f^{(l+2)}_{k,t}(x) + w^{(l+2)}_{k,s_{l+3},0} \right) \cdots w^{(L-2)}_{k,s_{L-1},s_{L-2}}$$
$$\cdot \sigma' \left(\sum_{t=1}^{r} w^{(L-2)}_{k,s_{L-1},t} \cdot f^{(L-2)}_{k,t}(x) + w^{(L-2)}_{k,s_{L-1},0} \right) \cdot w^{(L-1)}_{k,1,s_{L-1}}$$
$$\cdot \sigma' \left(\sum_{t=1}^{r} w^{(L-1)}_{k,1,t} \cdot f^{(L-1)}_{k,t}(x) + w^{(L-1)}_{k,1,0} \right) \cdot w^{(L)}_{1,1,k}, \tag{67}$$

where we have used the abbreviations

$$f^{(0)}_{k,j}(x) = \begin{cases} x^{(j)} & \text{if } j \in \{1, \dots, d\} \\ 1 & \text{if } j = 0 \end{cases}$$

and

$$f^{(l)}_{k,0}(x) = 1 \quad (l = 1, \dots, L-1).$$

Using the assumptions of Lemma 6 we can conclude

$$\max_{k,i,j,l,s} \left(\frac{\partial f_{\mathbf{w}}}{\partial w^{(l)}_{k,i,j}}(X_s) \right)^2 \cdot 1_{[-\alpha_n,\alpha_n]^d}(X_s) \leq c_{45} \cdot r^{2L} \cdot \max\{\|\sigma'\|^{2L}_\infty, 1\} \cdot B_n^{2L} \cdot (\gamma_n^*)^2 \cdot \alpha_n^2.$$

By the Lipschitz continuity of σ together with the assumptions of Lemma 6 we get for any $x \in [-\alpha_n, \alpha_n]^d$

$$|f_{\mathbf{w}}(x) - f_{\mathbf{v}}(x)| \leq 2 \cdot K_n \cdot \max\{\|\sigma'\|^L_\infty, 1\} \cdot \gamma_n^* \cdot (2r+1)^L \cdot B_n^L \cdot \alpha_n \cdot \max\{\|\sigma\|_\infty, 1\} \cdot \|\mathbf{w} - \mathbf{v}\|_\infty.$$

(cf., e.g., Lemma 5 in Kohler and Krzyżak (2021) for a related proof). This implies

$$\begin{aligned} &\frac{1}{n} \sum_{s=1}^{n} (Y_s - f_{\mathbf{w}}(X_s))^2 \cdot 1_{[-\alpha_n,\alpha_n]^d}(X_s) \\ &\leq 2 \cdot F_n(\mathbf{v}) + \frac{2}{n} \sum_{s=1}^{n} (f_{\mathbf{v}}(X_s) - f_{\mathbf{w}}(X_s))^2 \cdot 1_{[-\alpha_n,\alpha_n]^d}(X_s) \\ &\leq 2 \cdot F_n(\mathbf{v}) + 8 \cdot \max\{\|\sigma'\|^{2L}_\infty, 1\} \cdot K_n^2 \cdot \gamma_n^{*2} \cdot (2r+1)^{2L} \cdot B_n^{2L} \cdot \alpha_n^2 \cdot \max\{\|\sigma\|_\infty, 1\}^2 \\ &\quad \cdot \frac{2t_n}{L_n} \cdot \max\{F_n(\mathbf{v}), 1\}. \end{aligned}$$

Summarizing the above results, the proof is complete. □

Lemma 7 *Let $\sigma : \mathbb{R} \to \mathbb{R}$ be bounded and differentiable, and assume that its derivative is Lipschitz continuous and bounded. Let $\alpha_n \geq 1$, $t_n \geq L_n$, $\gamma_n^* \geq 1$, $B_n \geq 1$, $r \geq 2d$ and assume*

$$|\max\{(\mathbf{w}_1)^{(L)}_{1,1,k}, (\mathbf{w}_2)^{(L)}_{1,1,k}\}| \leq \gamma_n^* \quad (k = 1, \dots, K_n), \tag{68}$$

$$|\max\{(\mathbf{w}_1)^{(l)}_{k,i,j},(\mathbf{w}_2)^{(l)}_{k,i,j}\}|\leq B_n \quad for\ l=1,\dots,L-1 \tag{69}$$

and

$$\|\mathbf{w}_2-\mathbf{v}\|^2\leq 8\cdot\frac{t_n}{L_n}\cdot\max\{F_n(\mathbf{v}),1\}. \tag{70}$$

Then we have

$$\|(\nabla_{\mathbf{w}}F_n)(\mathbf{w}_1)-(\nabla_{\mathbf{w}}F_n)(\mathbf{w}_2)\|$$
$$\leq c_{46}\cdot\max\{\sqrt{F_n(\mathbf{v})},1\}\cdot(\gamma_n^*)^2\cdot B_n^{3L}\cdot\alpha_n^3\cdot K_n^{3/2}\cdot\sqrt{\frac{t_n}{L_n}}\cdot\|\mathbf{w}_1-\mathbf{w}_2\|.$$

Proof. We have

$$\|\nabla_{\mathbf{w}}F_n(\mathbf{w}_1)-\nabla_{\mathbf{w}}F_n(\mathbf{w}_2)\|^2$$
$$=\sum_{k,i,j,l}\Bigg(\frac{2}{n}\sum_{s=1}^n(Y_s-f_{\mathbf{w}_1}(X_s))\cdot 1_{[-\alpha_n,\alpha_n]^d}(X_s)\cdot\frac{\partial f_{\mathbf{w}_1}}{\partial w^{(l)}_{k,i,j}}(X_s)$$
$$+\frac{\partial}{\partial w^{(l)}_{k,i,j}}\left(c_2\cdot\sum_{r=1}^{K_n}|(\mathbf{w}_1)^{(L)}_{1,1,r}|^2\right)$$
$$-\sum_{k,i,j,l}\Bigg(\frac{2}{n}\sum_{s=1}^n(Y_s-f_{\mathbf{w}_2}(X_s))\cdot 1_{[-\alpha_n,\alpha_n]^d}(X_s)\cdot\frac{\partial f_{\mathbf{w}_2}}{\partial w^{(l)}_{k,i,j}}(X_s)$$
$$+\frac{\partial}{\partial w^{(l)}_{k,i,j}}\left(c_2\cdot\sum_{r=1}^{K_n}|(\mathbf{w}_2)^{(L)}_{1,1,r}|^2\right)\Bigg)^2$$
$$\leq 16\cdot\sum_{k,i,j,l}\left(\frac{1}{n}\sum_{s=1}^n(f_{\mathbf{w}_2}(X_s)-f_{\mathbf{w}_1}(X_s))\cdot 1_{[-\alpha_n,\alpha_n]^d}(X_s)\cdot\frac{\partial f_{\mathbf{w}_1}}{\partial w^{(l)}_{k,i,j}}(X_s)\right)^2$$
$$+16\cdot\sum_{k,i,j,l}\Bigg(\frac{1}{n}\sum_{s=1}^n(Y_s-f_{\mathbf{w}_2}(X_s))\cdot 1_{[-\alpha_n,\alpha_n]^d}(X_s)$$
$$\cdot\left(\frac{\partial f_{\mathbf{w}_1}}{\partial w^{(l)}_{k,i,j}}(X_s)-\frac{\partial f_{\mathbf{w}_2}}{\partial w^{(l)}_{k,i,j}}(X_s)\right)\Bigg)^2$$
$$+8\cdot c_2^2\cdot\|\mathbf{w}_1-\mathbf{w}_2\|^2$$
$$\leq 16\cdot\sum_{k,i,j,l}\max_{s=1,\dots,n}\left(\frac{\partial f_{\mathbf{w}_1}}{\partial w^{(l)}_{k,i,j}}(X_s)\right)^2\cdot 1_{[-\alpha_n,\alpha_n]^d}(X_s)$$
$$\cdot\frac{1}{n}\sum_{s=1}^n(f_{\mathbf{w}_2}(X_s)-f_{\mathbf{w}_1}(X_s))^2\cdot 1_{[-\alpha_n,\alpha_n]^d}(X_s)$$
$$+16\cdot\frac{1}{n}\sum_{s=1}^n(Y_s-f_{\mathbf{w}_2}(X_s))^2\cdot 1_{[-\alpha_n,\alpha_n]^d}(X_s)$$

$$\cdot \sum_{k,i,j,l} \max_{s=1,\dots,n} \left(\frac{\partial f_{\mathbf{w}_1}}{\partial w_{k,i,j}^{(l)}}(X_s) - \frac{\partial f_{\mathbf{w}_2}}{\partial w_{k,i,j}^{(l)}}(X_s) \right)^2 \cdot 1_{[-\alpha_n,\alpha_n]^d}(X_s)$$

$$+8 \cdot c_2^2 \cdot \|\mathbf{w}_1 - \mathbf{w}_2\|_\infty^2.$$

From the proof of Lemma 6 we can conclude

$$\sum_{k,i,j,l} \max_{s=1,\dots,n} \left(\frac{\partial f_{\mathbf{w}_1}}{\partial w_{k,i,j}^{(l)}}(X_s) \right)^2 \cdot 1_{[-\alpha_n,\alpha_n]^d}(X_s)$$
$$\leq c_{47} \cdot K_n \cdot L \cdot r^2 \cdot d \cdot r^{2L} \cdot \max\{\|\sigma'\|_\infty^{2L}, 1\} \cdot B_n^{2L} \cdot (\gamma_n^*)^2 \cdot \alpha_n^2,$$

$$\frac{1}{n} \sum_{s=1}^n (f_{\mathbf{w}_2}(X_s) - f_{\mathbf{w}_1}(X_s))^2 \cdot 1_{[-\alpha_n,\alpha_n]^d}(X_s)$$
$$\leq 4 \cdot \max\{\|\sigma'\|_\infty^{2L}, 1\} \cdot K_n^2 \cdot (2r+1)^{2L} \cdot (\gamma_n^*)^2 \cdot B_n^{2L} \cdot \alpha_n^2 \cdot \max\{\|\sigma\|_\infty, 1\}^2 \cdot \|\mathbf{w}_1 - \mathbf{w}_2\|^2$$

and

$$\frac{1}{n} \sum_{s=1}^n (Y_s - f_{\mathbf{w}_2}(X_s))^2 \cdot 1_{[-\alpha_n,\alpha_n]^d}(X_s)$$
$$\leq 2 \cdot F_n(\mathbf{v}) + 8 \cdot \max\{\|\sigma'\|_\infty^{2L}, 1\} \cdot K_n^2 \cdot (2r+1)^{2L} \cdot (\gamma_n^*)^2 \cdot B_n^{2L} \cdot \alpha_n^2 \cdot \max\{\|\sigma\|_\infty, 1\}^2$$
$$\cdot \frac{8 t_n}{L_n} \cdot \max\{F_n(v), 1\}.$$

So it remains to bound

$$\sum_{k,i,j,l} \max_{s=1,\dots,n} \left(\frac{\partial f_{\mathbf{w}_1}}{\partial w_{k,i,j}^{(l)}}(X_s) - \frac{\partial f_{\mathbf{w}_2}}{\partial w_{k,i,j}^{(l)}}(X_s) \right)^2 \cdot 1_{[-\alpha_n,\alpha_n]^d}(X_s).$$

By (67) we know that

$$\frac{\partial f_{\mathbf{w}}}{\partial w_{k,i,j}^{(l)}}(x)$$

is for fixed $x \in [-\alpha_n, \alpha_n]^d$ a sum of products of Lipschitz continuous functions (considered as functions of $\mathbf{w}$). Arguing as in the proof of Lemma 6 in Kohler and Krzyżak (2021) we can show that we have for any $x \in [-\alpha_n, \alpha_n]^d$

$$\left| \frac{\partial f_{\mathbf{w}_1}}{\partial w_{k,i,j}^{(l)}}(x) - \frac{\partial f_{\mathbf{w}_2}}{\partial w_{k,i,j}^{(l)}}(x) \right| \leq c_{48} \cdot B_n^{2L} \cdot \gamma_n^* \cdot \alpha_n \cdot \|\mathbf{w}_1 - \mathbf{w}_2\|,$$

which implies

$$\sum_{k,i,j,l} \max_{s=1,\dots,n} \left(\frac{\partial f_{\mathbf{w}_1}}{\partial w_{k,i,j}^{(l)}}(X_s) - \frac{\partial f_{\mathbf{w}_2}}{\partial w_{k,i,j}^{(l)}}(X_s) \right)^2 \cdot 1_{[-\alpha_n,\alpha_n]^d}(X_s)$$

$$\leq c_{49} \cdot K_n \cdot L \cdot r^2 \cdot d \cdot B_n^{4L} \cdot (\gamma_n^*)^2 \cdot \alpha_n^4 \cdot \|\mathbf{w}_1 - \mathbf{w}_2\|^2.$$

Summarizing the above results we get the assertion. □

In order to be able to formulate our next auxiliary result we need the following notation: Let $(x_1, y_1), \dots, (x_n, y_n) \in \mathbb{R}^d \times \mathbb{R}$, let $K \in \mathbb{N}$, let $B_1, \dots, B_K : \mathbb{R}^d \to \mathbb{R}$ and let $c_2 > 0$. In the next lemma we consider the problem to minimize

$$F(\mathbf{a}) = \frac{1}{n} \sum_{i=1}^{n} |\sum_{k=1}^{K} a_k \cdot B_k(x_i) - y_i|^2 + c_2 \cdot \sum_{k=1}^{K_n} a_k^2, \tag{71}$$

where $\mathbf{a} = (a_1, \dots, a_K)^T$, by gradient descent. To do this, we choose $\mathbf{a}^{(0)} \in \mathbb{R}^K$ and set

$$\mathbf{a}^{(t+1)} = \mathbf{a}^{(t)} - \lambda_n \cdot (\nabla_{\mathbf{a}} F)(\mathbf{a}^{(t)}) \tag{72}$$

for some properly chosen $\lambda_n > 0$.

Lemma 8 *Let F be defined by (71) and choose $\mathbf{a}_{opt}$ such that*

$$F(\mathbf{a}_{opt}) = \min_{\mathbf{a} \in \mathbb{R}^K} F(\mathbf{a}).$$

Then for any $\mathbf{a} \in \mathbb{R}^K$ we have

$$\|(\nabla_{\mathbf{a}} F)(\mathbf{a})\|^2 \geq 4 \cdot c_2 \cdot (F(\mathbf{a}) - F(\mathbf{a}_{opt})).$$

Proof. The proof is a modification of the proof of Lemma 3 in Braun, Kohler and Walk (2019).

Set

$$\mathbf{E} = c_2 \cdot \begin{pmatrix} 1 & 0 & 0 & \dots & 0 \\ 0 & 1 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & 1 \end{pmatrix},$$

$$\mathbf{B} = (B_j(x_i))_{1 \leq i \leq n, 1 \leq j \leq K} \quad \text{and} \quad \mathbf{A} = \frac{1}{n} \cdot \mathbf{B}^T \cdot \mathbf{B} + c_2 \cdot \mathbf{E}.$$

Then $\mathbf{A}$ is positive definite and hence regular, from which we can conclude

$$\begin{aligned} F(\mathbf{a}) &= \frac{1}{n} \cdot (\mathbf{B} \cdot \mathbf{a} - \mathbf{y})^T \cdot (\mathbf{B} \cdot \mathbf{a} - \mathbf{y}) + c_2 \cdot \mathbf{a}^T \cdot \mathbf{E} \cdot \mathbf{a} \\ &= \mathbf{a}^T \mathbf{A} \mathbf{a} - 2\mathbf{y}^T \frac{1}{n} \mathbf{B} \mathbf{a} + \frac{1}{n} \mathbf{y}^T \mathbf{y} \\ &= (\mathbf{a} - \mathbf{A}^{-1} \frac{1}{n} \mathbf{B}^T \mathbf{y})^T \mathbf{A} (\mathbf{a} - \mathbf{A}^{-1} \frac{1}{n} \mathbf{B}^T \mathbf{y}) + F(\mathbf{a}_{opt}), \end{aligned}$$

where

$$F(\mathbf{a}_{opt}) = \frac{1}{n} \mathbf{y}^T \mathbf{y} - \mathbf{y}^T \cdot \frac{1}{n} \cdot \mathbf{B} \cdot \mathbf{A}^{-1} \cdot \frac{1}{n} \cdot \mathbf{B}^T \mathbf{y}.$$

Using

$$\mathbf{b}^T \mathbf{A} \mathbf{b} \geq c_2 \cdot \mathbf{b}^T \mathbf{E} \mathbf{b} = c_2 \cdot \mathbf{b}^T \mathbf{b}$$

and $\mathbf{A}^T = \mathbf{A}$ we conclude

$$\begin{aligned}
&F(\mathbf{a}) - F(\mathbf{a}_{opt}) \\
&= ((\mathbf{A}^{1/2})^T(\mathbf{a} - \mathbf{A}^{-1}\frac{1}{n}\mathbf{B}^T\mathbf{y}))^T \mathbf{A}^{1/2}(\mathbf{a} - \mathbf{A}^{-1}\frac{1}{n}\mathbf{B}^T\mathbf{y}) \\
&\leq \frac{1}{c_2} \cdot ((\mathbf{A}^{1/2})^T(\mathbf{a} - \mathbf{A}^{-1}\frac{1}{n}\mathbf{B}^T\mathbf{y}))^T \mathbf{A}\mathbf{A}^{1/2}(\mathbf{a} - \mathbf{A}^{-1}\frac{1}{n}\mathbf{B}^T\mathbf{y}) \\
&= \frac{1}{c_2} \cdot ((\mathbf{A})^T(\mathbf{a} - \mathbf{A}^{-1}\frac{1}{n}\mathbf{B}^T\mathbf{y}))^T \mathbf{A}(\mathbf{a} - \mathbf{A}^{-1}\frac{1}{n}\mathbf{B}^T\mathbf{y}) \\
&= \frac{1}{c_2} \cdot (\mathbf{A}\mathbf{a} - \frac{1}{n}\mathbf{B}^T\mathbf{y})^T(\mathbf{A}\mathbf{a} - \frac{1}{n}\mathbf{B}^T\mathbf{y}) \\
&= \frac{1}{4 \cdot c_2} \cdot (2\mathbf{A}\mathbf{a} - \frac{2}{n}\mathbf{B}^T\mathbf{y})^T(2\mathbf{A}\mathbf{a} - \frac{2}{n}\mathbf{B}^T\mathbf{y}) \\
&= \frac{1}{4 \cdot c_2} \cdot \|(\nabla_{\mathbf{a}} F)(\mathbf{a})\|^2,
\end{aligned}$$

where the last equality follows from

$$(\nabla_{\mathbf{a}} F)(\mathbf{a}) = \nabla_{\mathbf{a}} \left(\mathbf{a}^T \mathbf{A} \mathbf{a} - 2\mathbf{y}^T \frac{1}{n}\mathbf{B}\mathbf{a} + \frac{1}{n}\mathbf{y}^T\mathbf{y} \right) = 2\mathbf{A}\mathbf{a} - \frac{2}{n}\mathbf{B}^T\mathbf{y}.$$

□

References

Kohler, M., and Krzyżak, A.(2021). Over-parametrized deep neural networks minimizing the empirical risk do not generalize well. *Bernoulli*, 27, 2564–2597

Braun, A., Kohler, M., and Walk, H. (2019). On the rate of convergence of a neural network regression estimate learned by gradient descent. *arXiv*: 1912.03921.