*Submitted to the Bernoulli*

# Over-Parametrized Deep Neural Networks Minimizing The Empirical Risk Do Not Generalize Well

MICHAEL KOHLER and ADAM KRZYŻAK

*Michael Kohler*
*Fachbereich Mathematik*
*TU Darmstadt*
*Schlossgartenstr. 7*
*64289 Darmstadt*
*E-mail:* kohler@mathematik.tu-darmstadt.de

*Adam Krzyżak*
*Department of Computer Science*
*Concordia University*
*1455 De Maisonneuve Blvd. West*
*Montreal, Quebec, Canada H3G 1M8*
*E-mail:* krzyzak@cs.concordia.ca

Recently it was shown in several papers that backpropagation is able to find the global minimum of the empirical risk on the training data using over-parametrized deep neural networks. In this paper a similar result is shown for deep neural networks with the sigmoidal squasher activation function in a regression setting, and a lower bound is presented which proves that these networks do not generalize well on a new data in the sense that networks which minimize the empirical risk do not achieve the optimal minimax rate of convergence in estimation of smooth regression functions.

Primary 62G05; secondary 62G20.
*Keywords:* neural networks, nonparametric regression, over-parametrization, rate of convergence.

## 1. Introduction

Deep neural networks are among the most successful approaches in multivariate statistical estimation applications, see, e.g., Schmidhuber (2015) and the literature cited therein. Motivated by the practical success of these networks there has been in recent years an increasing interest in studying the corresponding estimators both practically and theoretically. This is often done in the context of nonparametric regression with random design. Here, $(X, Y)$ is an $\mathbb{R}^d \times \mathbb{R}$–valued random vector satisfying $\mathbf{E}\{Y^2\} < \infty$, and given a sample of $(X, Y)$ of size $n$, i.e., given a data set

$$\mathcal{D}_n = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}, \tag{1}$$

where $(X,Y)$, $(X_1,Y_1)$, …, $(X_n,Y_n)$ are i.i.d. random variables, the aim is to construct an estimate

$$m_n(\cdot) = m_n(\cdot, \mathcal{D}_n) : \mathbb{R}^d \to \mathbb{R}$$

of the regression function $m : \mathbb{R}^d \to \mathbb{R}$, $m(x) = \mathbf{E}\{Y|X=x\}$ such that the $L_2$ error

$$\int |m_n(x) - m(x)|^2 \mathbf{P}_X(dx)$$

is "small" (see, e.g., Györfi et al. (2002) for a systematic introduction to nonparametric regression and a motivation for the $L_2$ error).

It is well–known that one needs smoothness assumptions on the regression function in order to derive non–trivial rate of convergence results for nonparametric regression estimates (cf., e.g., Theorem 7.2 and Problem 7.2 in Devroye, Györfi and Lugosi (1996) and Section 3 in Devroye and Wagner (1980)). To do this we will use the following definition.

**Definition 1** *Let $p = q + s$ for some $q \in \mathbb{N}_0$ and $0 < s \leq 1$, where $\mathbb{N}_0$ is the set of nonnegative integers. A function $f : \mathbb{R}^d \to \mathbb{R}$ is called $(p,C)$-**smooth**, if for every $\alpha = (\alpha_1, \ldots, \alpha_d) \in \mathbb{N}_0^d$ with $\sum_{j=1}^d \alpha_j = q$ the partial derivative $\frac{\partial^q f}{\partial x_1^{\alpha_1} \ldots \partial x_d^{\alpha_d}}$ exists and satisfies*

$$\left| \frac{\partial^q f}{\partial x_1^{\alpha_1} \ldots \partial x_d^{\alpha_d}}(x) - \frac{\partial^q f}{\partial x_1^{\alpha_1} \ldots \partial x_d^{\alpha_d}}(z) \right| \leq C \cdot \|x - z\|^s$$

*for all $x, z \in \mathbb{R}^d$, where $\|\cdot\|$ denotes the Euclidean norm.*

Stone (1982) showed that the optimal minimax rate of convergence in nonparametric regression for $(p,C)$-smooth functions is $n^{-2p/(2p+d)}$. In order to describe this result formally we need to introduce the following class of distributions.

**Definition 2** *Let $p, C, c_1 > 0$. We define $\mathcal{D}^{(p,C)}$ as the class of all distributions of $(X,Y)$ which satisfy*

1. $X \in [0,1]^d$ *a.s.*
2. $\sup_{x \in [0,1]^d} \mathbf{E}\{Y^2|X=x\} \leq c_1$
3. $m(\cdot) = \mathbf{E}\{Y|X=\cdot\}$ *is $(p,C)$–smooth.*

With this notation we can formulate the classical result of Stone (1982) as follows: Let $p, C, c_1 > 0$ be arbitrary and let $\mathcal{D}^{(p,C)}$ be the corresponding class of distributions. Then there exist estimates $m_n$ which satisfy

$$\limsup_{n\to\infty} \sup_{(X,Y)\in\mathcal{D}^{(p,C)}} \frac{\mathbf{E} \int |m_n(x) - m(x)|^2 \mathbf{P}_X(dx)}{n^{-2p/(2p+d)}} < \infty, \tag{2}$$

and no estimate can achieve for this class of distributions a better rate of convergence in the sense that it holds

$$\liminf_{n\to\infty} \inf_{\tilde{m}_n} \sup_{(X,Y)\in\mathcal{D}^{(p,C)}} \frac{\mathbf{E} \int |\tilde{m}_n(x) - m(x)|^2 \mathbf{P}_X(dx)}{n^{-2p/(2p+d)}} > 0, \tag{3}$$

where the above infimum is computed with respect to all possible estimates $\tilde{m}_n$ (cf., Stone (1982) or Sections 3.2, 5.3 and 19.4 in Györfi et al. (2002)).

The above result implies that $n^{-2p/(2p+d)}$ is the optimal rate of convergence for estimation of $(p, C)$–smooth regression functions. In case that $d$ is large compared to $p$ this rate of convergence is rather slow (so called curse of dimensionality). It is well-known that it is possible to circumvent this curse of dimensionality by imposing the additional constraints on the regression function like additivity (cf., Stone (1985, 1994)). Recently it was shown that under rather general compository assumptions on the regression function the curse of dimensionality can be avoided by the suitably defined least squares neural network regression estimates, which we want to explore next.

The starting point in defining a neural network is the choice of an activation function $\sigma : \mathbb{R} \to \mathbb{R}$. Here, we use in the sequel so–called squashing functions, which are nondecreasing and satisfy $\lim_{x\to-\infty} \sigma(x) = 0$ and $\lim_{x\to\infty} \sigma(x) = 1$. An example of a squashing function is the so-called sigmoidal or logistic squasher

$$\sigma(x) = \frac{1}{1 + \exp(-x)} \quad (x \in \mathbb{R}). \tag{4}$$

In applications, also unbounded activation functions are often used, e.g., the famous ReLU activation function

$$\sigma(x) = \max\{x, 0\}.$$

The network architecture $(L, \mathbf{k})$ depends on a positive integer $L$ called the *number of hidden layers* and a *width vector* $\mathbf{k} = (k_1, \ldots, k_L) \in \mathbb{N}^L$ that describes the number of neurons in the first, second, $\ldots$, $L$-th hidden layer. A multilayer feedforward neural network with architecture $(L, \mathbf{k})$ and activation function $\sigma$ is a real-valued function $f : \mathbb{R}^d \to \mathbb{R}$ defined by

$$f_{\mathbf{w}}(x) = \sum_{i=1}^{k_L} w_{1,i}^{(L)} \cdot f_i^{(L)}(x) + w_{1,0}^{(L)} \tag{5}$$

for some $w_{1,0}^{(L)}, \ldots, w_{1,k_L}^{(L)} \in \mathbb{R}$ and for $f_i^{(L)}$'s recursively defined by

$$f_i^{(l)}(x) = \sigma\left(\sum_{j=1}^{k_{l-1}} w_{i,j}^{(l-1)} \cdot f_j^{(l-1)}(x) + w_{i,0}^{(l-1)}\right) \tag{6}$$

for some $w_{i,0}^{(l-1)}, \ldots, w_{i,k_{l-1}}^{(l-1)} \in \mathbb{R}$ $(l = 2, \ldots, L)$ and

$$f_i^{(1)}(x) = \sigma\left(\sum_{j=1}^{d} w_{i,j}^{(0)} \cdot x^{(j)} + w_{i,0}^{(0)}\right) \tag{7}$$

for some $w_{i,0}^{(0)}, \ldots, w_{i,d}^{(0)} \in \mathbb{R}$.

In the sequel we want to use the data (1) in order to choose the weights $\mathbf{w} = (w_{i,j}^{(s)})_{i,j,s}$ of the neural network such that the resulting function $f_{\mathbf{w}}$ defined by (5)–(7) is a good estimate of the regression function. This can be done for instance by applying the principle of the least squares. Here one defines a suitable class $\mathcal{F}_n$ of neural networks and chooses that function from this class which minimizes the error on the training data, i.e., one defines the so–called least squares neural network estimate by

$$m_n(\cdot) = \arg \min_{f \in \mathcal{F}_n} \frac{1}{n} \sum_{i=1}^n |f(X_i) - Y_i|^2.$$

Recently it was shown in several articles, that such least squares estimates based on deep neural networks achieve rates of convergence independent on the dimension $d$ of $X$ if suitable compository constraints on the regression function are imposed, cf., e.g., Kohler and Krzyżak (2017), Bauer and Kohler (2019), Kohler and Langer (2019) and Schmidt-Hieber (2020a). Hence neural networks can circumvent the curse of dimensionality in case that rather general compository assumptions on the regression function hold. Eckle and Schmidt-Hieber (2019) and Kohler, Krzyżak and Langer (2019) showed that the least squares neural network regression estimates based on deep neural networks can achieve the rate of convergence results similar to piecewise polynomial partition estimates where the partition is chosen in an optimal way. Results concerning estimation by neural networks of piecewise polynomial regression functions with partitions having rather general smooth boundaries have been obtained by Imaizumi and Fukamizu (2019).

Unfortunately it is not possible to compute the least squares neural networks regression estimate exactly, because such computation requires minimization of the non-convex and nonlinear function

$$F_n(\mathbf{w}) = \frac{1}{n} \sum_{i=1}^n |f_{\mathbf{w}}(X_i) - Y_i|^2$$

with respect to the weight vector $\mathbf{w}$. In practice, one uses gradient descent in order to compute the minimum of the above function approximately. Here one chooses a random starting value $\mathbf{w}^{(0)}$ for the weight vector, and then defines

$$\mathbf{w}^{(t+1)} = \mathbf{w}^{(t)} - \lambda_n \cdot (\nabla_{\mathbf{w}} F_n)(\mathbf{w}^{(t)}) \quad (t = 0, \dots, t_n - 1)$$

for some suitably chosen stepsize $\lambda_n > 0$ and the number of gradient descent steps $t_n \in \mathbb{N}$. Then the regression estimate is defined by $m_n(x) = f_{\mathbf{w}^{(t_n)}}(x)$.

There are quite a few papers which try to prove that backpropagation works theoretically for deep neural networks. The most popular approach in this context is the so–called landscape approach. Choromanska et al. (2015) used random matrix theory to derive a heuristic argument showing that the risk of most of the local minima of the empirical $L_2$ risk $F_n(\mathbf{w})$ is not much larger than the risk of the global minimum. For neural networks with special activation function it was possible to validate this claim, see, e.g., Arora et al. (2018), Kawaguchi (2016), and Du and Lee (2018), which have analyzed gradient descent for neural networks with linear or quadratic activation function. But for such neural networks there do not exist good approximation results, consequently, one cannot derive from these results good rates of convergence for neural network regression estimates. Du et al. (2018) analyzed gradient descent applied to neural networks with one hidden layer in case of an input with a Gaussian distribution. They used the expected gradient instead of the gradient in their gradient descent routine, and therefore, their result cannot be used to derive the rate of convergence results for a neural network regression estimate learned by the gradient descent. Liang et al. (2018) applied gradient descent to a modified loss function in classification, where it is assumed that the data can be interpolated by a neural network. Here, as we will show in this paper (cf., Theorem 2 below), the last assumption does not lead to good rates of convergence in nonparametric regression, and it is unclear whether the main idea (of simplifying the estimation by a modification of the loss function) can also be used in a regression setting. Neural tangent kernel networks (NTK) were introduced by Jackot, Gabriel and Honger (2018). They showed that in the infinite-width limit case NTK converges to a deterministic limit kernel which stays constant during Gaussian descent training of the random weights initialized with the Gaussian distributions. These results were extended by Huang, Du and Xu (2020) to orthogonal initialization which was shown to speed up training of fully connected deep networks. Nitanda and Suzuki (2020) obtained global convergence rate for the averaged stochastic gradient descent for overparametrized two-layer neural networks.

Recently it was shown in several papers, see, e.g., Allen-Zhu, Li and Song (2019), Kawaguchi and Huang (2019) and the literature cited therein, that the gradient descent leads to a small empirical $L_2$ risk in over-parametrized neural networks. Here the results in Allen-Zhu, Li and Song (2019) are proven for the ReLU activation function and neural networks with a polynomial size in the sample size. The neural networks in Kawaguchi and Huang (2019) use squashing activation functions and are much smaller (in fact, they require only a linear size in the sample size). In contrast to Allen-Zhu, Li and Song (2019) there the learning rate is set to zero for all neurons except for neurons in the output layer and consequently in different layers of the network different learning rates are used. Actually, they compute a linear least squares estimate with the gradient descent, which is not used in practice. Related to these results are various recent studies which try to understand the capacity of neural networks. These works examine the ability of neural networks to fit the training data, either on a finite data set as in Bubeck et al. (2020), or with respect to neural networks trained by gradient methods as in Daniely (2019, 2020), or in the so–called neural tangent training as in Montanari and Zhong (2020).

The main results in this paper are twofold: Firstly, we show, that it is rather easy to show a corresponding result for a deep neural network regression estimate with the logistic squasher activation function, where the learning rate is the same for all neurons of the network. The main trick here is that we use a special topology of the neural network, where we compute a huge number of deep fully connected neural networks in parallel and use the final layer to compute a linear combination of all the outputs of these networks. Here we use a simple initialization of the inner weights from the uniform distribution and show that this results (with high probability) in an initial network, where for each $x$-value of the data points we have one neuron in the output layer which has output approximately one at this $x$-value and approximately zero at all other $x$-values of the sample. Since we use the logistic squasher as an activation function we are able to show that this implies that all the inner weights in the fully connected neural networks corresponding to these neurons do not change much during training and consequently the gradient descent performs for these parts of the neural network in a similar way as the gradient descent applied to a linear function space which we can easily analyze. Of course, in this result neither the topology of the network nor the random initialization of the weights is typical for deep learning, which is the place where our second result is important. This result shows that any estimate (including the above cited overparameterized neural networks estimates), which achieves more or less the minimal empirical $L_2$ risk on the training data, does not generalize well to new (independent) data in a sense that it does not achieve the optimal minimax rate of convergence $n^{-2p/(2p+d)}$ in case of the class $\mathcal{D}^{(p,C)}$ of distributions introduced above. Here the main trick is that we also allow discrete design distributions and prove a general result which shows that any estimate which achieves with high probability a very small error on the training data in case of such distributions does not achieve the optimal minimax error.

A result related to our second result was recently presented in Schmidt-Hieber (2020b). Here it was shown that gradient descent in a simplified over-parametrized regime converges to a spline interpolant and hence it is not consistent even if the distribution of $X$ is uniform on some compact cube (a case which is not covered by our lower bound in Theorem 2). However, it is unclear whether this result derived in a simplified setting also holds in general (as our result does).

Our second result contrasts the recent trend in machine learning, where one tries to argue that such estimates can achieve good rates of convergence (see, e.g., Bartlett et al. (2019), Belkin et al. (2019), Hastie et al. (2019) and the literature cited therein). We would like to emphasize that our result above does not contradict Belkin, Rakhlin and Tsybakov (2018), who show that learning method which interpolates the training data can achieve the optimal rates for nonparametric regression problems, because it is assumed there that the design variable has a density with respect to the Lebesgue-Borel measure, which is bounded away from zero and infinity. It also does not contradict Cao and Gu (2019), Neyshabur et al. (2019) or Allen-Zhu, Li and Liang (2019), who derive generalization bounds for over-parametrized neural networks. We can justify our claim as follows. Firstly, Cao and Gu (2019) consider a noiseless classification problem where the classes are separated according to marging condition, which is not a regression problem with noise as considered in our paper here and where interpolation of the training data is clearly a good idea. Secondly, Neyshabur et al. (2019) also consider a classification problem and show that there over-parametrized deep neural networks generalize well in case that they are contained in a special subclass of these networks (which has a small Rademacher complexity). It is argued by means of simulations that the over-parametrized neural network learned by the gradient descent is indeed contained in this subclass, but no formal proof of this claim is provided. And thirdly, the generalization bounds in Allen-Zhu, Li and Liang (2019) are derived in a PAC setting which is different from the classical regression setting considered in our paper and where in each gradient descent step a new independent data point is used which clearly helps to avoid the overfitting observed in the proof of our second result.

We would also like to point out that it has been observed in the experiments that the choice of the learning rate and of the initial initialization of the weights can bias the optimization trajectory to minima with poor generalization properties (cf., Chizat, Oyallon and Bach (2020) and Woodworth et al. (2020)). However, our second result is independent of this. It does not matter how the over-parametrized neural network is learned at all, as soon as it achieves a very small empirical risk, it cannot achieve the optimal minimax rate of convergence for the classes $\mathcal{D}^{(p,C)}$ introduced above (cf., (2) and (3)).

Throughout the paper, the following notation is used: The sets of natural numbers, natural numbers including 0, and real numbers are denoted by $\mathbb{N}$, $\mathbb{N}_0$ and $\mathbb{R}$, respectively. The Euclidean norm of $x \in \mathbb{R}^d$ is denoted by $\|x\|$ and $\|x\|_\infty$ denotes its supremum norm. For $f : \mathbb{R}^d \to \mathbb{R}$

$$\|f\|_\infty = \sup_{x \in \mathbb{R}^d} |f(x)|$$

is its supremum norm. Let $p = q + s$ for some $q \in \mathbb{N}_0$ and $0 < s \leq 1$. A function $f : \mathbb{R}^d \to \mathbb{R}$ is called $(p, C)$-smooth, if for every $\alpha = (\alpha_1, \dots, \alpha_d) \in \mathbb{N}_0^d$ with $\sum_{j=1}^d \alpha_j = q$ the partial derivative $\frac{\partial^q f}{\partial x_1^{\alpha_1} \dots \partial x_d^{\alpha_d}}$ exists and satisfies

$$\left| \frac{\partial^q f}{\partial x_1^{\alpha_1} \dots \partial x_d^{\alpha_d}}(x) - \frac{\partial^q f}{\partial x_1^{\alpha_1} \dots \partial x_d^{\alpha_d}}(z) \right| \leq C \cdot \|x - z\|^s$$

for all $x, z \in \mathbb{R}^d$.

The outline of this paper is as follows: In Section 2 the over-parametrized neural network regression estimate is defined. The main results are presented in Section 3 and proven in Section 4.

# 2. Over-parametrized neural network regression estimator

In the sequel we use the logistic squasher $\sigma(x) = 1/(1 + e^{-x})$ as the activation function, and we use a network topology where we compute the linear combination of $k_n$ fully connected neural networks with $L$ layers and $r_0$ neurons per layer. Thus we define our neural networks by

$$f_{\mathbf{w}}(x) = \sum_{j=1}^{k_n} w_{1,1,j}^{(L)} \cdot f_{j,1}^{(L)}(x) + w_{1,1,0}^{(L)} \tag{8}$$

for some $w_{1,1,0}^{(L)}, \dots, w_{1,1,k_n}^{(L)} \in \mathbb{R}$, where $f_{j,1}^{(L)}$ are recursively defined by

$$f_{k,i}^{(l)}(x) = \sigma\left( \sum_{j=1}^{r_0} w_{k,i,j}^{(l-1)} \cdot f_{k,j}^{(l-1)}(x) + w_{k,i,0}^{(l-1)} \right) \tag{9}$$

for some $w_{k,i,0}^{(l-1)}, \dots, w_{k,i,r_0}^{(l-1)} \in \mathbb{R}$ $(l = 2, \dots, L)$ and

$$f_{k,i}^{(1)}(x) = \sigma\left( \sum_{j=1}^{d} w_{k,i,j}^{(0)} \cdot x^{(j)} + w_{k,i,0}^{(0)} \right) \tag{10}$$

for some $w_{k,i,0}^{(0)}, \dots, w_{k,i,d}^{(0)} \in \mathbb{R}$.

The above neural network consists of $k_n$ fully connected neural networks with depth $L$, which are computed in parallel. These networks have $r_0$ neurons in all layers except for the last layer, where they only have one neuron. In the $k$-th such network we denote the output of neuron $i$ in the $l$-th layer by $f_{k,i}^{(l)}$, and the weight between neuron $j$-th in the $(l-1)$-th layer and neuron $i$ in the $l$-th layer is denoted by $w_{k,i,j}^{(l-1)}$.

We learn the weight vector $\mathbf{w} = (w_{k,i,j}^{(s)})_{k,i,j,s}$ of our neural nework by the gradient descent. We initialize $\mathbf{w}^{(0)}$ by setting

$$w_{1,1,j}^{(L)} = 0 \quad \text{for } j = 0, \dots, k_n, \tag{11}$$

and by choosing all others weights randomly such that all weights $w_{k,i,j}^{(s)}$ with $s < L$ are independent uniformly distributed on $[-n^4, n^4]$, and we set

$$\mathbf{w}^{(t+1)} = \mathbf{w}^{(t)} - \lambda_n \cdot (\nabla_{\mathbf{w}} F_n)(\mathbf{w}^{(t)}) \quad (t = 0, \dots, t_n - 1)$$

where

$$F_n(\mathbf{w}) = \frac{1}{n} \sum_{i=1}^n |f_{\mathbf{w}}(X_i) - Y_i|^2$$

is the empirical $L_2$ risk of the network $f_{\mathbf{w}}$ on the training data. The step size $\lambda_n > 0$ and the number $t_n$ of gradient descent steps will be chosen below.

Because of (11) we have

$$F_n(\mathbf{w}^{(0)}) = \frac{1}{n} \sum_{i=1}^n |Y_i|^2.$$

# 3. Main results

Our first result shows that our estimate is able to achieve with high probability a very small error on the training data in case that $k_n$, $\lambda_n$ and $t_n$ are suitably chosen.

**Theorem 1** *Let $r_0 \in \mathbb{N}$ with $r_0 \geq 2 \cdot d$, let $L \in \mathbb{N}$ with $L \geq 2$, set*

$$k_n = n^{5 \cdot (L-2) \cdot (r_0^2 + r_0) + 5 \cdot r_0 \cdot (d+2) + 7},$$

$$\lambda_n = \frac{1}{n^{8(L-2) \cdot (r_0^2 + r_0) + 8 \cdot r_0 \cdot (d+2) + 16L + 15}}$$

*and*

$$t_n = 2 \cdot n^{8 \cdot (L-2) \cdot (r_0^2 + r_0) + 8 \cdot r_0 \cdot (d+2) + 16L + 17},$$

*and define the estimate as in Section 2. Then for sufficiently large $n$ we have on the event*

$$\inf\{\|X_i - X_j\|_\infty : 1 \leq i, j \leq n, X_i \neq X_j\} \geq \frac{1}{(n+1)^3},$$

$$\max\{\|X_i\|_\infty \,:\, 1 \le i \le n\} \le 1 \quad \textit{and} \quad \max\{|Y_i| \,:\, 1 \le i \le n\} \le n^2$$

*that with probability at least $1 - 1/n$ the random choice of $\mathbf{w}^{(0)}$ leads to*

$$\frac{1}{n}\sum_{i=1}^{n} |f_{\mathbf{w}^{(t_n)}}(X_i) - Y_i|^2 \le \min_{g:\mathbb{R}^d \to \mathbb{R}} \frac{1}{n}\sum_{i=1}^{n} |g(X_i) - Y_i|^2 + \frac{1}{n \cdot \log n}. \tag{12}$$

**Remark 1. a)** A corresponding result was shown in Kawaguchi and Huang (2019) for a fully connected network of much smaller size (linear instead of polynomial in the sample size as in Theorem 1 above), however there the learning rate of the gradient descent was set to zero for all weights $w_{k,i,j}^{(r)}$ with $r < L$. In contrast, our learning rate is positive for all weights and the same learning rate is used for all the weights in the network.

**b)** If the design points in Theorem 1 are different, they are assumed to be $(n+1)^{-3}$-separated from each other with respect to the maximum norm. The possibility of observing the same design point twice (which does not happen for continuous design a.s.) also then explains why the first term on the right hand side of (12) is not zero.

As our next result shows that any estimate which (as our estimate from Theorem 1) achieves with high probability a very small error on the training data does not, in general generalize well on a new independent data (provided we allow the distributions of $X$ which are concentrated on finite sets).

**Theorem 2** *Let $(X, Y)$, $(X_1, Y_1)$, … be independent and identically distributed $\mathbb{R}^d \times \mathbb{R}$-valued random variables with $\mathbf{E}Y^2 < \infty$, and let $U$ be an $\mathbb{R}^K$–valued random variable independent of the random variables above. Let $\mathcal{P}_n$ be a subset of $\mathbb{R}^K$, and let*

$$m_n(\cdot) = m_n(\cdot, (X_1, Y_1), \dots, (X_n, Y_n), U) : \mathbb{R}^d \to \mathbb{R}$$

*be an estimate of $m$. Let $\kappa_n > 0$ and let $\delta_n \le 1/(n+1)^3$ and assume that $m_n$ satisfies for any data set $(X_1, Y_1)$, …, $(X_n, Y_n)$*

$$\frac{1}{n}\sum_{i=1}^{n} |m_n(X_i) - Y_i|^2 \le \min_{g:\mathbb{R}^d \to \mathbb{R}} \frac{1}{n}\sum_{i=1}^{n} |g(X_i) - Y_i|^2 + \kappa_n$$

*whenever*

$$\inf\{\|X_i - X_j\|_\infty \,:\, 1 \le i, j \le n,\, X_i \ne X_j\} \ge \delta_n \quad \textit{and} \quad U \in \mathcal{P}_n.$$

*Then there exists a distribution of $(X, Y)$ such that $X \in [0,1]^d$ a.s., $Y \in \{-1, 1\}$ a.s., $m(x) = 0$ for all $x \in [0,1]^d$ and such that we have for $n \ge 10$*

$$\mathbf{E}\int |m_n(x) - m(x)|^2 \mathbf{P}_X(dx) \ge \frac{1}{10} - n \cdot \kappa_n - \frac{1}{2} \cdot \mathbf{P}_U(\mathcal{P}_n^c).$$

In Theorem 2 above $U$ is the randomness in the procedure (e.g. random initialization) and $\mathcal{P}_n$ is the subset of "good" initializations.

**Corollary 1** *Let $m_n$ be either the estimate of Theorem 1 or an arbitrary estimates which satisfies the assumptions of Theorem 2 for $\kappa_n = 1/(n \cdot \log n)$ and some set $\mathcal{P}_n$ with $\mathbf{P}_U(\mathcal{P}_n^c) \leq 1/n$. Let $p, C, c_1 > 0$ and let $\mathcal{D}^{(p,C)}$ be the class of all distributions of $(X, Y)$ introduced in Definition 2. Then we have for $n$ sufficiently large*

$$\sup_{(X,Y) \in \mathcal{D}^{(p,C)}} \mathbf{E} \int |m_n(x) - m(x)|^2 \mathbf{P}_X(dx) \geq \frac{1}{11}.$$

**Proof.** Let $U$ be the values for the random initialization of the weights of the estimate in Theorem 1. By Theorem 1 we know that there exists a set $\mathcal{P}_n$ of weights such that (12) holds for $n$ sufficiently large whenever $U \in \mathcal{P}_n$, where $\mathbf{P}_U(\mathcal{P}_n^c) \leq 1/n$. Hence the assumptions of Theorem 2 are satisfied with $\kappa_n = 1/(n \cdot \log n)$. Let $(X, Y)$ be the distribution from Theorem 2. Then for $n$ sufficiently large

$$\begin{aligned} \sup_{(X,Y) \in \mathcal{D}^{(p,C)}} \mathbf{E} \int |m_n(x) - m(x)|^2 \mathbf{P}_X(dx) &\geq \mathbf{E} \int |m_n(x) - m(x)|^2 \mathbf{P}_X(dx) \\ &\geq \frac{1}{10} - n \cdot \frac{1}{n \cdot \log n} - \frac{1}{2} \cdot \frac{1}{n} \\ &\geq \frac{1}{11}. \end{aligned}$$

□

**Remark 2.** Corollary 1 implies that the estimate of Theorem 1 does not achieve the optimal minimax rate of convergence for the class $\mathcal{D}^{(p,C)}$ (cf., (2) and (3)). In fact, the minimax $L_2$ error for this class does not even converge to zero, let alone to the optimal value.

**Remark 3.** The neural network estimate in Corollary 1 behaves bad for the discrete distributions in $\mathcal{D}^{(p,C)}$. Also it is important in our proof that we allow the bad distribution to change with sample size. However, we think the reason for this bad behaviour of our estimate is the over-parametrization. Because in case that we restrict the number of weights in the neural network accordingly it follows from Theorem 1 in Kohler and Langer (2019) that a least squares regression estimates based on a set of fully connected feedforward neural network achieves the optimal rate of convergence for the class $\mathcal{D}^{(p,C)}$ up to some logarithmic factor. Hence in this case the neural network estimate is also good for discrete estimates.

**Remark 4.** As mentioned in the introduction it is shown in Belkin, Rakhlin and Tsybakov (2018) that estimates which interpolate the data can nevertheless achieve the optimal rates for nonparametric regression problems in case that the design variable has a density with respect to the Lebesgue-Borel measure which is bounded away from zero and infinity. In Mücke and Steinwart (2019) it is shown that there even exists neural network regression estimates which have this property. In the same paper it is also shown that there exists neural network regression estimates which minimize the empirical $L_2$ risk and which do not achieve the optimal rate of convergence (and which are in fact not even consistent). So even if we assume that the distribution of the design variable is nice, a neural network estimate which interpolates the training data might generalize not well.

**Remark 5.** In light of the previous two remarks the takeaway message for practitioners from our results is that it is not clear whether an overparameterized neural network which minimizes the empirical $L_2$ risk generalizes well on new data.

# 4. Proofs

## 4.1. Proof of Theorem 1

**Lemma 1** *Let $F : \mathbb{R}^K \to \mathbb{R}$ be differentiable, let $C_{Lip,n} > 0$, set*

$$\lambda_n = \frac{1}{C_{Lip,n}},$$

*let $\mathbf{a}_1 \in \mathbb{R}^K$ and set*

$$\mathbf{a}_2 = \mathbf{a}_1 - \lambda_n \cdot (\nabla_{\mathbf{a}} F)(\mathbf{a}_1).$$

*Then*

$$\|(\nabla_{\mathbf{a}} F)(\mathbf{a}) - (\nabla_{\mathbf{a}} F)(\mathbf{a}_1)\| \le C_{Lip,n} \cdot \|\mathbf{a} - \mathbf{a}_1\| \tag{13}$$

*for all $\mathbf{a} = \mathbf{a}_1 + s \cdot (\mathbf{a}_2 - \mathbf{a}_1)$, $s \in [0,1]$ implies*

$$F(\mathbf{a}_2) \le F(\mathbf{a}_1) - \frac{1}{2 \cdot C_{Lip,n}} \cdot \|(\nabla_{\mathbf{a}} F)(\mathbf{a}_1)\|^2.$$

**Proof.** See proof of Lemma 1 in Braun, Kohler and Walk (2019). □

Set

$$F_n(\mathbf{w}) = \frac{1}{n} \sum_{i=1}^n |f_{\mathbf{w}}(X_i) - Y_i|^2$$

where $f_{\mathbf{w}}$ is defined by (8)–(10).

**Lemma 2** *Let $f_{\mathbf{w}}$ be defined by (8)–(10) and assume that for any $i \in \{1, \dots, n\}$ there exists $j_i \in \{1, \dots, k_n\}$ such that*

$$f_{j_i,1}^{(L)}(X_i) \ge 1 - \frac{2}{n^2} \quad \text{and} \quad \sup_{t \in \{1,\dots,n\}, X_t \neq X_i} f_{j_i,1}^{(L)}(X_t) \le \frac{2}{n^2} \tag{14}$$

*hold. Then we have for any $n \ge 5$*

$$\|(\nabla_{\mathbf{w}} F_n(\mathbf{w}))\|^2 \quad \ge \quad \frac{1}{n} \cdot \left( \frac{1}{n} \sum_{i=1}^n |f_{\mathbf{w}}(X_i) - Y_i|^2 - \min_{g:\mathbb{R}^d \to \mathbb{R}} \frac{1}{n} \sum_{i=1}^n |g(X_i) - Y_i|^2 \right).$$

**Proof.** Set

$$\bar{m}_n(x) = \frac{\sum_{i=1}^n Y_i \cdot I_{\{X_i = x\}}}{\sum_{i=1}^n I_{\{X_i = x\}}} \quad (x \in \mathbb{R}^d),$$

where we use the convention $0/0=0$. We have

$$\frac{1}{n}\sum_{i=1}^n |f(X_i)-Y_i|^2 = \frac{1}{n}\sum_{i=1}^n |f(X_i)-\bar{m}_n(X_i)|^2 + \frac{1}{n}\sum_{i=1}^n |\bar{m}_n(X_i)-Y_i|^2,$$

since

$$\begin{aligned} &\frac{1}{n}\sum_{i=1}^n (f(X_i)-\bar{m}_n(X_i))\cdot(\bar{m}_n(X_i)-Y_i) \\ &= \frac{1}{n}\sum_{x\in\{X_1,\dots,X_n\}} (f(x)-\bar{m}_n(x))\cdot \sum_{1\le i\le n: X_i=x} (\bar{m}_n(X_i)-Y_i) = 0. \end{aligned}$$

This implies

$$\min_{g:\mathbb{R}^d\to\mathbb{R}} \frac{1}{n}\sum_{i=1}^n |g(X_i)-Y_i|^2 = \frac{1}{n}\sum_{i=1}^n |\bar{m}_n(X_i)-Y_i|^2$$

and

$$\frac{1}{n}\sum_{i=1}^n |f(X_i)-\bar{m}_n(X_i)|^2 = \frac{1}{n}\sum_{i=1}^n |f(X_i)-Y_i|^2 - \min_{g:\mathbb{R}^d\to\mathbb{R}} \frac{1}{n}\sum_{i=1}^n |g(X_i)-Y_i|^2$$

for any $f:\mathbb{R}^d\to\mathbb{R}$.

Next we observe

$$\begin{aligned} \|(\nabla_{\mathbf{w}} F_n(\mathbf{w}))\|^2 &= \sum_{k,i,j,s} \left| \frac{\partial}{\partial w_{k,j,i}^{(s)}} F_n(\mathbf{w}) \right|^2 \\ &\ge \sum_{i\in\{1,\dots,n\}, j_i\neq j_s \text{ for all } s<i} \left| \frac{\partial}{\partial w_{1,1,j_i}^{(L)}} F_n(\mathbf{w}) \right|^2 \\ &= \sum_{i\in\{1,\dots,n\}, j_i\neq j_s \text{ for all } s<i} \left| \frac{2}{n}\cdot\sum_{t=1}^n (f_{\mathbf{w}}(X_t)-Y_t)\cdot \frac{\partial}{\partial w_{1,1,j_i}^{(L)}} f_{\mathbf{w}}(X_t) \right|^2 \\ &= \sum_{i\in\{1,\dots,n\}, j_i\neq j_s \text{ for all } s<i} \left| \frac{2}{n}\cdot\sum_{t=1}^n (f_{\mathbf{w}}(X_t)-Y_t)\cdot f_{j_i,1}^{(L)}(X_t) \right|^2 \\ &\ge \sum_{i\in\{1,\dots,n\}, j_i\neq j_s \text{ for all } s<i} \left( \frac{1}{2}\cdot \left| \frac{2}{n}\cdot \sum_{t\in\{1,\dots,n\}, X_t=X_i} (f_{\mathbf{w}}(X_i)-Y_t)\cdot f_{j_i,1}^{(L)}(X_i) \right|^2 \right. \\ &\qquad \left. - \left| \frac{2}{n}\cdot \sum_{t\in\{1,\dots,n\}, X_t\neq X_i} (f_{\mathbf{w}}(X_t)-Y_t)\cdot f_{j_i,1}^{(L)}(X_t) \right|^2 \right), \end{aligned}$$

where the last inequality followed from $b^2 \leq 2(b-a)^2 + 2a^2$ which implies

$$a^2 \geq \frac{1}{2} b^2 - (a-b)^2 \quad (a, b \in \mathbb{R}).$$

Using

$$\sum_{t \in \{1,\dots,n\}, X_t = X_i} (f_{\mathbf{w}}(X_i) - Y_t) = |\{1 \leq k \leq n \, : \, X_k = X_i\}| \cdot (f_{\mathbf{w}}(X_i) - \bar{m}_n(X_i)),$$

$$\sum_{t \in \{1,\dots,n\}, X_t \neq X_i} (f_{\mathbf{w}}(X_t) - Y_t) \cdot f_{j_i,1}^{(L)}(X_t)$$
$$= \sum_{t \in \{1,\dots,n\}, X_t \notin \{X_i, X_1, \dots, X_{t-1}\}} |\{1 \leq k \leq n \, : \, X_k = X_t\}| \cdot (f_{\mathbf{w}}(X_t) - \bar{m}_n(X_t)) \cdot f_{j_i,1}^{(L)}(X_t),$$

(14) and the inequality of Jensen we conclude

$$\begin{aligned}
&\|(\nabla_{\mathbf{w}} F_n(\mathbf{w}))\|^2 \\
&\geq \frac{2}{n^2} \cdot \sum_{i \in \{1,\dots,n\}, j_i \neq j_s \text{ for all } s<i} |\{1 \leq k \leq n \, : \, X_k = X_i\}|^2 \cdot (f_{\mathbf{w}}(X_i) - \bar{m}_n(X_i))^2 \cdot \left(1 - \frac{2}{n^2}\right)^2 \\
&\qquad -4 \cdot \sum_{i \in \{1,\dots,n\}, j_i \neq j_s \text{ for all } s<i} \; \sum_{t \in \{1,\dots,n\}, X_t \notin \{X_i, X_1, \dots, X_{t-1}\}} \\
&\qquad\qquad \frac{|\{1 \leq k \leq n \, : \, X_k = X_t\}|}{n} \cdot |f_{\mathbf{w}}(X_t) - \bar{m}_n(X_t)|^2 \cdot \frac{4}{n^4} \\
&\geq \frac{2}{n^2} \cdot \sum_{i \in \{1,\dots,n\}, j_i \neq j_s \text{ for all } s<i} |\{1 \leq k \leq n \, : \, X_k = X_i\}| \cdot (f_{\mathbf{w}}(X_i) - \bar{m}_n(X_i))^2 \cdot \left(1 - \frac{2}{n^2}\right)^2 \\
&\qquad -4 \cdot \sum_{i \in \{1,\dots,n\}, j_i \neq j_s \text{ for all } s<i} \; \sum_{t \in \{1,\dots,n\}, X_t \notin \{X_i, X_1, \dots, X_{t-1}\}} \\
&\qquad\qquad \frac{|\{1 \leq k \leq n \, : \, X_k = X_t\}|}{n} \cdot |f_{\mathbf{w}}(X_t) - \bar{m}_n(X_t)|^2 \cdot \frac{4}{n^4} \\
&\geq \frac{2}{n} \cdot \left(1 - \frac{2}{n^2}\right)^2 \cdot \frac{1}{n} \sum_{t=1}^{n} (f_{\mathbf{w}}(X_t) - \bar{m}_n(X_t))^2 \\
&\qquad\qquad -4 \cdot n \cdot \frac{4}{n^4} \cdot \frac{1}{n} \cdot \sum_{t=1}^{n} |f_{\mathbf{w}}(X_t) - \bar{m}_n(X_t)|^2 \\
&= \left( \frac{2}{n} \cdot \left(1 - \frac{2}{n^2}\right)^2 - \frac{16}{n^3} \right) \cdot \frac{1}{n} \cdot \sum_{t=1}^{n} |f_{\mathbf{w}}(X_t) - \bar{m}_n(X_t)|^2 \\
&= \left( \frac{2}{n} - \frac{8}{n^3} + \frac{8}{n^5} - \frac{16}{n^3} \right) \cdot \frac{1}{n} \cdot \sum_{t=1}^{n} |f_{\mathbf{w}}(X_t) - \bar{m}_n(X_t)|^2
\end{aligned}$$

$$
\begin{aligned}
&\geq \frac{1}{n} \cdot \frac{1}{n} \cdot \sum_{t=1}^{n} (f_{\mathbf{w}}(X_t) - \bar{m}_n(X_t))^2 \\
&= \frac{1}{n} \cdot \left( \frac{1}{n} \sum_{i=1}^{n} |f_{\mathbf{w}}(X_i) - Y_i|^2 - \min_{g:\mathbb{R}^d \to \mathbb{R}} \frac{1}{n} \sum_{i=1}^{n} |g(X_i) - Y_i|^2 \right).
\end{aligned}
$$

□

**Lemma 3** *Define* $\mathbf{w}^{(t)}$ *by*

$$
\mathbf{w}^{(t+1)} = \mathbf{w}^{(t)} - \lambda_n \cdot (\nabla_{\mathbf{w}} F_n)(\mathbf{w}^{(t)}) \quad (t = 0, \ldots, t_n - 1)
$$

*for some fixed* $\mathbf{w}^{(0)}$ *and*

$$
\lambda_n = \frac{1}{C_{Lip,n}}.
$$

*Assume that (13) holds for* $F = F_n$ *and all* $\mathbf{a}_1 = \mathbf{w}^{(t)}$ *and* $\mathbf{a}_2 = \mathbf{w}^{(t+1)}$ *and any* $t \in \{0, 1, \ldots, t_n - 1\}$*. Furthermore assume that (14) holds for all* $\mathbf{w} = \mathbf{w}^{(t)}$ $(t \in \{0, 1, \ldots, t_n - 1\})$*. Then we have for any* $n \geq 5$

$$
\begin{aligned}
&F_n(\mathbf{w}^{(t_n)}) - \min_{g:\mathbb{R}^d \to \mathbb{R}} \frac{1}{n} \sum_{i=1}^{n} |g(X_i) - Y_i|^2 \\
&\leq \left(1 - \frac{1}{2 \cdot n \cdot C_{Lip,n}}\right)^{t_n} \cdot \left( F_n(\mathbf{w}^{(0)}) - \min_{g:\mathbb{R}^d \to \mathbb{R}} \frac{1}{n} \sum_{i=1}^{n} |g(X_i) - Y_i|^2 \right).
\end{aligned}
$$

**Proof.** Application of Lemma 1 and Lemma 2 implies for any $t \in \{0, \ldots, t_n - 1\}$

$$
\begin{aligned}
&F_n(\mathbf{w}^{(t+1)}) - \min_{g:\mathbb{R}^d \to \mathbb{R}} \frac{1}{n} \sum_{i=1}^{n} |g(X_i) - Y_i|^2 \\
&\leq F_n(\mathbf{w}^{(t)}) - \frac{1}{2 \cdot C_{Lip,n}} \cdot \|(\nabla_{\mathbf{w}} F_n)(\mathbf{w}^{(t)})\|^2 - \min_{g:\mathbb{R}^d \to \mathbb{R}} \frac{1}{n} \sum_{i=1}^{n} |g(X_i) - Y_i|^2 \\
&\leq \left(1 - \frac{1}{2 \cdot n \cdot C_{Lip,n}}\right) \cdot \left( F_n(\mathbf{w}^{(t)}) - \min_{g:\mathbb{R}^d \to \mathbb{R}} \frac{1}{n} \sum_{i=1}^{n} |g(X_i) - Y_i|^2 \right).
\end{aligned}
$$

From this we can conclude

$$
\begin{aligned}
&F_n(\mathbf{w}^{(t_n)}) - \min_{g:\mathbb{R}^d \to \mathbb{R}} \frac{1}{n} \sum_{i=1}^{n} |g(X_i) - Y_i|^2 \\
&\leq \left(1 - \frac{1}{2 \cdot n \cdot C_{Lip,n}}\right) \cdot \left( F_n(\mathbf{w}^{(t_n - 1)}) - \min_{g:\mathbb{R}^d \to \mathbb{R}} \frac{1}{n} \sum_{i=1}^{n} |g(X_i) - Y_i|^2 \right) \\
&\leq \left(1 - \frac{1}{2 \cdot n \cdot C_{Lip,n}}\right)^2 \cdot \left( F_n(\mathbf{w}^{(t_n - 2)}) - \min_{g:\mathbb{R}^d \to \mathbb{R}} \frac{1}{n} \sum_{i=1}^{n} |g(X_i) - Y_i|^2 \right)
\end{aligned}
$$

$$\leq \dots$$
$$\leq \left(1 - \frac{1}{2 \cdot n \cdot C_{Lip,n}}\right)^{t_n} \cdot \left(F_n(\mathbf{w}^{(0)}) - \min_{g:\mathbb{R}^d \to \mathbb{R}} \frac{1}{n} \sum_{i=1}^{n} |g(X_i) - Y_i|^2\right).$$

□

**Lemma 4** *Let $F : \mathbb{R}^K \to \mathbb{R}_+$ be differentiable, let $t_n \in \mathbb{N}$, and let $C_{Lip,n} > 0$ be such that*

$$\|(\nabla_{\mathbf{a}} F)(\mathbf{a})\|_\infty \leq C_{Lip,n} \cdot c_3 \cdot n^{c_4} \quad \textit{holds for all } \mathbf{a} \textit{ with } \|\mathbf{a}\|_\infty \leq 2 \cdot c_3 \cdot n^{c_4} \tag{15}$$

*and*

$$\|(\nabla_{\mathbf{a}} F)(\mathbf{a}_1) - (\nabla_{\mathbf{a}} F)(\mathbf{a}_2)\| \leq C_{Lip,n} \cdot \|\mathbf{a}_1 - \mathbf{a}_2\| \tag{16}$$

*holds for all $\mathbf{a}_1$, $\mathbf{a}_2$ with $\|\mathbf{a}_1\|_\infty \leq 3 \cdot c_3 \cdot n^{c_4}$ and $\|\mathbf{a}_2\|_\infty \leq 3 \cdot c_3 \cdot n^{c_4}$. Let $\mathbf{a}^{(0)}$ be such that*

$$\|\mathbf{a}^{(0)}\| \leq c_3 \cdot n^{c_4} \tag{17}$$

*and*

$$\sqrt{\frac{2 \cdot t_n}{C_{Lip,n}} \cdot F(\mathbf{a}^{(0)})} \leq c_3 \cdot n^{c_4}, \tag{18}$$

*and set*

$$\mathbf{a}^{(t+1)} = \mathbf{a}^{(t)} - \lambda_n \cdot (\nabla_{\mathbf{a}} F)(\mathbf{a}^{(t)}) \quad (t \in \{0, 1, \dots, t_n - 1\}),$$

*where*

$$\lambda_n = \frac{1}{C_{Lip,n}}.$$

*Then we have*

$$\|\mathbf{a}^{(t)}\|_\infty \leq 2 \cdot c_3 \cdot n^{c_4} \quad (t \in \{0, 1, \dots, t_n\}.$$

**Proof.** We show

$$\|\mathbf{a}^{(s)}\|_\infty \leq 2 \cdot c_3 \cdot n^{c_4} \quad (s \in \{0, \dots, t\}) \tag{19}$$

for all $t \in \{0, 1, \dots, t_n\}$ by induction.

For $t = 0$ the assertion follows from (17). So assume that (19) holds for some $t \in \{0, 1, \dots, t_n - 1\}$. Then this together with (15) implies that we have

$$\|\mathbf{a}^{(t+1)}\|_\infty \leq \|\mathbf{a}^{(t)}\|_\infty + \frac{1}{C_{Lip,n}} \cdot \|(\nabla_{\mathbf{a}} F)(\mathbf{a}^{(t)})\|_\infty \leq 3 \cdot c_3 \cdot n^{c_4}.$$

From this, the induction hypothesis and Lemma 1 we can conclude

$$0 \leq F(\mathbf{a}^{(s)}) \leq F(\mathbf{a}^{(s-1)}) - \frac{1}{2 \cdot C_{Lip,n}} \cdot \|(\nabla_{\mathbf{a}} F)(\mathbf{a}^{(s-1)})\|^2$$

for all $s \in \{0, \dots, t+1\}$ which implies

$$0 \leq F(\mathbf{a}^{(t+1)}) \leq F(\mathbf{a}^{(0)}) - \sum_{s=1}^{t+1} \frac{1}{2 \cdot C_{Lip,n}} \cdot \|(\nabla_{\mathbf{a}} F)(\mathbf{a}^{(s-1)})\|^2.$$

Consequently we have

$$\sum_{s=1}^{t+1} \frac{1}{2 \cdot C_{Lip,n}} \cdot \|(\nabla_{\mathbf{a}} F)(\mathbf{a}^{(s-1)})\|^2 \leq F(\mathbf{a}^{(0)}),$$

which implies

$$\begin{aligned}
\|\mathbf{a}^{(t+1)}\|_\infty &\leq \|\mathbf{a}^{(t+1)}\| \\
&\leq \|\mathbf{a}^{(0)}\| + \sum_{s=1}^{t+1} \frac{1}{C_{Lip,n}} \cdot \|(\nabla_{\mathbf{a}} F)(\mathbf{a}^{(s-1)})\| \\
&\leq \|\mathbf{a}^{(0)}\| + \sqrt{\frac{t+1}{C_{Lip,n}}} \cdot \sqrt{\sum_{s=1}^{t+1} \frac{1}{C_{Lip,n}} \cdot \|(\nabla_{\mathbf{a}} F)(\mathbf{a}^{(s-1)})\|^2} \\
&\leq \|\mathbf{a}^{(0)}\| + \sqrt{\frac{t+1}{C_{Lip,n}} \cdot 2 \cdot F(\mathbf{a}^{(0)})} \\
&\leq 2 \cdot c_3 \cdot n^{c_4},
\end{aligned}$$

where the last inequality followed from (17) and (18). □

**Lemma 5** *Let $\sigma$ be the logistic squasher. Let $k_n \in \mathbb{N}$ and $r_0 \in \mathbb{N}$ with $2 \cdot r_0 \geq d$. Let $\mathbf{w} = (w_{k,i,j}^{(s)})_{k,i,j,s}$ and $\bar{\mathbf{w}} = (\bar{w}_{k,i,j}^{(s)})_{k,i,j,s}$ be weight vectors and define $f_{\mathbf{w}}$ and $f_{\bar{\mathbf{w}}}$ by*

$$f_{\mathbf{w}}(x) = \sum_{i=1}^{k_n} w_{1,1,i}^{(L)} \cdot f_{i,1}^{(L)}(x) + w_{1,1,0}^{(L)} \quad \text{and} \quad f_{\bar{\mathbf{w}}}(x) = \sum_{i=1}^{k_n} \bar{w}_{1,1,i}^{(L)} \cdot \bar{f}_{i,1}^{(L)}(x) + \bar{w}_{1,1,0}^{(L)} \tag{20}$$

*for $f_{i,i}^{(L)}$'s and $\bar{f}_{i,i}^{(L)}$'s recursively defined by*

$$f_{k,i}^{(l)}(x) = \sigma\left(\sum_{j=1}^{r_0} w_{k,i,j}^{(l-1)} \cdot f_{k,j}^{(l-1)}(x) + w_{k,i,0}^{(l-1)}\right) \tag{21}$$

*$(l = 2, \ldots, L)$ and*

$$\bar{f}_{k,i}^{(l)}(x) = \sigma\left(\sum_{j=1}^{r_0} \bar{w}_{k,i,j}^{(l-1)} \cdot \bar{f}_{k,j}^{(l-1)}(x) + \bar{w}_{k,i,0}^{(l-1)}\right) \tag{22}$$

*$(l = 2, \ldots, L)$ and*

$$f_{k,i}^{(1)}(x) = \sigma\left(\sum_{j=1}^{d} w_{k,i,j}^{(0)} \cdot x^{(j)} + w_{k,i,0}^{(0)}\right) \quad \text{and} \quad \bar{f}_{k,i}^{(1)}(x) = \sigma\left(\sum_{j=1}^{d} \bar{w}_{k,i,j}^{(0)} \cdot x^{(j)} + \bar{w}_{k,i,0}^{(0)}\right). \tag{23}$$

**a)** *For any* $k \in \{1, \dots, k_n\}$ *and any* $x \in \mathbb{R}^d$ *we have*

$$|f_{k,1}^{(L)}(x) - \bar{f}_{k,1}^{(L)}(x)| \leq (2 \cdot r_0 + 1)^L \cdot (\max\{\|\mathbf{w}\|_\infty, \|x\|_\infty, 1\})^L \cdot \max_{i,j,s:s<L} |w_{k,i,j}^{(s)} - \bar{w}_{k,i,j}^{(s)}|.$$

**b)** *For any* $x \in \mathbb{R}^d$ *we have*

$$|f_{\mathbf{w}}(x) - f_{\bar{\mathbf{w}}}(x)| \leq (2 \cdot k_n + 1) \cdot (2 \cdot r_0 + 1)^L \cdot (\max\{\|\mathbf{w}\|_\infty, \|x\|_\infty, 1\})^{L+1} \cdot \|\mathbf{w} - \bar{\mathbf{w}}\|_\infty.$$

**Proof. a)** We show by induction

$$|f_{r,k}^{(l)}(x) - \bar{f}_{r,k}^{(l)}(x)| \leq (2 \cdot r_0 + 1)^l \cdot (\max\{\|\mathbf{w}\|_\infty, \|x\|_\infty, 1\})^l \cdot \max_{i,j,s:s<L} |w_{k,i,j}^{(s)} - \bar{w}_{k,i,j}^{(s)}| \quad (24)$$

($l \in \{1, \dots, L\}$). The logistic squasher satisfies $|\sigma'(x)| = |\sigma(x) \cdot (1 - \sigma(x))| \leq 1$, hence it is Lipschitz continuous with Lipschitz constant one. This implies

$$\begin{aligned} \left| f_{k,i}^{(1)}(x) - \bar{f}_{k,i}^{(1)}(x) \right| &\leq \sum_{j=1}^{d} |w_{k,i,j}^{(0)} - \bar{w}_{k,i,j}^{(0)}| \cdot |x^{(j)}| + |w_{k,i,0}^{(0)} - \bar{w}_{k,i,0}^{(0)}| \\ &\leq (2 \cdot r_0 + 1) \cdot \max\{\|x\|_\infty, 1\} \cdot \max_{i,j,s:s<L} |w_{k,i,j}^{(s)} - \bar{w}_{k,i,j}^{(s)}|. \end{aligned}$$

Assume now that (24) holds for some $l - 1$, where $l \in \{2, \dots, L\}$. Then

$$\begin{aligned} &\left| f_{k,i}^{(l)}(x) - \bar{f}_{k,i}^{(l)}(x) \right| \\ &\leq \sum_{j=1}^{r_0} |w_{k,i,j}^{(l-1)}| \cdot |f_{k,j}^{(l-1)}(x) - \bar{f}_{k,j}^{(l-1)}(x)| + \sum_{j=1}^{r_0} |w_{k,i,j}^{(l-1)} - \bar{w}_{k,i,j}^{(l-1)}(x)| \cdot |\bar{f}_{k,j}^{(l-1)}(x)| \\ &\quad + |w_{k,i,0}^{(l-1)} - \bar{w}_{k,i,0}^{(l-1)}| \\ &\leq r_0 \cdot \|\mathbf{w}\|_\infty \cdot \max_{j=1,\dots,r_0} |f_{k,j}^{(l-1)}(x) - \bar{f}_{k,j}^{(l-1)}(x)| + (r_0 + 1) \cdot \max_{i,j,s:s<L} |w_{k,i,j}^{(s)} - \bar{w}_{k,i,j}^{(s)}| \\ &\leq (2r_0 + 1) \cdot \max\{\|\mathbf{w}\|_\infty, \|x\|_\infty, 1\} \\ &\qquad \cdot \max\left\{ \max_{j=1,\dots,r_0} |f_{k,j}^{(l-1)}(x) - \bar{f}_{k,j}^{(l-1)}(x)|, \max_{i,j,s:s<L} |w_{k,i,j}^{(s)} - \bar{w}_{k,i,j}^{(s)}| \right\} \\ &\leq (2 \cdot r_0 + 1)^r \cdot (\max\{\|\mathbf{w}\|_\infty, \|x\|_\infty, 1\})^r \cdot \max_{i,j,s:s<L} |w_{k,i,j}^{(s)} - \bar{w}_{k,i,j}^{(s)}|. \end{aligned}$$

**b)** Because of

$$\begin{aligned} &|f_{\mathbf{w}}(x) - \bar{f}_{\mathbf{w}}(x)| \\ &= \left| \sum_{i=1}^{k_n} w_{1,1,i}^{(L)} \cdot f_{i,1}^{(L)}(x) + w_{1,1,0}^{(L)} - \sum_{i=1}^{k_n} \bar{w}_{1,1,i}^{(L)} \cdot \bar{f}_{i,1}^{(L)}(x) - \bar{w}_{1,1,0}^{(L)} \right| \\ &\leq \left| \sum_{i=1}^{k_n} w_{1,1,i}^{(L)} \cdot (f_{i,1}^{(L)}(x) - \bar{f}_{i,1}^{(L)}(x)) \right| + \left| \sum_{i=1}^{k_n} (w_{1,1,i}^{(L)} - \bar{w}_{1,1,i}^{(L)}) \cdot \bar{f}_{i,1}^{(L)}(x) + w_{1,1,0}^{(L)} - \bar{w}_{1,1,0}^{(L)} \right| \end{aligned}$$

$$\leq k_n \cdot \max_i |w_{1,1,i}^{(L)}| \cdot \max_i |f_{i,1}^{(L)}(x) - \bar{f}_{i,1}^{(L)}(x)| + (k_n + 1) \cdot \max_i |w_{1,1,i}^{(L)} - \bar{w}_{1,1,i}^{(L)}|,$$

the assertion follows from a). $\square$

**Lemma 6** *Let $\sigma$ be the logistic squasher. Define $f_{\mathbf{w}}$ by (8)-(10) and set*

$$F_n(\mathbf{w}) = \frac{1}{n} \sum_{i=1}^n |f_{\mathbf{w}}(X_i) - Y_i|^2.$$

*Let $c_3, c_4 \geq 1$. Assume $\|\mathbf{w}_1\|_\infty \leq c_3 \cdot n^{c_4}$, $\|\mathbf{w}_2\|_\infty \leq c_3 \cdot n^{c_4}$ and*

$$\max_{i=1,\dots,n} \|X_i\|_\infty \leq c_3 \cdot n^{c_4} \quad \textit{and} \quad \max_{i=1,\dots,n} |Y_i| \leq c_3 \cdot n^{c_4}.$$

*Set*

$$C_{Lip,n} = 45 \cdot L \cdot 3^L \cdot (\max\{r_0, L, d\})^{3/2} \cdot r_0^{2L} \cdot k_n^{3/2} \cdot (c_3 \cdot n^{c_4})^{4L+1}.$$

*Then we have*

$$\|(\nabla_{\mathbf{w}} F_n)(\mathbf{w}_1)\|_\infty \leq C_{Lip,n} \cdot c_3 \cdot n^{c_4} \tag{25}$$

*and*

$$\|(\nabla_{\mathbf{w}} F_n)(\mathbf{w}_1) - (\nabla_{\mathbf{w}} F_n)(\mathbf{w}_2)\| \leq C_{Lip,n} \cdot \|\mathbf{w}_1 - \mathbf{w}_2\|. \tag{26}$$

**Proof.** *In the first step of the proof* we compute the partial derivatives of $F_n(\mathbf{w})$. We have

$$\frac{\partial}{\partial w_{k,i,j}^{(r)}} F_n(\mathbf{w}) = \frac{2}{n} \sum_{l=1}^n (f_{\mathbf{w}}(X_l) - Y_l) \cdot \frac{\partial f_{\mathbf{w}}}{\partial w_{k,i,j}^{(r)}}(X_l).$$

The recursive definition of $f_{\mathbf{w}}$ together with the chain rule imply

$$\frac{\partial f_{\mathbf{w}}}{\partial w_{1,1,i}^{(L)}}(X_l) = f_{i,1}^{(L)}(X_l)$$

(where we have set $f_{0,0}^{(L)}(x) = 1$) and in case $\bar{r} < L$

$$\frac{\partial f_{\mathbf{w}}}{\partial w_{\bar{k},\bar{i},\bar{j}}^{(\bar{r})}}(X_l) = \sum_{i=1}^{k_n} w_{1,1,i}^{(L)} \cdot \frac{\partial f_{i,1}^{(L)}}{\partial w_{\bar{k},\bar{i},\bar{j}}^{(\bar{r})}}(X_l) = w_{1,1,\bar{k}}^{(L)} \cdot \frac{\partial f_{\bar{k},1}^{(L)}}{\partial w_{\bar{k},\bar{i},\bar{j}}^{(\bar{r})}}(X_l).$$

In case $0 \leq \bar{r} < r$ and $r > 1$ we have

$$\begin{aligned} & \frac{\partial f_{k,i}^{(r)}}{\partial w_{k,\bar{i},\bar{j}}^{(\bar{r})}}(X_l) \\ = \quad & \sigma' \left( \sum_{j=1}^{r_0} w_{k,i,j}^{(r-1)} \cdot f_{k,j}^{(r-1)}(X_l) + w_{k,i,0}^{(r-1)} \right) \end{aligned}$$

$$\cdot \frac{\partial}{\partial w_{k,\bar{i},\bar{j}}^{(\bar{r})}} \left( \sum_{j=1}^{r_0} w_{k,i,j}^{(r-1)} \cdot f_{k,j}^{(r-1)}(X_l) + w_{k,i,0}^{(r-1)} \right)$$

$$= \sigma \left( \sum_{j=1}^{r_0} w_{k,i,j}^{(r-1)} \cdot f_{k,j}^{(r-1)}(X_l) + w_{k,i,0}^{(r-1)} \right)$$

$$\cdot \left( 1 - \sigma \left( \sum_{j=1}^{r_0} w_{k,i,j}^{(r-1)} \cdot f_{k,j}^{(r-1)}(X_l) + w_{k,i,0}^{(r-1)} \right) \right)$$

$$\cdot \frac{\partial}{\partial w_{k,\bar{i},\bar{j}}^{(\bar{r})}} \left( \sum_{j=1}^{r_0} w_{k,i,j}^{(r-1)} \cdot f_{k,j}^{(r-1)}(X_l) + w_{k,i,0}^{(r-1)} \right).$$

Next we explain how we can compute

$$\frac{\partial}{\partial w_{k,\bar{i},\bar{j}}^{(\bar{r})}} \left( \sum_{j=1}^{r_0} w_{k,i,j}^{(r-1)} \cdot f_{k,j}^{(r-1)}(X_l) + w_{k,i,0}^{(r-1)} \right).$$

In case $\bar{r} = r - 1 > 0$ we have

$$\frac{\partial}{\partial w_{k,\bar{i},\bar{j}}^{(r-1)}} \left( \sum_{j=1}^{r_0} w_{k,i,j}^{(r-1)} \cdot f_{k,j}^{(r-1)}(X_l) + w_{k,i,0}^{(r-1)} \right) = f_{k,\bar{j}}^{(r-1)}(X_l) \cdot 1_{\{\bar{i}=i\}}$$

(where we have set $f_{\bar{k},0}^{(r-1)}(x) = 1$), and in case $\bar{r} < r - 1$ we get

$$\frac{\partial}{\partial w_{k,\bar{i},\bar{j}}^{(\bar{r})}} \left( \sum_{j=1}^{r_0} w_{k,i,j}^{(r-1)} \cdot f_{k,j}^{(r-1)}(X_l) + w_{k,i,0}^{(r-1)} \right) = \sum_{j=1}^{r_0} w_{k,i,j}^{(r-1)} \cdot \frac{\partial}{\partial w_{k,\bar{i},\bar{j}}^{(\bar{r})}} f_{k,j}^{(r-1)}(X_l).$$

And in case $r = 2$ and $\bar{r} = 0$ we have

$$\frac{\partial f_{k,i}^{(1)}}{\partial w_{k,\bar{i},\bar{j}}^{(0)}}(X_l)$$

$$= \sigma' \left( \sum_{j=1}^{d} w_{k,i,j}^{(0)} \cdot X_l^{(j)} + w_{k,i,0}^{(0)} \right) \cdot X_l^{(\bar{j})} \cdot 1_{\{\bar{i}=i\}}$$

$$= \sigma \left( \sum_{j=1}^{d} w_{k,i,j}^{(0)} \cdot X_l^{(j)} + w_{k,i,0}^{(0)} \right) \cdot \left( 1 - \sigma \left( \sum_{j=1}^{d} w_{k,i,j}^{(0)} \cdot X_l^{(j)} + w_{k,i,0}^{(0)} \right) \right) \cdot X_l^{(\bar{j})} \cdot 1_{\{\bar{i}=i\}},$$

where we have set $X_l^{(0)} = 1$.

*In the second step of the proof* we show for $x \in \mathbb{R}^d$ with $\|x\|_\infty \leq c_3 \cdot n^{c_4}$ and $\mathbf{w}$, $\mathbf{w}_1$, $\mathbf{w}_2$ with $\|\mathbf{w}\|_\infty \leq c_3 \cdot n^{c_4}$, $\|\mathbf{w}_1\|_\infty \leq c_3 \cdot n^{c_4}$ and $\|\mathbf{w}_2\|_\infty \leq c_3 \cdot n^{c_4}$,

$$\left| \frac{\partial f_{\mathbf{w}}(x)}{\partial w_{k,i,j}^{(r)}} \right| \leq r_0^L \cdot (c_3 \cdot n^{c_4})^{L+1} \tag{27}$$

and

$$\left| \frac{\partial f_{\mathbf{w}_1}(x)}{\partial w_{k,i,j}^{(r)}} - \frac{\partial f_{\mathbf{w}_2}(x)}{\partial w_{k,i,j}^{(r)}} \right| \leq \bar{C}_{Lip,n} \cdot \|\mathbf{w}_1 - \mathbf{w}_2\|_\infty, \tag{28}$$

where

$$\bar{C}_{Lip,n} = 4L \cdot 3^L \cdot r_0^{2L-2} \cdot (c_3 \cdot n^{c_4})^{4L}.$$

It is easy to see that the first step of the proof implies

$$\begin{aligned}
\frac{\partial f_{\mathbf{w}}(x)}{\partial w_{k,i,j}^{(r)}} &= \sum_{s_{r+1}=1}^{r_0} \cdots \sum_{s_{L-2}=1}^{r_0} f_{k,j}^{(r)}(x) \cdot f_{k,i}^{(r+1)}(x) \cdot (1 - f_{k,i}^{(r+1)}(x)) \\
&\cdot w_{k,s_{r+1},i}^{(r+1)} \cdot f_{k,s_{r+1}}^{(r+2)}(x) \cdot (1 - f_{k,s_{r+1}}^{(r+2)}(x)) \cdot w_{k,s_{r+2},s_{r+1}}^{(r+2)} \cdot f_{k,s_{r+2}}^{(r+3)}(x) \cdot (1 - f_{k,s_{r+2}}^{(r+3)}(x)) \\
&\cdots w_{k,s_{L-2},s_{L-3}}^{(L-2)} \cdot f_{k,s_{L-2}}^{(L-1)}(x) \cdot (1 - f_{k,s_{L-2}}^{(L-1)}(x)) \cdot w_{k,k,s_{L-2}}^{(L-1)} \cdot f_{k,1}^{(L)}(x) \cdot (1 - f_{k,1}^{(L)}(x)) \\
&\cdot w_{1,1,k}^{(L)},
\end{aligned} \tag{29}$$

where we have used the abbreviations

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

and

$$f_{k,0}^{(r)}(x) = 1.$$

Because of

$$f_{k,i}^{(r)}(x) \in [0,1] \quad \text{if } r > 0$$

and

$$|f_{k,i}^{(0)}(x)| \leq c_3 \cdot n^{c_4}$$

and

$$\|\mathbf{w}\|_\infty \leq c_3 \cdot n^{c_4}$$

this implies (27).

Next we prove (28). The right-hand side of (29) is a sum of at most $r_0^{L-2}$ products, where each product contains at most $3L+1$ factors. In the worst case from these $3L+1$ factors $L$ are Lipschitz continuous functions with Lipschitz constant bounded by one, which are bounded in absolute value by $c_3 \cdot n^{c_4}$. And according to the proof of Lemma 5 (cf., (24)) the remaining $2L+1$ factors are Lipschitz continuous functions with Lipschitz constant bounded by

$$(2r_0+1)^L \cdot (c_3 \cdot n^{c_4})^L,$$

which are bounded in absolute value by $c_3 \cdot n^{c_4}$.

If $g_1, \dots, g_s : \mathbb{R} \to \mathbb{R}$ are Lipschitz continuous functions with Lipschitz constants $C_{Lip,g_1}, \dots, C_{Lip,g_s}$, then

$$\prod_{l=1}^{s} g_l \quad \text{and} \quad \sum_{l=1}^{s} g_l$$

are Lipschitz continuous functions with Lipschitz constant bounded by

$$\sum_{l=1}^{s} C_{Lip,g_l} \cdot \prod_{k \in \{1,\dots,s\} \setminus \{l\}} \|g_k\|_\infty \leq s \cdot \max_l C_{Lip,g_l} \cdot (\max_k \|g_k\|_\infty)^{s-1}$$

and by

$$\sum_{l=1}^{s} C_{Lip,g_l} \leq s \cdot \max_l C_{Lip,g_l},$$

respectively. This implies that (29) is Lipschitz continuous with Lipschitz constant bounded by

$$r_0^{L-2} \cdot (3L+1) \cdot (2r_0+1)^L \cdot (c_3 n^{c_4})^L \cdot (c_3 n^{c_4})^{3L}.$$

In the *third step of the proof* we show (25). We have

$$\begin{aligned}
\|(\nabla_{\mathbf{w}} F_n)(\mathbf{w})\|_\infty &= \max_{k,i,j,r} \left| \frac{2}{n} \sum_{l=1}^{n} (f_{\mathbf{w}}(X_l) - Y_l) \cdot \frac{\partial f_{\mathbf{w}}}{\partial w_{k,i,j}^{(r)}}(X_i) \right| \\
&\leq 2 \cdot \left( (k_n+1) \cdot \|\mathbf{w}\|_\infty + \max_{i=1,\dots,n} |Y_i| \right) \cdot \max_{l,k,i,j,r} \left| \frac{\partial f_{\mathbf{w}}}{\partial w_{k,i,j}^{(r)}}(X_l) \right| \\
&\leq 6 \cdot k_n \cdot c_3 n^{c_4} \cdot \max_{l,k,i,j,r} \left| \frac{\partial f_{\mathbf{w}}}{\partial w_{k,i,j}^{(r)}}(X_l) \right|.
\end{aligned}$$

From this the result follows by (27).

*In the fourth step of the proof* we show (26). Because of

$$\|(\nabla_{\mathbf{w}} F_n)(\mathbf{w}_1) - (\nabla_{\mathbf{w}} F_n)(\mathbf{w}_2)\| = \left( \sum_{k,i,j,r} \left| \frac{\partial F_n}{\partial w_{k,i,j}^{(r)}}(\mathbf{w}_1) - \frac{\partial F_n}{\partial w_{k,i,j}^{(r)}}(\mathbf{w}_2) \right|^2 \right)^{1/2}$$

and

$$\frac{\partial F_n}{\partial w_{k,i,j}^{(r)}}(\mathbf{w}) = \frac{2}{n} \sum_{l=1}^{n} (f_{\mathbf{w}}(X_l) - Y_l) \cdot \frac{\partial f_{\mathbf{w}}}{\partial w_{k,i,j}^{(r)}}(X_l)$$

we have

$$\begin{aligned}
&\|(\nabla_{\mathbf{w}} F_n)(\mathbf{w}_1) - (\nabla_{\mathbf{w}} F_n)(\mathbf{w}_2)\| \\
&\leq \sqrt{k_n \cdot (r_0 + 1 + (L-2) \cdot (r_0^2 + r_0) + r_0 \cdot (d+1)) + k_n + 1}
\end{aligned}$$

$$\cdot 2 \cdot \max_{k,i,j,r,l} \left| (f_{\mathbf{w}_1}(X_l) - Y_l) \cdot \frac{\partial f_{\mathbf{w}_1}}{\partial w_{k,i,j}^{(r)}}(X_l) - (f_{\mathbf{w}_2}(X_l) - Y_l) \cdot \frac{\partial f_{\mathbf{w}_2}}{\partial w_{k,i,j}^{(r)}}(X_l) \right|. \quad (30)$$

By Lemma 5 we know

$$|f_{\mathbf{w}_1}(X_l) - f_{\mathbf{w}_2}(X_l)| \leq (2k_n + 1) \cdot (2r_0 + 1)^L \cdot (c_3 n^{c_4})^{L+1} \cdot \|\mathbf{w}_1 - \mathbf{w}_2\|_\infty. \quad (31)$$

Trivially,

$$|f_{\mathbf{w}}(X_l) - Y_l| \leq (k_n + 1) \cdot c_3 n^{c_4} + c_3 n^{c_4} = (k_n + 2) \cdot c_3 n^{c_4}. \quad (32)$$

If $g_i$ are Lipschitz continuous functions with Lipschitz constants $C_{Lip,g_i}$, then $g_1 \cdot g_2$ is Lipschitz continuous with Lipschitz constant

$$\|g_1\|_\infty \cdot C_{Lip,g_2} + \|g_2\|_\infty \cdot C_{Lip,g_1}.$$

Combining this with (27), (28), (31) and (32) we get that

$$\mathbf{w} \mapsto (f_{\mathbf{w}}(X_l) - Y_l) \cdot \frac{\partial f_{\mathbf{w}}}{\partial w_{k,i,j}^{(r)}}(X_l)$$

is Lipschitz continuous with Lipschitz constant bounded by

$$\begin{aligned} &(k_n + 2) \cdot c_3 \cdot n^{c_4} \cdot 4L \cdot 3^L \cdot r_0^{2L-2} \cdot (c_3 \cdot n^{c_4})^{4L} \\ &\qquad + r_0^L \cdot (c_3 n^{c_4})^{L+1} \cdot (2k_n + 1) \cdot (2r_0 + 1)^L \cdot (c_3 n^{c_4})^{L+1} \\ &\leq 15 \cdot k_n \cdot L \cdot 3^L \cdot r_0^{2L} \cdot (c_3 n^{c_4})^{4L+1}. \end{aligned}$$

This together with (30) implies the assertion. □

**Lemma 7** *Let $\sigma$ be the logistic squasher and let $n, d, r_0, L \in \mathbb{N}$ with $r_0 \geq 2 \cdot d$ and $L \geq 2$. Define $f_{1,1}^{(L)} : \mathbb{R} \to \mathbb{R}$ recursively by*

$$f_{1,k}^{(r)}(x) = \sigma \left( \sum_{j=1}^{r_0} w_{1,k,j}^{(r-1)} \cdot f_{1,j}^{(r-1)}(x) + w_{1,k,0}^{(r-1)} \right)$$

*for some $w_{1,k,0}^{(r-1)}, \dots, w_{1,k,r_0}^{(r-1)} \in \mathbb{R}$ $(r = 2, \dots, L)$ and*

$$f_{1,k}^{(1)}(x) = \sigma \left( \sum_{j=1}^{d} w_{1,k,j}^{(0)} \cdot x^{(j)} + w_{1,k,0}^{(0)} \right)$$

*for some $w_{1,k,0}^{(0)}, \dots, w_{1,k,d}^{(0)} \in \mathbb{R}$. Let $\delta > 0$ and let $\mathbf{a}$, $\mathbf{b} \in \mathbb{R}^d$ such that*

$$b^{(l)} - a^{(l)} \geq 2 \cdot \delta \quad \text{for all } l \in \{1, \dots, d\}.$$

*Assume*

$$w_{1,1,1}^{(L-1)} \leq -4 \cdot (n + 1), \quad (33)$$

$$|w_{1,1,j}^{(L-1)} - w_{1,1,1}^{(L-1)}| \leq \frac{1}{2r_0} \quad \textit{for } j = 2, \dots, d, \tag{34}$$

$$|w_{1,1,j}^{(L-1)}| \leq \frac{1}{2r_0} \quad \textit{for } j = 2d+1, \dots, r_0, \tag{35}$$

$$|w_{1,k,0}^{(L-1)} + \frac{1}{2} \cdot w_{1,1,1}^{(L-1)}| \leq \frac{1}{2} \quad \textit{for } k \in \{1, \dots, 2d\}, \tag{36}$$

$$w_{1,k,k}^{(r-1)} \geq 8 \cdot \log(8d-1) \quad \textit{for } k \in \{1, \dots, 2d\} \textit{ and } r \in \{2, \dots, L-1\}, \tag{37}$$

$$|w_{1,k,0}^{(r-1)} + \frac{1}{2} \cdot w_{1,k,k}^{(r-1)}| \leq \frac{\log(8d-1)}{r_0} \quad \textit{for } k \in \{1, \dots, 2d\} \textit{ and } r \in \{2, \dots, L-1\}, \tag{38}$$

$$|w_{1,k,j}^{(r-1)}| \leq \frac{\log(8d-1)}{r_0} \quad \textit{for } j \in \{1, \dots, r_0\} \setminus \{k\}, k \in \{1, \dots, 2d\}, r \in \{2, \dots, L-1\}, \tag{39}$$

$$w_{1,k,k}^{(0)} \leq -\frac{2}{\delta} \cdot \log(8d-1) \quad \textit{for } k \in \{1, \dots, d\}, \tag{40}$$

$$|w_{1,k,0}^{(0)} + a^{(k)} \cdot w_{1,k,k}^{(0)}| \leq \frac{\log(8d-1)}{d} \quad \textit{for } k \in \{1, \dots, d\}, \tag{41}$$

$$|w_{1,k,j}^{(0)}| \leq \frac{\log(8d-1)}{d} \quad \textit{for } k \in \{1, \dots, d\}, j \in \{1, \dots, d\} \setminus \{k\} \tag{42}$$

$$w_{1,d+k,k}^{(0)} \geq \frac{2}{\delta} \cdot \log(8d-1) \quad \textit{for } k \in \{1, \dots, d\}, \tag{43}$$

$$|w_{1,d+k,0}^{(0)} + b^{(k)} \cdot w_{1,d+k,k}^{(0)}| \leq \frac{\log(8d-1)}{d} \quad \textit{for } k \in \{1, \dots, d\} \tag{44}$$

*and*

$$|w_{1,d+k,j}^{(0)}| \leq \frac{\log(8d-1)}{d} \quad \textit{for } k \in \{1, \dots, d\}, j \in \{1, \dots, d\} \setminus \{k\}. \tag{45}$$

*Then* $f_{1,1}^{(L)}$ *satisfies for any* $x \in [-1,1]^d$

$$f_{1,1}^{(L)}(x) \geq 1 - e^{-n} \quad \textit{if } x \in [a^{(1)} + \delta, b^{(1)} - \delta] \times \dots \times [a^{(d)} + \delta, b^{(d)} - \delta] \tag{46}$$

*and*

$$f_{1,1}^{(L)}(x) \leq e^{-n} \quad \textit{if } x \notin [a^{(1)} - \delta, b^{(1)} + \delta] \times \dots \times [a^{(d)} - \delta, b^{(d)} + \delta]. \tag{47}$$

**Proof.** Let $x \in [a^{(1)} + \delta, b^{(1)} - \delta] \times \dots \times [a^{(d)} + \delta, b^{(d)} - \delta] \cap [-1,1]^d$. Then we get for any $k \in \{1, \dots, d\}$ by (40), (41) and (42)

$$\begin{aligned} &\sum_{j=1}^{d} w_{1,k,j}^{(0)} \cdot x^{(j)} + w_{1,k,0}^{(0)} \\ &= w_{1,k,k}^{(0)} \cdot (x^{(k)} - a^{(k)}) + w_{1,k,0}^{(0)} + w_{1,k,k}^{(0)} \cdot a^{(k)} + \sum_{j \in \{1,\dots,d\} \setminus \{k\}} w_{1,k,j}^{(0)} \cdot x^{(j)} \end{aligned}$$

$$\leq -2 \cdot \log(8d-1) + |w_{1,k,0}^{(0)} + w_{1,k,k}^{(0)} \cdot a^{(k)}| + \sum_{j \in \{1,\dots,d\} \setminus \{k\}} |w_{1,k,j}^{(0)}|$$
$$\leq -\log(8d-1).$$

And by (43), (44) and (45) we get for any $k \in \{1, \dots, d\}$

$$\sum_{j=1}^{d} w_{1,d+k,j}^{(0)} \cdot x^{(j)} + w_{1,d+k,0}^{(0)}$$
$$= -w_{1,d+k,k}^{(0)} \cdot (b^{(k)} - x^{(k)}) + w_{1,d+k,0}^{(0)} + w_{1,d+k,k}^{(0)} \cdot b^{(k)} + \sum_{j \in \{1,\dots,d\} \setminus \{k\}} w_{1,d+k,j}^{(0)} \cdot x^{(j)}$$
$$\leq -2 \cdot \log(8d-1) + |w_{1,d+k,0}^{(0)} + w_{1,d+k,k}^{(0)} \cdot b^{(k)}| + \sum_{j \in \{1,\dots,d\} \setminus \{k\}} |w_{1,d+k,j}^{(0)}|$$
$$\leq -\log(8d-1).$$

It is easy to see that the logistic squasher satisfies

$$\sigma(x) \geq 1-\kappa \quad \text{if} \quad x \geq \log\left(\frac{1}{\kappa} - 1\right) \quad \text{and} \quad \sigma(x) \leq \kappa \quad \text{if} \quad x \leq -\log\left(\frac{1}{\kappa} - 1\right). \tag{48}$$

Using this we get for any $k \in \{1, \dots, 2d\}$

$$f_{1,k}^{(1)}(x) \leq \sigma(-\log(8d-1)) = \sigma\left(-\log\left(\frac{1}{1/(8d)} - 1\right)\right) \leq \frac{1}{8d} \leq \frac{1}{4}.$$

Using (37), (38) and (39), we can recursively conclude for $r = 2, \dots, L-1$ that we have for any $k \in \{1, \dots, 2d\}$

$$\sum_{j=1}^{r_0} w_{1,k,j}^{(r-1)} \cdot f_{1,j}^{(r-1)}(x) + w_{1,k,0}^{(r-1)}$$
$$= w_{1,k,k}^{(r-1)} \cdot \left(f_{1,k}^{(r-1)}(x) - \frac{1}{2}\right) + w_{1,k,k}^{(r-1)} \cdot \frac{1}{2} + w_{1,k,0}^{(r-1)} + \sum_{j \in \{1,\dots,r_0\} \setminus \{k\}} w_{1,k,j}^{(r-1)} \cdot f_{1,j}^{(r-1)}(x)$$
$$\leq -2 \cdot \log(8d-1) + |w_{1,k,k}^{(r-1)} \cdot \frac{1}{2} + w_{1,k,0}^{(r-1)}| + \sum_{j \in \{1,\dots,r_0\} \setminus \{k\}} |w_{1,k,j}^{(r-1)}|$$
$$\leq -\log(8d-1)$$

and

$$f_{1,k}^{(r)}(x) \leq \sigma(-\log(8d-1)) \leq \frac{1}{8d} \leq \frac{1}{4}.$$

From this together with (33), (34), (35) and (36) we conclude

$$\sum_{j=1}^{r_0} w_{1,1,j}^{(L-1)} \cdot f_{1,j}^{(L-1)}(x) + w_{1,1,0}^{(L-1)}$$

$$
\begin{aligned}
&= w_{1,1,1}^{(L-1)} \cdot (\sum_{j=1}^{2d} f_{1,j}^{(L-1)}(x) - \frac{1}{2}) + w_{1,1,0}^{(L-1)} + \frac{1}{2} \cdot w_{1,1,1}^{(L-1)} \\
&\quad + \sum_{j=1}^{2d} (w_{1,1,j}^{(L-1)} - w_{1,1,1}^{(L-1)}) \cdot f_{1,j}^{(L-1)}(x) + \sum_{j=2d+1}^{r_0} w_{1,1,j}^{(L-1)} \cdot f_{1,j}^{(L-1)}(x) \\
&\geq w_{1,1,1}^{(L-1)} \cdot (\sum_{j=1}^{2d} f_{1,j}^{(L-1)}(x) - \frac{1}{2}) - |w_{1,1,0}^{(L-1)} + \frac{1}{2} \cdot w_{1,1,1}^{(L-1)}| \\
&\quad - \sum_{j=1}^{2d} |w_{1,1,j}^{(L-1)} - w_{1,1,1}^{(L-1)}| - \sum_{j=2d+1}^{r_0} |w_{1,1,j}^{(L-1)}| \\
&\geq -4 \cdot (n+1) \cdot (2d \cdot \frac{1}{8d} - \frac{1}{2}) - \frac{1}{2} - \sum_{j=1}^{2d} \frac{1}{2r_0} - \sum_{j=2d+1}^{r_0} \frac{1}{2r_0} \\
&\geq n \geq \log(1/e^{-n} - 1),
\end{aligned}
$$

which implies (46).

In order to prove (47) we assume that $x \in [-1,1]^d$ satisfies $x^{(k)} \notin [a^{(k)} - \delta, b^{(k)} + \delta]$ for some $k \in \{1, \dots, d\}$. In case $x^{(k)} < a^{(k)} - \delta$ we can argue similarly as above and conclude recursively from (48) and (33)-(45)

$$
\begin{aligned}
&\sum_{j=1}^{d} w_{1,k,j}^{(0)} \cdot x^{(j)} + w_{1,k,0}^{(0)} \\
&= w_{1,k,k}^{(0)} \cdot (x^{(k)} - a^{(k)}) + w_{1,k,0}^{(0)} + w_{1,k,k}^{(0)} \cdot a^{(k)} + \sum_{j \in \{1,\dots,d\} \setminus \{k\}} w_{1,k,j}^{(0)} \cdot x^{(j)} \\
&\geq 2 \cdot \log(8d - 1) - |w_{1,k,0}^{(0)} + w_{1,k,k}^{(0)} \cdot a^{(k)}| - \sum_{j \in \{1,\dots,d\} \setminus \{k\}} |w_{1,k,j}^{(0)}| \\
&\geq \log(8d - 1),
\end{aligned}
$$

which implies

$$
f_{1,k}^{(1)}(x) \geq \sigma(\log(8d-1)) = \sigma(\log(1/(1/(8d)) - 1)) \geq 1 - \frac{1}{8d} \geq \frac{3}{4}.
$$

Recursively we can conclude for $r = 2, \dots, L-1$

$$
\begin{aligned}
&\sum_{j=1}^{r_0} w_{1,k,j}^{(r-1)} \cdot f_{1,j}^{(r-1)}(x) + w_{1,k,0}^{(r-1)} \\
&= w_{1,k,k}^{(r-1)} \cdot \left( f_{1,k}^{(r-1)}(x) - \frac{1}{2} \right) + w_{1,k,k}^{(r-1)} \cdot \frac{1}{2} + w_{1,k,0}^{(r-1)} + \sum_{j \in \{1,\dots,r_0\} \setminus \{k\}} w_{1,k,j}^{(r-1)} \cdot f_{1,j}^{(r-1)}(x) \\
&\geq 2 \cdot \log(8d - 1) - |w_{1,k,k}^{(r-1)} \cdot \frac{1}{2} + w_{1,k,0}^{(r-1)}| - \sum_{j \in \{1,\dots,r_0\} \setminus \{k\}} |w_{1,k,j}^{(r-1)}|
\end{aligned}
$$

$$\geq \log(8d-1)$$

and

$$f_{1,k}^{(r)}(x) \geq \sigma(\log(8d-1)) \geq 1 - \frac{1}{8d} \geq \frac{3}{4}.$$

This yields

$$\begin{aligned}
&\sum_{j=1}^{r_0} w_{1,1,j}^{(L-1)} \cdot f_{1,j}^{(L-1)}(x) + w_{1,1,0}^{(L-1)} \\
&= w_{1,1,1}^{(L-1)} \cdot (\sum_{j=1}^{2d} f_{1,j}^{(L-1)}(x) - \frac{1}{2}) + w_{1,1,0}^{(L-1)} + \frac{1}{2} \cdot w_{1,1,1}^{(L-1)} \\
&\quad + \sum_{j=1}^{2d} (w_{1,1,j}^{(L-1)} - w_{1,1,1}^{(L-1)}) \cdot f_{1,j}^{(L-1)}(x) + \sum_{j=2d+1}^{r_0} w_{1,1,j}^{(L-1)} \cdot f_{1,j}^{(L-1)}(x) \\
&\leq w_{1,1,1}^{(L-1)} \cdot (\sum_{j=1}^{2d} f_{1,j}^{(L-1)}(x) - \frac{1}{2}) + |w_{1,1,0}^{(L-1)} + \frac{1}{2} \cdot w_{1,1,1}^{(L-1)}| \\
&\quad + \sum_{j=1}^{2d} |w_{1,1,j}^{(L-1)} - w_{1,1,1}^{(L-1)}| + \sum_{j=2d+1}^{r_0} |w_{1,1,j}^{(L-1)}| \\
&\leq -4 \cdot (n+1)(\frac{3}{4} - \frac{1}{2}) + \frac{1}{2} + \sum_{j=1}^{2d} \frac{1}{2r_0} + \sum_{j=2d+1}^{r_0} \frac{1}{2r_0} \\
&= -n \leq -\log(1/e^{-n} - 1),
\end{aligned}$$

which implies (47).

In the same way we get the assertion in case $x^{(k)} > b^{(k)} + \delta$. □

**Remark 3.** It is easy to see that the number of weights of the neural network $f_{1,1}^{(L)}$ is given by

$$(L-2) \cdot (r_0^2 + r_0) + r_0 \cdot (d+2) + 1.$$

**Lemma 8** *Let $\sigma$ be the logistic squasher. Let $f_{\mathbf{w}}$ be defined by (8)–(10), let $k \in \{1, \ldots, k_n\}$ and assume that*

$$\max_{t \in \{1,\ldots,n\}} f_{k,1}^{(L)}(X_t) \cdot (1 - f_{k,1}^{(L)}(X_t)) \leq e^{-n} \tag{49}$$

*holds. Assume furthermore $\|\mathbf{w}\|_\infty \leq c_3 \cdot n^{c_4}$ and*

$$\max_{j=1,\ldots,n} \|X_j\|_\infty \leq c_3 \cdot n^{c_4} \quad \text{and} \quad \max_{j=1,\ldots,n} |Y_j| \leq c_3 \cdot n^{c_4}. \tag{50}$$

*Set*

$$F_n(\mathbf{w}) = \frac{1}{n} \sum_{i=1}^{n} |f_{\mathbf{w}}(X_i) - Y_i|^2.$$

*Then we have for all $r < L$ and all $i$, $j$*

$$\left| \frac{\partial F_n}{\partial w_{k,i,j}^{(r)}}(\mathbf{w}) \right| \leq 2 \cdot \sqrt{F_n(\mathbf{w})} \cdot r_0^L \cdot (c_3 \cdot n^{c_4})^{L+1} \cdot e^{-n}$$

**Proof.** By the Cauchy-Schwarz inequality we get

$$\begin{aligned} \left| \frac{\partial}{\partial w_{k,i,j}^{(r)}} F_n(\mathbf{w}) \right| &= \left| \frac{2}{n} \sum_{l=1}^{n} (f_{\mathbf{w}}(X_l) - Y_l) \cdot \frac{\partial f_{\mathbf{w}}}{\partial w_{k,i,j}^{(r)}}(X_l) \right| \\ &\leq 2 \cdot \sqrt{F_n(\mathbf{w})} \cdot \max_{l=1,\dots,n} \left| \frac{\partial f_{\mathbf{w}}}{\partial w_{k,i,j}^{(r)}}(X_l) \right|. \end{aligned}$$

Using the recursive definition of $f_{\mathbf{w}}$ together with (49), $r < L$ and $\sigma'(x) = \sigma(x) \cdot (1 - \sigma(x))$ we get

$$\begin{aligned} &\left| \frac{\partial f_{\mathbf{w}}}{\partial w_{k,i,j}^{(r)}}(X_l) \right| \\ &= \left| \sum_{\bar{i}=1}^{k_n} w_{1,1,\bar{i}}^{(L)} \cdot \frac{\partial f_{\bar{i},1}^{(L)}}{\partial w_{k,i,j}^{(r)}}(X_l) \right| \\ &= |w_{1,1,k}^{(L)}| \cdot \left| \frac{\partial f_{k,1}^{(L)}}{\partial w_{k,i,j}^{(r)}}(X_l) \right| \\ &= |w_{1,1,k}^{(L)}| \cdot f_{k,1}^{(L)}(X_l) \cdot (1 - f_{k,1}^{(L)}(X_l)) \cdot \left| \frac{\partial}{\partial w_{k,i,j}^{(r)}} \left( \sum_{\bar{j}=1}^{r_0} w_{k,1,\bar{j}}^{(L-1)} \cdot f_{k,\bar{j}}^{(L-1)}(X_l) + w_{k,1,0}^{(L-1)} \right) \right| \\ &\leq |w_{1,1,k}^{(L)}| \cdot e^{-n} \cdot \left| \frac{\partial}{\partial w_{k,i,j}^{(r)}} \left( \sum_{\bar{j}=1}^{r_0} w_{k,1,\bar{j}}^{(L-1)} \cdot f_{k,\bar{j}}^{(L-1)}(X_l) + w_{k,1,0}^{(L-1)} \right) \right|. \end{aligned}$$

As in the proof of Lemma 6 (cf., proof of (27)) it is possible to show

$$|w_{1,1,k}^{(L)}| \cdot \left| \frac{\partial}{\partial w_{k,i,j}^{(r)}} \left( \sum_{\bar{j}=1}^{r_0} w_{k,1,\bar{j}}^{(L-1)} \cdot f_{k,\bar{j}}^{(L-1)}(X_l) + w_{k,1,0}^{(L-1)} \right) \right| \leq r_0^L \cdot (c_3 \cdot n^{c_4})^{L+1},$$

which implies the assertion. □

**Proof of Theorem 1.** The proof is divided into six steps.

In the *first step of the proof* we show that for every $l \in \{1, \dots, n\}$ there exist (random)

$$(\bar{w}_{1,i,j}^{(r)})_{i,j,r\,:r<L} \in \left[-n^4, n^4\right]^{(L-2)\cdot(r_0^2+r_0)+r_0\cdot(d+2)+1}$$

such that for any $(w_{1,i,j}^{(r)})_{i,j,r\,:r<L}$ with

$$\max_{i,j,r:r<L} |w_{1,i,j}^{(r)} - \bar{w}_{1,i,j}^{(r)}| < \min\left\{\frac{1}{16r_0}, \frac{\log(8d-1)}{24r_0}\right\} \tag{51}$$

we have that any function $f_{1,1}^{(L)}$ corresponding to any $(\tilde{w}_{1,i,j}^{(r)})_{i,j,r\,:r<L}$ with

$$\max_{i,j,r:r<L} |\tilde{w}_{1,i,j}^{(r)} - w_{1,i,j}^{(r)}| < \min\left\{\frac{1}{16r_0}, \frac{\log(8d-1)}{24r_0}\right\} \tag{52}$$

satisfies in case $\min\{\|X_i - X_j\|_\infty : 1 \le i,j \le n, X_i \neq X_j\} \ge 1/(n+1)^3$

$$f_{1,1}^{(L)}(X_l) \ge 1 - e^{-n} \quad \text{and} \quad \max_{t\in\{1,\dots,n\},\, X_t \neq X_l} f_{1,1}^{(L)}(X_t) \le e^{-n}. \tag{53}$$

Set $\delta_n = 1/(n+1)^3$ and $a^{(i)} = X_l^{(i)} - \frac{\delta_n}{2}$ and $b^{(i)} = X_l^{(i)} + \frac{\delta_n}{2}$ $(i = 1,\dots,d)$. Then we have

$$X_l \in \left[a^{(1)} + \frac{\delta_n}{4}, b^{(1)} - \frac{\delta_n}{4}\right] \times \cdots \times \left[a^{(d)} + \frac{\delta_n}{4}, b^{(d)} - \frac{\delta_n}{4}\right],$$

and

$$\min\{\|X_i - X_j\|_\infty : 1 \le i,j \le n, X_i \neq X_j\} \ge 1/(n+1)^3$$

implies that we also have

$$X_t \notin \left[a^{(1)} - \frac{\delta_n}{4}, b^{(1)} + \frac{\delta_n}{4}\right] \times \cdots \times \left[a^{(d)} - \frac{\delta_n}{4}, b^{(d)} + \frac{\delta_n}{4}\right]$$

for all $t \in \{1,\dots,n\}$ with $X_t \neq X_l$. If $(\bar{w}_{1,i,j}^{(r)})_{i,j,r\,:r<L}$ satisfies

$$\bar{w}_{1,1,1}^{(L-1)} \le -8 \cdot (n+1),$$

$$|\bar{w}_{1,1,j}^{(L-1)} - \bar{w}_{1,1,1}^{(L-1)}| \le \frac{1}{4r_0} \quad \text{for } j = 2,\dots,d,$$

$$|\bar{w}_{1,1,j}^{(L-1)}| \le \frac{1}{4r_0} \quad \text{for } j = 2d+1,\dots,r_0,$$

$$|\bar{w}_{1,k,0}^{(L-1)} + \frac{1}{2} \cdot \bar{w}_{1,1,1}^{(L-1)}| \le \frac{1}{4} \quad \text{for } k \in \{1,\dots,2d\},$$

$$\bar{w}_{1,k,k}^{(r-1)} \ge 16 \cdot \log(8d-1) \quad \text{for } k \in \{1,\dots,2d\} \text{ and } r \in \{2,\dots,L-1\},$$

$$|\bar{w}_{1,k,0}^{(r-1)} + \frac{1}{2} \cdot \bar{w}_{1,k,k}^{(r-1)}| \le \frac{\log(8d-1)}{2r_0} \quad \text{for } k \in \{1,\dots,2d\} \text{ and } r \in \{2,\dots,L-1\},$$

$$|\bar{w}_{1,k,j}^{(r-1)}| \le \frac{\log(8d-1)}{2r_0} \quad \text{for } j \in \{1,\dots,r_0\} \setminus \{k\}, k \in \{1,\dots,2d\}, r \in \{2,\dots,L-1\},$$

$$\bar{w}_{1,k,k}^{(0)} \leq -\frac{4}{\delta_n} \cdot \log(8d-1) \quad \text{ for } k \in \{1,\dots,d\},$$

$$|\bar{w}_{1,k,0}^{(0)} + a^{(k)} \cdot \bar{w}_{1,k,k}^{(0)}| \leq \frac{\log(8d-1)}{2d} \quad \text{ for } k \in \{1,\dots,d\},$$

$$|\bar{w}_{1,k,j}^{(0)}| \leq \frac{\log(8d-1)}{2d} \quad \text{ for } k \in \{1,\dots,d\}, j \in \{1,\dots,d\} \setminus \{k\}$$

$$\bar{w}_{1,d+k,k}^{(0)} \geq \frac{4}{\delta_n} \cdot \log(8d-1) \quad \text{ for } k \in \{1,\dots,d\},$$

$$|\bar{w}_{1,d+k,0}^{(0)} + b^{(k)} \cdot \bar{w}_{1,d+k,k}^{(0)}| \leq \frac{\log(8d-1)}{2d} \quad \text{ for } k \in \{1,\dots,d\}$$

and

$$|\bar{w}_{1,d+k,j}^{(0)}| \leq \frac{\log(8d-1)}{2d} \quad \text{ for } k \in \{1,\dots,d\}, j \in \{1,\dots,d\} \setminus \{k\},$$

then it is easy to see that for any $(w_{1,i,j}^{(r)})_{i,j,r\,:r<L}$ which satisfies (51) we have that any $(\tilde{w}_{1,i,j}^{(r)})_{i,j,r\,:r<L}$ which satisfies (52) also satisfies (33)-(45). Application of Lemma 7 yields (53).

In the *second step of the proof* we show that for $n$ sufficiently large with probability at least $1 - n \cdot e^{-n}$ the weights in the random initialization of the weights are chosen such that for each $l \in \{1,\dots,n\}$ the weights for some index $k_l$ satisfy (51) (and hence all functions with weights satisfying (52) satisfy (53)). We assume in the sequel that $n$ is sufficiently large. If we sample the weight vector from the uniform distribution on

$$\left[-n^4, n^4\right]^{(L-2)\cdot(r_0^2+r_0)+r_0(d+2)+1},$$

then condition (51) is satisfied for a weight vector $\bar{w}$ corresponding to $X_1$ with probability at least

$$\left(\frac{1}{n^5}\right)^{(L-2)\cdot(r_0^2+r_0)+r_0\cdot(d+2)+1} = \frac{1}{n^{5\cdot(L-2)\cdot(r_0^2+r_0)+5\cdot r_0\cdot(d+2)+5}} =: \eta_n.$$

Hence after $\rho_n = n \cdot \lceil \frac{1}{\eta_n} \rceil$ of such independent choices (51) is never satisfied with probability less than or equal to

$$(1-\eta_n)^{\rho_n} \leq \left(1 - \frac{n}{\rho_n}\right)^{\rho_n} \leq \exp\left(-\frac{n}{\rho_n} \cdot \rho_n\right) = e^{-n}.$$

Now we consider $n$–times successively $\rho_n$ choices of the weights, i.e.,

$$k_n = n^2 \cdot \lceil \frac{1}{\eta_n} \rceil = n^{5\cdot(L-2)\cdot(r_0^2+r_0)+5\cdot r_0\cdot(d+2)+7}$$

such choices. Then the probability that in the first series of weights there are no weights corresponding to $X_1$ chosen, or in the second no weights corresponding to $X_2$, ..., or in the $n$-th no weights corresponding to $X_n$ is bounded from above by

$$\sum_{i=1}^{n} e^{-n} = n \cdot e^{-n}.$$

Set

$$C_{Lip,n} = n^{8\cdot(L-2)\cdot(r_0^2+r_0)+8\cdot r_0\cdot(d+2)+16\cdot L+15}.$$

In the *third step of the proof* we show that we have for $n$ sufficiently large

$$\|\mathbf{w}^{(t)}\|_\infty \leq 2 \cdot n^4 \quad \text{for } t = 0, 1, \ldots, t_n \tag{54}$$

and

$$\|(\nabla_{\mathbf{w}} F_n)(\mathbf{w}) - (\nabla_{\mathbf{w}} F_n)(\mathbf{w}^{(t)})\| \leq C_{Lip,n} \cdot \|\mathbf{w} - \mathbf{w}^{(t)}\| \tag{55}$$

for all $\mathbf{w} = \mathbf{w}^{(t)} + s \cdot (\mathbf{w}^{(t+1)} - \mathbf{w}^{(t)})$ and all $s \in [0,1]$, for all $t = 0, 1, \ldots, t_n - 1$.

By Lemma 6 we know that for $n$ sufficiently large (15) and (16) hold for $c_3 = 1$ and $c_4 = 4$. The initial choice of our weights implies furthermore (17) and (18) for $n$ sufficiently large. Application of Lemma 4 yields (54). And (54) together with another application of Lemma 6 implies (55).

In the *fourth step of the proof* we show for $n$ sufficiently large

$$F_n(\mathbf{w}^{(t)}) \leq n^4 \quad \text{for } t = 0, 1, \ldots, t_n. \tag{56}$$

Because of (55) we can conclude from Lemma 1 that we have for $n$ sufficiently large

$$F_n(\mathbf{w}^{(t+1)}) \leq F_n(\mathbf{w}^{(t)}) \quad \text{for } t = 0, 1, \ldots, t_n - 1.$$

But the initial choice of the weights implies

$$F_n(\mathbf{w}^{(0)}) = \frac{1}{n} \sum_{i=1}^{n} Y_i^2 \leq n^4.$$

In the *fifth step of the proof* we show that for $n$ sufficiently large and with probability at least $1 - n \cdot e^{-n}$ (14) holds for all $\mathbf{w} = \mathbf{w}^{(t)}$ $(t = 0, 1, \ldots, t_n - 1)$. Because of the first and the second step of the proof it suffices to show

$$|\bar{w}_{j_i,k,l}^{(r)} - w_{j_i,k,l}^{(r)}| \leq \frac{1}{n}$$

for all $i \in \{1, \ldots, n\}$ and all $k, l, r$ with $r < L$, where $\bar{w}_{j_i,k,l}^{(r)}$ and $w_{j_i,k,l}^{(r)}$ are the corresponding components of $\mathbf{w}^{(t)}$ and $\mathbf{w}^{(0)}$. Here $j_i$ is chosen such that

$$f_{j_i,1}^{(L)}(X_i) \geq 1 - e^{-n} \quad \text{and} \quad \max_{t \in \{1,\ldots,n\},\, X_t \neq X_i} f_{j_i,1}^{(L)}(X_t) \leq e^{-n}. \tag{57}$$

By Lemma 8 and the result of the fourth step of the proof we can successively conclude for $n$ sufficiently large that we have for $t = 0, 1, \dots, t_n - 1$

$$\left| \frac{\partial}{\partial^{(r)}_{j_i,k,l}} F_n(\mathbf{w}^{(t)}) \right| \leq 2 \cdot n^2 \cdot r_0^L \cdot (n^4)^{L+1} \cdot e^{-n} \leq \frac{1}{2n^3} = \frac{1}{t_n \cdot \lambda_n} \cdot \frac{1}{n}$$

and that consequently (57) holds for $\mathbf{w}^{(t)}$.

In the *sixth step of the proof* we show the assertion of Theorem 1. By the results of the third and the fifth step of the proof we know that the assumptions of Lemma 3 are satisfied. Application of Lemma 3 yields

$$\begin{aligned}
&F_n(\mathbf{w}^{(t_n)}) - \min_{g:\mathbb{R}^d \to \mathbb{R}} \frac{1}{n} \sum_{i=1}^n |g(X_i) - Y_i|^2 \\
&\leq \left(1 - \frac{1}{2 \cdot n \cdot C_{Lip,n}}\right)^{t_n} \cdot \left( F_n(\mathbf{w}^{(0)}) - \min_{g:\mathbb{R}^d \to \mathbb{R}} \frac{1}{n} \sum_{i=1}^n |g(X_i) - Y_i|^2 \right) \\
&\leq \exp\left( - \frac{t_n}{2 \cdot n \cdot C_{Lip,n}} \right) \cdot \left( F_n(\mathbf{w}^{(0)}) - \min_{g:\mathbb{R}^d \to \mathbb{R}} \frac{1}{n} \sum_{i=1}^n |g(X_i) - Y_i|^2 \right) \\
&= \exp(-n) \cdot \left( F_n(\mathbf{w}^{(0)}) - \min_{g:\mathbb{R}^d \to \mathbb{R}} \frac{1}{n} \sum_{i=1}^n |g(X_i) - Y_i|^2 \right).
\end{aligned}$$

With

$$F_n(\mathbf{w}^{(0)}) \leq n^4$$

we get the assertion. □

## 4.2. Proof of Theorem 2

**Lemma 9** *Let $n \in \mathbb{N}$, $(x_1, y_1)$, ..., $(x_n, y_n) \in \mathbb{R}^d \times \mathbb{R}$, $f : \mathbb{R}^d \to \mathbb{R}$, $\kappa_n > 0$ and assume*

$$\frac{1}{n} \sum_{i=1}^n |f(x_i) - y_i|^2 \leq \min_{g:\mathbb{R}^d \to \mathbb{R}} \frac{1}{n} \sum_{i=1}^n |g(x_i) - y_i|^2 + \kappa_n. \tag{58}$$

*Set*

$$\bar{m}_n(x) = \frac{\sum_{i=1}^n y_i \cdot I_{\{x_i = x\}}}{\sum_{i=1}^n I_{\{x_i = x\}}} \quad (x \in \mathbb{R}^d),$$

*where we use the convention $0/0 = 0$. Then we have for any $i \in \{1, \dots, n\}$*

$$|f(x_i) - \bar{m}_n(x_i)| \leq \sqrt{n \cdot \kappa_n}.$$

**Proof.** We have

$$\frac{1}{n} \sum_{i=1}^n |f(x_i) - y_i|^2 = \frac{1}{n} \sum_{i=1}^n |f(x_i) - \bar{m}_n(x_i)|^2 + \frac{1}{n} \sum_{i=1}^n |\bar{m}_n(x_i) - y_i|^2,$$

since

$$\frac{1}{n}\sum_{i=1}^{n}(f(x_i)-\bar{m}_n(x_i))\cdot(\bar{m}_n(x_i)-y_i)$$
$$=\frac{1}{n}\sum_{x\in\{x_1,\dots,x_n\}}(f(x)-\bar{m}_n(x))\cdot\sum_{1\leq i\leq n: x_i=x}(\bar{m}_n(x_i)-y_i)=0.$$

Application of (58) yields

$$\frac{1}{n}\sum_{i=1}^{n}|f(x_i)-\bar{m}_n(x_i)|^2\leq\kappa_n,$$

which implies the assertion. □

**Proof of Theorem 2.** Set

$$p_k=\frac{1}{n}\quad(k\in\{1,\dots,n\})$$

and set $p_k=0$ for $k>n$. Set $x_k=(k/n,0,\dots,0)^T$ and define the distribution of $(X,Y)$ by

1. $\mathbf{P}[X=x_k]=p_k$ $(k\in\mathbb{N})$,
2. $Y=m(X)+\epsilon$ where $X$, $\epsilon$ are independent and $m:\mathbb{R}^d\to\mathbb{R}$,
3. $\mathbf{P}\{\epsilon=-1\}=\frac{1}{2}=\mathbf{P}\{\epsilon=1\}$,
4. $m(x)=0$ $(x\in\mathbb{R}^d)$.

Then $m$ is the regression function of $(X,Y)$ and the distribution of $(X,Y)$ satisfies the assumptions of Theorem 2.

Set

$$\bar{m}_n(x)=\frac{\sum_{i=1}^n Y_i\cdot I_{\{X_i=x\}}}{\sum_{i=1}^n I_{\{X_i=x\}}}\quad(x\in\mathbb{R}^d).$$

Using

$$|\bar{m}_n(x)|^2\leq 2\cdot|m_n(x_k)|^2+2\cdot|m_n(x_k)-\bar{m}_n(x_k)|^2$$

together with Lemma 9 we get

$$\mathbf{E}\int|m_n(x)-m(x)|^2\mathbf{P}_X(dx)$$
$$\geq\mathbf{E}\left\{\sum_{k=1}^n|m_n(x_k)|^2\cdot p_k\cdot I_{\{\sum_{i=1}^n I_{\{X_i=x_k\}}>0\}}\cdot I_{\{U\in\mathcal{P}_n\}}\right\}$$
$$\geq\mathbf{E}\left\{\sum_{k=1}^n\left(\frac{1}{2}|\bar{m}_n(x_k)|^2-|m_n(x_k)-\bar{m}_n(x_k)|^2\right)\cdot p_k\cdot I_{\{\sum_{i=1}^n I_{\{X_i=x_k\}}>0\}}\cdot I_{\{U\in\mathcal{P}_n\}}\right\}$$
$$=\mathbf{E}\left\{\sum_{k=1}^n\frac{1}{2}|\bar{m}_n(x_k)|^2\cdot p_k\cdot I_{\{\sum_{i=1}^n I_{\{X_i=x_k\}}>0\}}\right\}$$

$$
\begin{aligned}
&-\mathbf{E}\left\{\sum_{k=1}^{n}\frac{1}{2}|\bar{m}_n(x_k)|^2\cdot p_k\cdot I_{\{\sum_{i=1}^n I_{\{X_i=x_k\}}>0\}}\cdot I_{\{U\in\mathcal{P}_n^c\}}\right\}\\
&-\mathbf{E}\left\{\sum_{k=1}^{n}|m_n(x_k)-\bar{m}_n(x_k)|^2\cdot p_k\cdot I_{\{\sum_{i=1}^n I_{\{X_i=x_k\}}>0\}}\cdot I_{\{U\in\mathcal{P}_n\}}\right\}\\
\geq\ &\frac{1}{2}\cdot\sum_{k=1}^{n}\mathbf{E}\left\{|\bar{m}_n(x_k)|^2\cdot I_{\{\sum_{i=1}^n I_{\{X_i=x_k\}}>0\}}\right\}\cdot p_k-\frac{1}{2}\cdot\mathbf{P}_U(\mathcal{P}_n^c)-n\cdot\kappa_n,
\end{aligned}
$$

where in the last inequality we used $|\tilde{m}_n(x)|\leq 1$ which holds because of $Y_i\in\{-1,1\}$. The definition of $\bar{m}_n$ implies

$$
\begin{aligned}
&\sum_{k=1}^{n}\mathbf{E}\left\{|\bar{m}_n(x_k)|^2\cdot I_{\{\sum_{i=1}^n I_{\{X_i=x_k\}}>0\}}\right\}\cdot p_k\\
&\geq\sum_{k=1}^{n}\mathbf{E}\left\{\mathbf{E}\left\{|\bar{m}_n(x_k)|^2\big|X_1,\dots,X_n\right\}\cdot I_{\{\sum_{i=1}^n I_{\{X_i=x_k\}}>0\}}\right\}\cdot p_k\\
&=\sum_{k=1}^{n}\mathbf{E}\left\{\frac{1}{\sum_{i=1}^n I_{\{X_i=x_k\}}}\cdot I_{\{\sum_{i=1}^n I_{\{X_i=x_k\}}>0\}}\right\}\cdot p_k.
\end{aligned}
$$

Using the fact that $\sum_{i=1}^n I_{\{X_i=x_k\}}$ is binomially distributed with $n$ degrees of freedom and probability of success $p_k$ we get

$$
\begin{aligned}
&\sum_{k=1}^{n}\mathbf{E}\left\{\frac{1}{\sum_{i=1}^n I_{\{X_i=x_k\}}}\cdot I_{\{\sum_{i=1}^n I_{\{X_i=x_k\}}>0\}}\right\}\cdot p_k\\
&=\sum_{k=1}^{n}\sum_{i=1}^{n}\frac{1}{i}\cdot\binom{n}{i}p_k^i\cdot(1-p_k)^{n-i}\cdot p_k\\
&\geq\sum_{k=1}^{n}\sum_{i=1}^{n}\frac{1}{i+1}\cdot\binom{n}{i}p_k^i\cdot(1-p_k)^{n-i}\cdot p_k\\
&=\frac{1}{n+1}\cdot\sum_{k=1}^{n}\sum_{i=1}^{n}\binom{n+1}{i+1}p_k^{i+1}\cdot(1-p_k)^{n+1-(i+1)}\\
&=\frac{n}{n+1}\cdot\left(1-\left(1-\frac{1}{n}\right)^{n+1}-(n+1)\cdot\frac{1}{n}\cdot\left(1-\frac{1}{n}\right)^{n}\right)\\
&\geq\frac{n}{n+1}\cdot\left(1-\frac{2n+1}{n}\cdot\left(1-\frac{1}{n}\right)^{n}\right)\\
&\geq\frac{10}{11}\cdot\left(1-\frac{21}{10}\cdot\frac{1}{e}\right),
\end{aligned}
$$

where the last inequality holds for $n\geq 10$.

Putting together the above results implies the assertion. □

## Acknowledgments

The authors would like to thank the Associate Editor and three anonymous referees for various invaluable comment, which helped very much to improve the presentation. This work was supported in part by the Natural Sciences and Engineering Research Council of Canada under Grant RGPIN-2015-06412 and the second author would like to thank NSERC for funding this work.

## References

[1] Allen-Zhu, Z., Li, Y., and Liang, Y. (2019). Learning and generalization in overparameterized neural networks, going beyond two layers. In *Advances in neural information processing systems*, pages 6155-6166.

[2] Allen-Zhu, Z., Li, Y., and Song, Z. (2019). A convergence theory for deep kearning via over-parameterization. *Proceedings of the 36th International Conference on Machine Learning (PMLR 2019)*, **97**, pp. 242-252. Long Beach, California.

[3] Arora, S., Cohen, N., Golowich, N., and Hu, W. (2018). A convergence analysis of gradient descent for deep linear neural networks. *International Conference on Learning Representations (ICLR 2019).* New Orleans, Louisiana.

[4] Bartlett, P. L., Long, P. M., and Lugosi, G. (2019). Beningn overfitting in linear regression. arXiv: 1906.11300v1.

[5] Bauer, B., and Kohler, M. (2019). On deep learning as a remedy for the curse of dimensionality in nonparametric regression. *Annals of Statistics* **47**, pp. 2261-2285.

[6] Belkin, M., Rakhlin, A., and Tsybakov, A. B. (2018). Does data interpolation contradict statistical optimality? arXiv: 1806.09471v1.

[7] Belkin, M., Hsu, D., Ma, S., and Mandal, S. (2019). Reconciling modern machine learning practice and the bias-variance trade-off. arXiv: 1812.11118v2.

[8] Bubeck, S., Eldan, R., Lee, Y. T., and Mikulincer, D. (2020) Network size and weights size for memorization with two-layers neural networks. arXiv: 2006.02855.

[9] Braun, A., Kohler, M., and Walk, H. (2019). On the rate of convergence of a neural network regression estimate learned by gradient descent. Submitted for publication.

[10] Cao, Y., and Gu, Q. (2019). Generalization error bounds of gradient descent for learning overparameterized deep relu networks. arXiv: 1902.01384.

[11] Chizat, L., Oyallon, E., and Bach, F. (2020). On Lazy Training in Differentiable Programming. arXiv: 1812.07956.

[12] Choromanska, A., Henaff, M., Mathieu, M., Arous, G. B., and LeCun, Y. (2015) The loss surface of multilayer networks. International Conference on Articial Intelligence and Statistics (AISTATS) 2015, San Diego, CA, USA. Proceeding of Machine Learning Research, volume 38, pp. 192-204.

[13] Daniely, A. (2019) Neural networks learning and memorization with (almost) no over-parameterization. arXiv: 1911.09873.

[14] Daniely, A. (2020) Memorizing gaussians with no over-parameterizaion via gradient decent on neural networks. arXiv: 2003.12895.

[15] Devroye, L., Györfi, L., and Lugosi, G. (1996). *A Probabilistic Theory of Pattern Recognition*. Springed, New York, USA.

[16] Devroye, L., and Wagner, T. J. (1980). Distribution-free consistency results in nonparametric discrimination and regression function estimation. *Annals of Statistics*, **8**, pp. 231-239.

[17] Du, S., and Lee, J. (2018). On the power of over-parametrization in neural networks with quadratic activation. *Proceedings of the 35th International Conference on Machine Learning (PMLR 2018)*, **80**, pp. 1329-1338. Stockholm, Sweden.

[18] Du, S., Lee, J., Tian, Y., Poczos, B., and Singh, A. (2018). Gradient descent learns one-hidden-layer CNN: don't be afraid of spurious local minima. *Proceedings of the 35th International Conference on Machine Learning (PMLR 2018)*, **80**, pp. 1339-1348. Stockholm, Sweden.

[19] Eckle, K., and Schmidt-Hieber, J. (2019). A comparison of deep networks with ReLU activation function and linear spline-type methods. *Neural Networks*, **110**, pp. 232-242.

[20] Györfi, L., Kohler, M., Krzyżak, A., and Walk, H. (2002). *A Distribution–Free Theory of Nonparametric Regression*. Springer.

[21] Hastie, T., Montanari, A., Rosset, S., and Tibshirani, R. J. (2019). Surprises in high-dimensional ridgeless least squares interpolation. *arXiv: 1903.08560v4*

[22] Huang, J. and Yau H-T. (2019). Dynamics of deep neural networks and neural tangent hierarchy. *arXiv:1909.08156v1*

[23] Imaizumi, M., and Fukamizu, K. (2019). Deep neural networks learn non-smooth functions effectively. *Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics (AISTATS 2019)*. Naha, Okinawa, Japan.

[24] Jackot, A., Gabriel, F. and Hongler, C. (2018). Neural tangent kernel: convergence and generalization in neural networks. *Advances in Neural Information Processing Systems*, pp. 8571-8580.

[25] Kawaguchi, K. (2016). Deep learning without poor local minima. *30th Conference on Neural Information Processing Systems (NIPS 2016)*. Barcelona, Spain.

[26] Kawaguchi, K, and Huang, J. (2019). Gradient descent finds global minima for generalizable deep neural networks of practical sizes. *57th IEEE Annual Allerton Conference on Communication, Control, and Computing*, Allerton, IL, pp. 92-99.

[27] Kohler, M., and Krzyżak, A. (2017). Nonparametric regression based on hierarchical interaction models. *IEEE Transaction on Information Theory*, **63**, pp. 1620-1630.

[28] Kohler, M., Krzyżak, A., and Langer, S. (2019). Estimation of a function of low local dimensionality by deep neural networks. Submitted for publication. arXiv: 1908.11140.

[29] Kohler, M., and Langer, S. (2019). On the rate of convergence of fully connected deep neural network regression estimates. To appear in *Annals of Statistics*, 2021. arXiv: 1908.11133.

[30] Liang, S., Sun, R., Lee, J., and Srikant, R. (2018). Adding one neuron can eliminate all bad local minima. *Proceedings of the 32nd Conference on Neural Information Processing Systems (NIPS 2018)*, pp. 4355 - 4365. Montreal, Canada.

[31] Montanari, A. and Zhong, Y. (2020) The interpolation phase transition in neural networks: Mem- orization and generalization under lazy training. arXiv: 2007.12826.

[32] Mücke, N., and Steinwart, I. (2019). Global Minima of DNNs: The Plenty Pantry.

arXiv: 1905.10686.

[33] Neyshabur, B., Li, Z., Bhojanapalli, S., LeCun, Y., and Srebro, N. (2019). The role of over-parametrization in generalization of neural networks. In *7th International Conference on Learning Representations, ICLR 2019.*

[34] Nitanda, A. and Suzuki, T. (2020). Optimal rates for averaged stochastic gradient descent under neural tangent kernel regime. *arXiv:2006.12297v1.*

[35] Schmidhuber, J. (2015). Deep learning in neural networks: an overview. *Neural Networks*, **61**, pp. 85-117.

[36] Schmidt-Hieber, J. (2020a). Nonparametric regression using deep neural networks with ReLU activation function. *Annals of Statistics*, **48**, pp. 1875-1897.

[37] Schmidt-Hieber, J. (2020b). Rejoinder to discussions of "Nonparametric regression using deep neural networks with ReLU activation function". *Annals of Statistics*, **48**, pp. 1916-1921.

[38] Stone, C. J. (1982). Optimal global rates of convergence for nonparametric regression. *Annals of Statistics*, **10**, pp. 1040-1053.

[39] Stone, C. J. (1985). Additive regression and other nonparametric models. *Annals of Statistics*, **13**, pp. 689-705.

[40] Stone, C. J. (1994). The use of polynomial splines and their tensor products in multivariate function estimation. *Annals of Statistics*, **22**, pp. 118-184.

[41] Woodworth, B., Gunasekar, S., Lee, J., Moroshko, E., Savarese, P., Golan, I., Soudry, D., and Srebro, N. (2020). Kernel and rich regimes in overparametrized models. arXiv: 2002.09277