# On the rate of convergence of an over-parametrized Transformer classifier learned by gradient descent *


Michael Kohler[1] and Adam Krzyżak[2,†]

[1] *Fachbereich Mathematik, Technische Universität Darmstadt, Schlossgartenstr. 7, 64289 Darmstadt, Germany, email: kohler@mathematik.tu-darmstadt.de*

[2] *Department of Computer Science and Software Engineering, Concordia University, 1455 De Maisonneuve Blvd. West, Montreal, Quebec, Canada H3G 1M8, email: krzyzak@cs.concordia.ca*


December 3, 2023


**Abstract**

Classification from independent and identically distributed random variables is considered. Classifiers based on over-parametrized transformer encoders are defined where all the weights are learned by gradient descent. Under suitable conditions on the a posteriori probability an upper bound on the rate of convergence of the difference of the misclassification probability of the estimate and the optimal misclassification probability is derived.



*AMS classification:* Primary 62G08; secondary 62G20.

*Key words and phrases:* Gradient descent, over-parametrization, pattern recognition, rate of convergence, Transformer networks.


## 1 Introduction

### 1.1 Scope of this article

One of the most recent and fascinating breakthroughs in artificial intelligence is ChatGPT, a chatbot which can simulate human conversation. ChatGPT is an instance of GPT4, which is a language model based on **g**enerative **p**redictive **t**ransformers. So if one wants to study from a theoretical point of view, how powerful such artificial intelligence can be, one approach is to consider transformer networks and to study which problems one can solve with these networks theoretically. Here it is not only important what kind of models these network can *approximate*, or how they can *generalize* their knowledge learned by choosing the best possible approximation to a concrete data set, but also how well *optimization* of such transformer network based on concrete data set works. In this article we consider all these three different aspects simultaneously and show a theoretical

*Running title: *Over-parametrized transformers*


†Corresponding author. Tel: +1-514-848-2424 ext. 3007, Fax:+1-514-848-2830

upper bound on the missclassification probability of a transformer network fitted to the observed data. For simplicity we focus in this context on transformer encoder networks which can be applied to define an estimate in the context of a classification problem involving natural language.

### 1.2 Pattern recognition

We study these estimates in the context of pattern recognition. Given $(X, Y)$, $(X_1, Y_1)$, $\dots$, $(X_n, Y_n)$ independent and identically distributed random variables with values in $\mathbb{R}^{d \cdot l} \times \{-1, 1\}$, and given the data set

$$\mathcal{D}_n = \{(X_1, Y_1), \dots, (X_n, Y_n)\}$$

the goal is to construct a classifier

$$\eta_n(\cdot) = \eta_n(\cdot, \mathcal{D}_n) : \mathbb{R}^{d \cdot l} \to \{-1, 1\}$$

such that its misclassification probability

$$\mathbf{P}\{\eta_n(X) \neq Y | \mathcal{D}_n\}$$

is as small as possible. Here the predictor variable $X$ describes the encoding of a sequence of length $l$ consisting of words or tokens, and each word or token is encoded by a value in $\mathbb{R}^d$. The goal is to predict the label $Y$ corresponding to the sentence described by $X$.

Let

$$m(x) = \mathbf{P}\{Y = 1 | X = x\} \quad (x \in \mathbb{R}^{d \cdot l}) \tag{1}$$

be the a posteriori probability of class 1. Then

$$\eta^*(x) = \begin{cases} 1, & \text{if } m(x) \geq \frac{1}{2} \\ -1, & \text{elsewhere} \end{cases}$$

is the Bayes classifier, i.e., the classifier satisfying

$$\mathbf{P}\{\eta^*(X) \neq Y\} = \min_{\eta : \mathbb{R}^{d \cdot l} \to \{-1,1\}} \mathbf{P}\{\eta(X) \neq Y\}$$

(cf., e.g., Theorem 2.1 in Devroye, Györfi and Lugosi (1996)).

In this paper we derive upper bounds on

$$\begin{aligned} &\mathbf{E}\left\{\mathbf{P}\{\eta_n(X) \neq Y | \mathcal{D}_n\} - \mathbf{P}\{\eta^*(X) \neq Y\}\right\} \\ &= \mathbf{P}\{\eta_n(X) \neq Y\} - \min_{\eta : \mathbb{R}^{d \cdot l} \to \{-1,1\}} \mathbf{P}\{\eta(X) \neq Y\}. \end{aligned} \tag{2}$$

It is well-known that in order to derive nontrivial rate of convergence results on the difference between the misclassification probability of any estimate and the minimal possible value it is necessary to restrict the class of distributions (cf., e.g., Section 3.1 in Györfi et al. (2002)). In this context we will assume that the a posteriori probability is smooth in the following sense:

**Definition 1** *Let $p = q + s$ for some $q \in \mathbb{N}_0$ and $0 < s \leq 1$. A function $m : \mathbb{R}^{d \cdot l} \to \mathbb{R}$ is called $(p, C)$-smooth, if for every $\alpha = (\alpha_1, \dots, \alpha_{d \cdot l}) \in \mathbb{N}_0^{d \cdot l}$ with $\sum_{j=1}^{d \cdot l} \alpha_j = q$ the partial derivative $\partial^q m / (\partial x_1^{\alpha_1} \dots \partial x_{d \cdot l}^{\alpha_{d \cdot l}})$ exists and satisfies*

$$\left| \frac{\partial^q m}{\partial x_1^{\alpha_1} \dots \partial x_{d \cdot l}^{\alpha_{d \cdot l}}}(x) - \frac{\partial^q m}{\partial x_1^{\alpha_1} \dots \partial x_{d \cdot l}^{\alpha_{d \cdot l}}}(z) \right| \leq C \|\mathbf{x} - \mathbf{z}\|^s$$

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

In order to show good rates of convergence even for high-dimensional predictors we use a hierarchical composition model as in Schmidt-Hieber (2020), where the a posteriori probability is represented by a composition of several functions and where each of these functions depends only on a few variables. We use the following definition of Kohler and Langer (2021) to formalize this assumption.

**Definition 2** *Let $d, l \in \mathbb{N}$, $m : \mathbb{R}^{d \cdot l} \to \mathbb{R}$ and let $\mathcal{P}$ be a subset of $(0, \infty) \times \mathbb{N}$.*
***a)*** *We say that $m$ satisfies a hierarchical composition model of level $0$ with order and smoothness constraint $\mathcal{P}$, if there exists $K \in \{1, \dots, d \cdot l\}$ such that*

$$m(\mathbf{x}) = x^{(K)} \quad \textit{for all } \mathbf{x} = (x^{(1)}, \dots, x^{(d \cdot l)})^\top \in \mathbb{R}^{d \cdot l}.$$

***b)*** *Let $\kappa \in \mathbb{N}_0$. We say that $m$ satisfies a hierarchical composition model of level $\kappa + 1$ with order and smoothness constraint $\mathcal{P}$, if there exist $(p, K) \in \mathcal{P}$, $C > 0$, $g : \mathbb{R}^K \to \mathbb{R}$ and $f_1, \dots, f_K : \mathbb{R}^{d \cdot l} \to \mathbb{R}$, such that $g$ is $(p, C)$–smooth, $f_1, \dots, f_K$ satisfy a hierarchical composition model of level $\kappa$ with order and smoothness constraint $\mathcal{P}$ and*

$$m(\mathbf{x}) = g(f_1(\mathbf{x}), \dots, f_K(\mathbf{x})) \quad \textit{for all } \mathbf{x} \in \mathbb{R}^{d \cdot l}.$$

Let $\mathcal{H}(\kappa, \mathcal{P})$ be the set of all functions $m : \mathbb{R}^{d \cdot l} \to \mathbb{R}$ which satisfy a hierarchical composition model of level $\kappa$ with order and smoothness constraint $\mathcal{P}$.

A motivation of hierarchical models from an applied point of view can be found in Kohler and Langer (2020a).

### 1.3 Learning of a transformer encoder

We apply gradient descent to an over-parametrized model of a transformer encoder in order to learn its parameter. More precisely, let $\mathbf{\Theta}$ be the set of parameters of the transformer networks $\{f_\vartheta : \vartheta \in \mathbf{\Theta}\}$ (which we will introduce in detail in Section 2 below), and consider a linear combination

$$f(x) = f_{(w_k)_{k=1,\dots,K}, (\vartheta_k)_{k=1,\dots,K}}(x) = \sum_{k=1}^K w_k \cdot f_{\vartheta_k}(x)$$

of transformer networks $f_{\vartheta_k}$. Here $(w_k)_{k=1,\dots,K}$ are weights satisfying

$$w_k \geq 0 \quad \text{and} \sum_{k=1}^K w_k = 1, \tag{3}$$

$\vartheta_k$ are the weights of the transformer networks $f_{\vartheta_k}$ $(k = 1, \dots, K_n)$, and by choosing $K$ very large our model becomes over-parametrized in the sense that the number of its parameters is much larger than the sample size. We will use

$$\eta_n(X) = sgn(f(X))$$

as our prediction of $Y$, and in order to achieve a small missclassification probability our aim will be to choose the parameters $(w_k)_{k=1,\dots,K}$ and $(\vartheta_k)_{k=1,\dots,K}$ of $f$ such that its logistic loss

$$\mathbf{E}\left\{\log\left(1 + \exp(-Y \cdot f(X)\right)\right\}$$

is small. To do this, we will randomly initialize its parameter in a proper way and then perform $t_n$ gradient descent steps in view of minimization of the empirical logistic loss

$$\frac{1}{n}\sum_{i=1}^{n} \log\left(1 + \exp(-Y_i \cdot f(X_i)\right),$$

where proper projection steps will ensure that (3) is satisfied and that the parameters $\vartheta_k$ will not move away too far from their random starting values (see Section 2 for details).

### 1.4 Main results

We show, that in case that the a posteriori probability satisfies a hierarchical composition model with smoothness and order constraint $\mathcal{P}$, the corresponding estimate $\eta_n$ satisfies

$$\mathbf{P}\{\eta_n(X) \neq Y\} - \min_{\eta:\mathbb{R}^{d \cdot l} \to \{-1,1\}} \mathbf{P}\{\eta(X) \neq Y\} \leq c_1 \cdot (\log n)^3 \cdot \max_{(p,K)\in\mathcal{P}} n^{-\min\left\{\frac{p}{2\cdot(2p+K)}, \frac{1}{6}\right\}}.$$

And if, in addition,

$$\mathbf{P}\left\{\max\left\{\frac{\mathbf{P}\{Y=1|X\}}{1-\mathbf{P}\{Y=1|X\}}, \frac{1-\mathbf{P}\{Y=1|X\}}{\mathbf{P}\{Y=1|X\}}\right\} > n^{1/3}\right\} \geq 1 - \frac{1}{n^{1/3}} \quad (n \in \mathbb{N})$$

holds (which implies that with high probability $\mathbf{P}\{Y = 1|X\}$ is either close to one or close to zero) then we show that the estimates achieve the improved rate of convergence

$$\mathbf{P}\{\eta_n(X) \neq Y\} - \min_{\eta:\mathbb{R}^{d \cdot l} \to \{-1,1\}} \mathbf{P}\{\eta(X) \neq Y\} \leq c_2 \cdot (\log n)^3 \cdot \max_{(p,K)\in\mathcal{P}} n^{-\min\left\{\frac{p}{2p+K}, \frac{1}{3}\right\}}.$$

In order to prove these results we derive a general result which gives an upper bound on the expected logistic loss of an over-parametrized linear combination of deep networks learned by minimizing an empirical logistic loss via gradient descent. In the proof of this result we show that the projection of the outer weights enables us to bound the generalization error of our over-parametrized linear combination of deep networks by the Rademacher complexity of a class of single deep networks. And in the application of this general result, we derive new approximation properties of Transformer networks with slightly disturbed weight matrices.

## 1.5 Discussion of related results

Transformers have been introduced by Vaswani et al. (2017). In applications they are usually combined with unsupervised pre-training and the same pre-trained transformer encoder is then fine-tuned to a variety of natural language processing tasks, see Devlin et al. (2019).

Approximation and generalization of Transformer encoder networks has been studied in Gurevych et al. (2022). There estimates are defined as plug-in classifiers of the least squares estimates based on transformer networks, and similar rate of convergence results as in the current paper are shown. The main difference between our result in this paper and the result in Gurevych et al. (2022) is that we define our estimates using gradient descent, and consequently we have to take the optimization error into account too, which forces us to derive technically much more complex approximation results for transformer networks.

Much more is known about the deep neural network estimates. There exist quite a few approximation results for neural networks (cf., e.g., Yarotsky (2018), Yarotsky and Zhevnerchute (2019), Lu et al. (2020), Langer (2021) and the literature cited therein), and generalization of deep neural networks can either be analyzed within the framework of the classical VC theory (using e.g. the result of Bartlett et al. (2019) to bound the VC dimension of classes of neural networks) or in case of over-parametrized deep neural networks (where the number of free parameters adjusted to the observed data set is much larger than the sample size) by using bounds on the Rademacher complexity (cf., e.g., Liang, Rakhlin and Sridharan (2015), Golowich, Rakhlin and Shamir (2019), Lin and Zhang (2019), Wang and Ma (2022) and the literature cited therein).

Combining such results leads to a rich theory showing that owing to the network structure the least squares neural network estimates can achieve suitable dimension reduction in hierarchical composition models for the function to be estimated. For a simple model this was first shown by Kohler and Krzyżak (2017) for Hölder smooth function and later extended to arbitrary smooth functions by Bauer and Kohler (2019). For a more complex hierarchical composition model and the ReLU activation function this was shown in Schmidt-Hieber (2020) under the assumption that the networks satisfy some sparsity constraint. Kohler and Langer (2021) showed that this also possible for fully connected neural networks, i.e., without imposing a sparsity constraint on the network. Adaptation of deep neural network to especially weak smoothness assumptions was shown in Imaizumi and Fukamizu (2018), Suzuki (2018) and Suzuki and Nitanda (2019).

Less well understood is the optimization of deep neural networks. As was shown, e.g., in Zou et al. (2018), Du et al. (2019), Allen-Zhu, Li and Song (2019) and Kawaguchi and Huang (2019) application of gradient descent to over-parameterized deep neural networks leads to neural network which (globally) minimizes the empirical risk considered. However, as was shown in Kohler and Krzyżak (2021), the corresponding estimates do not behave well on new independent data. So the main question is why gradient descent (and its variants like stochastic gradient descent) can be used to fit a neural network to observed data in such a way that the resulting estimate achieves good results on new independent data. The challenge here is not only to analyze optimization but to consider it simultaneously with approximation and generalization.

In case of shallow neural networks (i.e., neural networks with only one hidden layer) this has been done successfully in Braun et al. (2023). Here it was possible to show that the classical dimension free rate of convergence of Barron (1994) for estimation of a regression function where its Fourier transform has a finite moment can also be achieved by shallow neural networks learned by gradient descent. The main idea here is that the gradient descent selects a subset of the neural network where random initialization of the inner weights leads to values with good approximation properties, and that it adjusts the outer weights for these neurons properly. A similar idea was also applied in Gonon (2021). Kohler and Krzyżak (2022) applied this idea in the context of over-parametrized deep neural networks where a linear combination of a huge number of deep neural networks of fixed size are computed in parallel. Here the gradient descent selects again a subset of the neural networks computed in parallel and chooses a proper linear combination of the networks. By using metric entropy bounds (cf., e.g., Birman and Solomnjak (1967) and Li, Gu and Ding (2021)) it is possible to control generalization of the over-parametrized neural networks, and as a result the rate of convergence of order close to $n^{-1/(1+d)}$ (or $n^{1/(1+d^*)}$ in case of interaction models, where it is assumed that the regression function is a sum of functions applied to only $d^*$ of the $d$ components of the predictor variable) can be shown for Hölder-smooth regression function with Hölder exponent $p \in [1/2, 1]$. Universal consistency of such estimates for bounded $X$ was shown in Drews and Kohler (2022).

In all those results adjusting the inner weights with gradient descent is not important. In fact, Gonon (2021) does not do this at all, while Braun et al. (2023) and Kohler and Krzyżak (2022) use that the relevant inner weights do not move too far away from their starting values during gradient descent. Similar ideas have also been applied in Andoni et al. (2014) and Daniely (2017). This whole approach is related to random feature networks (cf., e.g., Huang, Chen and Siew (2006) and Rahimi and Recht (2008a, 2008b, 2009)), where the inner weights are chosen randomly and only the outer weights are learned during gradient descent. Yehudai and Shamir (2022) present a lower bound which implies that either the number of neurons or the absolute value of the coefficients must grow exponential in the dimension in order to learn a single ReLU neuron with random feature networks. But since Braun et al. (2023) was able to prove a useful rate of convergence result for networks similar to random feature networks, the practical relevance of this lower bound is not clear.

The estimates in Kohler and Krzyżak (2022) use a $L_2$ regularization on the outer weights during gradient descent. As was shown in Drews and Kohler (2023), it is possible to achieve similar results without $L_2$ regularization.

Often gradient descent in neural networks is studied in the neural tangent kernel setting proposed by Jacot, Gabriel and Hongler (2020), where instead of a neural network estimate a kernel estimate is studied and its error is used to bound the error of the neural network estimate. For further results in this context see Hanin and Nica (2019) and the literature cited therein. Nitanda and Suzuki (2021) were able to analyze the global error of an over-parametrized shallow neural network learned by gradient descent based on this approach. However, due to the use of the neural tangent kernel, also the smoothness assumption of the function to be estimated has to be defined with the aid of a norm involving the kernel, which does not lead to the classical smoothness conditions of our paper. Another approach where the estimate is studied in some asymptotically equivalent model is the mean field approach, cf., Mei, Montanari, and Nguyen (2018), Chizat and Bach (2018) or Nguyen and Pham (2020). A survey of various results on over-parametrized deep neural network estimates learned by gradient descent can be found in Bartlett, Montanari and Rakhlin (2021).

### 1.6 Notation

The sets of natural numbers, natural numbers including zero, real numbers and non-negative real numbers are denoted by $\mathbb{N}$, $\mathbb{N}_0$, $\mathbb{R}$ and $\mathbb{R}_+$, respectively. We set $\bar{\mathbb{R}} = \mathbb{R} \cup \{-\infty, \infty\}$. For $z \in \mathbb{R}$, we denote the smallest integer greater than or equal to $z$ by $\lceil z \rceil$, and we set $z_+ = \max\{z, 0\}$ and $z_- = \max\{-z, 0\}$. The Euclidean norm of $x \in \mathbb{R}^d$ is denoted by $\|x\|$ and for $x, z \in \mathbb{R}^d$ its scalar product is denoted by $< x, z >$. For a closed and convex set $A \subseteq \mathbb{R}^d$ we denote by $Proj_A x$ that element $Proj_A x \in A$ such that

$$\|x - Proj_A x\| = \min_{z \in A} \|x - z\|.$$

For $f : \mathbb{R}^d \to \mathbb{R}$

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

is its supremum norm, and for $A \subseteq \mathbb{R}^d$ we set

$$\|f\|_{\infty, A} = \sup_{x \in A} |f(x)|.$$

For a vector $x = (x^{(1)}, \dots, x^{(d)})^T$ we denote by

$$\|x\|_\infty = \max_{i=1,\dots,n} |x^{(i)}|$$

its supremum norm, and if $A = (a_{i,j})_{i=1,\dots,I, j=1,\dots,J}$ we set

$$\|A\|_\infty = \max_{i=1,\dots,I, j=1,\dots,J} |a_{i,j}|.$$

For $\mathbf{j} = (j^{(1)}, \dots, j^{(d)}) \in \mathbb{N}_0^d$ we write

$$\|\mathbf{j}\|_1 = j^{(1)} + \dots + j^{(d)}$$

and for $f : \mathbb{R}^d \to \mathbb{R}$ we set

$$\partial^{\mathbf{j}} f = \frac{\partial^{\|\mathbf{j}\|_1} f}{(\partial x^{(1)})^{j^{(1)}} \dots (\partial x^{(d)})^{j^{(d)}}}.$$

For $q \in \mathbb{N}_0$ and $f : \mathbb{R}^d \to \mathbb{R}$ we set

$$\|f\|_{C^q(\mathbb{R}^d)} = \max\left\{ \|\partial^{\mathbf{j}} f\|_\infty \, : \, \mathbf{j} \in \mathbb{N}_0^d, \, \|\mathbf{j}\|_1 \le q \right\}.$$

Let $\mathcal{F}$ be a set of functions $f : \mathbb{R}^d \to \mathbb{R}$, let $x_1, \dots, x_n \in \mathbb{R}^d$, set $x_1^n = (x_1, \dots, x_n)$ and let $p \geq 1$. A finite collection $f_1, \dots, f_N : \mathbb{R}^d \to \mathbb{R}$ is called an $L_p$ $\varepsilon$–packing in $\mathcal{F}$ on $x_1^n$ if $f_1, \dots, f_N \in \mathcal{F}$ and

$$\min_{1 \leq i < j \leq N} \left( \frac{1}{n} \sum_{k=1}^{n} |f_i(x_k) - f_j(x_k)|^p \right)^{1/p} \geq \varepsilon$$

holds. The $L_p$ $\varepsilon$–packing number of $\mathcal{F}$ on $x_1^n$ is the size $N$ of the largest $L_p$ $\varepsilon$–packing of $\mathcal{F}$ on $x_1^n$ and is denoted by $\mathcal{M}_p(\varepsilon, \mathcal{F}, x_1^n)$.

For $z \in \mathbb{R}$ and $\beta > 0$ we define $T_\beta z = \max\{-\beta, \min\{\beta, z\}\}$. If $f : \mathbb{R}^d \to \mathbb{R}$ is a function then we set $(T_\beta f)(x) = T_\beta\,(f(x))$. For $z \in \bar{\mathbb{R}}$ we denote by

$$sgn(z) = \begin{cases} 1 & \text{if } z > 0, \\ 0 & \text{if } z = 0, \\ -1 & \text{if } z < 0 \end{cases}$$

its sign. For $i, j \in \mathbb{N}_0$ we set

$$\delta_{i,j} = \begin{cases} 1 & \text{if } i = j, \\ 0 & \text{if } i \neq j. \end{cases}$$

### 1.7 Outline

The over-parametrized transformer classifiers considered in this paper are introduced in Section 2. The main result is presented in Section 3. In Section 4 we present a general result concerning the expected logistic loss of an over-parametrized estimate defined by a linear combination of deep networks. The proof of our main result is given in Section 5.

## 2 Definition of the estimate

### 2.1 Topology of the Transformer networks

Let $K_n \in \mathbb{N}$ be the number of transformer networks which we compute in parallel. The over-parametrized transformer networks which we use for our classifier are of the form

$$f_{(w_k)_{k=1,\dots,K_n}, (\mathbf{W}_k)_{k=1,\dots,K_n}, (\mathbf{V}_k)_{k=1,\dots,K_n}}(x) = \sum_{k=1}^{K_n} w_k \cdot T_{\beta_n}(f_{\mathbf{W}_k, \mathbf{V}_k}(x)), \tag{4}$$

where the outer weights $(w_k)_{k=1,\dots,K_n}$ will be chosen such that

$$w_k \geq 0 \quad (k = 1, \dots, K_n) \quad \text{and} \quad \sum_{k=1}^{K_n} w_k \leq 1 \tag{5}$$

hold and where $\mathbf{W}_k$ and $\mathbf{V}_k$ are the weights used in the $k$-th Transformer network $T_{\beta_n}(f_{\mathbf{W}_k, \mathbf{V}_k})$ and $\beta_n = c_3 \cdot \log n$. This Transformer network is defined as follows:

$$x = \begin{pmatrix} x_1^{(1)} & x_2^{(1)} & x_3^{(1)} & x_4^{(1)} \\ x_1^{(2)} & x_2^{(2)} & x_3^{(2)} & x_4^{(2)} \end{pmatrix} \quad \mapsto \quad z_{k,0} = \begin{pmatrix} x_1^{(1)} & x_2^{(1)} & x_3^{(1)} & x_4^{(1)} \\ x_1^{(2)} & x_2^{(2)} & x_3^{(2)} & x_4^{(2)} \\ 1 & 1 & 1 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 \end{pmatrix} \in \mathbb{R}^{20 \times 4}$$

Figure 1: Illustration of the transformation of the input in case $d = 2$, $l = 4$, $I = 10$ and $h = 2$.

For input

$$x = (x_1, \dots, x_l) \in \mathbb{R}^{l \cdot d}$$

it computes in a first step a new representation

$$z_{k,0} = (z_{k,0,1}, \dots, z_{k,0,l}) \in \mathbb{R}^{d_{model} \times l}$$

for some $d_{model} \in \mathbb{N}$ (which will be done in the same way for all $k \in \{1, \dots, K_n\}$). Here $z_{k,0,j}$ is a new representation of $x_j \in \mathbb{R}^d$ of dimension

$$d_{model} = h \cdot I \tag{6}$$

(where $h, I \in \mathbb{N}$ with $I \geq d + l + 4$) which includes the original data, coding of the position and additional auxiliary values used for later computation of function values. More precisely, we set for $s \in \{1, \dots, h \cdot I\}$

$$z_{k,0,j}^{(s)} = \begin{cases} x_j^{(s)} & \text{if } s \in \{1, \dots, d\} \\ 1 & \text{if } s = d + 1 \\ \delta_{s-d-1,j} & \text{if } s \in \{d + 2, \dots, d + 1 + l\} \\ 0 & \text{if } s \in \{d + l + 2, d + l + 3, \dots, h \cdot I\} \end{cases}$$

For $d = 2$, $l = 4$, $I = 10$ and $h = 2$ the transformation of the input is illustrated in Figure 1.

After that we compute successive representations

$$z_{k,r} = (z_{k,r,1}, \dots, z_{k,r,l}) \in \mathbb{R}^{d_{model} \times l} \tag{7}$$

of the input for $r = 1, \dots, N$, and apply a feedforward neural network to $z_{k,N}$. Here $z_{k,r}$ is the representation of the input in the $k$-th transformer network in level $r$. It depends on $l$ parts which correspond to $x_1, \ldots, x_l$. And $N$ is the number of pairs of attention and pointwise feedforward layers of our transformer encoder.

Given $z_{k,r-1}$ for some $r \in \{1, \ldots, N\}$ we compute $z_{k,r}$ by applying first a multi-head attention and afterwards a pointwise feedforward neural network with one hidden layer. Both times we will use an additional residual connection.

The computation of the multi-head attention depends on matrices

$$W_{query,k,r,s}, W_{key,k,r,s} \in \mathbb{R}^{d_{key} \times d_{model}} \quad \text{and} \quad W_{value,k,r,s} \in \mathbb{R}^{d_v \times d_{model}} \quad (s = 1, \ldots, h), \tag{8}$$

where $h \in \mathbb{N}$ is the number of attentions which we compute in parallel, where $d_{key} \in \mathbb{N}$ is the dimension of the queries and the keys, and where $d_v = d_{model}/h = I$ is the dimension of the values. Here each of the $h$ attention heads will be used to compute a new part of length $d_v = I$ of the representation $z_{k,r,i}$ of $x_i$ for $i = 1, \ldots, l$. We use the above matrices to compute for each component $z_{k,r-1,i}$ of $z_{k,r-1}$ (i.e., for each representation of $x_i$ at level $r-1$ $(i = 1, \ldots, l)$) corresponding queries

$$q_{k,r-1,s,i} = W_{query,k,r,s} \cdot z_{k,r-1,i}, \tag{9}$$

keys

$$k_{k,r-1,s,i} = W_{key,k,r,s} \cdot z_{k,r-1,i} \tag{10}$$

and values

$$v_{k,r-1,s,i} = W_{value,k,r,s} \cdot z_{k,r-1,i} \tag{11}$$

$(s \in \{1, \ldots, h\}, i \in \{1, \ldots, l\})$. Then the so-called attention between the component $i$ of $z_{k,r-1}$ and the component $j$ of $z_{k,r-1}$ (i.e., between the representations of $x_i$ and $x_j$ at level $r-1$) is defined as the scalar product

$$< q_{k,r-1,s,i}, k_{k,r-1,s,j} > \tag{12}$$

of the corresponding query and key, and the index $\hat{j}_{k,r-1,s,i}$ for which the maximal value occurs, i.e.,

$$\hat{j}_{k,r-1,s,i} = \arg \max_{j \in \{1,\ldots,l\}} < q_{k,r-1,s,i}, k_{k,r-1,s,j} >, \tag{13}$$

is determined. The value corresponding to this index is multiplied with the maximal attention in (12) in order to define

$$\begin{aligned} \bar{y}_{k,r,s,i} &= v_{k,r-1,s,\hat{j}_{k,r-1,s,i}} \cdot \max_{j \in \{1,\ldots,l\}} < q_{k,r-1,s,i}, k_{k,r-1,s,j} > \\ &= v_{k,r-1,s,\hat{j}_{r-1,s,i}} \cdot < q_{k,r-1,s,i}, k_{k,r-1,s,\hat{j}_{k,r-1,s,i}} > \end{aligned} \tag{14}$$

$(s \in \{1, \ldots, h\}, i \in \{1, \ldots, l\})$. Using a residual connection we compute the output of the multi-head attention by

$$y_{k,r} = z_{k,r-1} + (\bar{y}_{k,r,1}, \ldots, \bar{y}_{k,r,l}) \tag{15}$$

where

$$\bar{y}_{k,r,i} = (\bar{y}_{k,r,1,i}, \ldots, \bar{y}_{k,r,h,i}) \in \mathbb{R}^{d_v \cdot h} = \mathbb{R}^{d_{model}} \quad (i \in \{1, \ldots, l\}).$$

Here $y_{k,r} \in \mathbb{R}^{d_{model} \times l}$ has the same dimension as $z_{k,r-1}$.

The output of the pointwise feedforward neural network depends on parameters

$$W_{k,r,1} \in \mathbb{R}^{d_{ff} \times d_{model}}, b_{k,r,1} \in \mathbb{R}^{d_{ff}}, W_{k,r,2} \in \mathbb{R}^{d_{model} \times d_{ff}}, b_{k,r,2} \in \mathbb{R}^{d_{model}}, \tag{16}$$

which describe the weights in a feedforward neural network with one hidden layer and $d_{ff} \in \mathbb{N}$ hidden neurons. This feedforward neural network is applied to each component of (15) (which is analogous to a convolutionary neural network), i.e., to each representation of $x_1, \ldots, x_l$ computed up to this point on level $r$, and computes

$$z_{k,r,s} = y_{k,r,s} + W_{k,r,2} \cdot \sigma \left( W_{k,r,1} \cdot y_{k,r,s} + b_{k,r,1} \right) + b_{k,r,2} \quad (s \in \{1, \ldots, l\}), \tag{17}$$

where we use again a residual connection. Here

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

is the ReLU activation function, which is applied to a vector by applying it to each component of the vector separately. After computing $z_{k,r,s}$ $(s \in \{1, \ldots, l\})$ we define $z_{k,r}$ by (7).

Given the output $z_{k,N}$ of the sequence of $N$ multi-head attention and pointwise feedforward layers, we apply a (shallow) feedforward neural network with one hidden layer and $J_n$ neurons to $z_{k,N,1}^{(d+l+2)}$, i.e., we set

$$f_{\mathbf{V}_k}(z_{k,N,1}^{(d+l+2)}) = f_{net,J_n,\mathbf{V}_k}(z_{k,N,1}^{(d+l+2)}),$$

where for $z \in \mathbb{R}$ we define

$$f_{net,J_n,\mathbf{V}_k}(z) = \sum_{j=1}^{J_n} v_{k,j}^{(1)} \cdot \sigma \left( v_{k,j,1}^{(0)} \cdot z + v_{k,j,0}^{(0)} \right).$$

Here $\sigma(x) = \max\{x, 0\}$ is again the ReLU activation function, and

$$\mathbf{V}_k = \left( (v_{k,j}^{(1)})_{j=1,\ldots,J_n}, (v_{k,j,1}^{(0)})_{j=1,\ldots,J_n}, (v_{k,j,0}^{(0)})_{j=1,\ldots,J_n} \right)$$

is the matrix of the weights of this feedforward neural network.

Because of

$$\begin{aligned} T_{\beta_n}(z) &= \max\{-\beta_n, \min\{\beta_n, z\}\} = \max\{0, \beta_n - \max\{-\beta_n, -z\}\} - \beta_n \\ &= \max\{0, 2\beta_n - \max\{0, -z + \beta_n\}\} - \beta_n = \sigma(2\beta_n - \sigma((-1) \cdot z + \beta_n)) - \beta_n, \end{aligned}$$

$z \mapsto T_{\beta_n}(z)$ is a neural network with two layers, one hidden neuron per layer and ReLU activation function. This implies that $T_{\beta_n}(f_{\mathbf{V}_k}(z))$ is a feedforward neural network with 3 hidden layers, $J_n$ neurons in layer 1 and one hidden neuron in layers 2 and 3, resp.

The output of our $k$-th transformer network is then

$$T_{\beta_n}(f_{\mathbf{W}_k,\mathbf{V}_k}(x)) = T_{\beta_n}(f_{\mathbf{V}_k}(z_{k,N,1}^{(d+l+2)})),$$

where $z_{k,N,1}^{(d+l+2)}$ is one component of the output $z_{k,N}$ of the $N$ pairs of attention layers and pointwise feedforward layers.

## 2.2 Initialization of the weights

We initialize the weights $\mathbf{w}^{(0)} = (w_k^{(0)}, \mathbf{W}_k^{(0)}, \mathbf{V}_k^{(0)})_{k=1,\dots,K_n}$ as follows: We set

$$w_k^{(0)} = 0 \quad (k = 1, \dots, K_n)$$

and choose the components of all other weight matrices independently from uniform distributions on the interval

$$[-c_4 \cdot n^{c_5}, c_4 \cdot n^{c_5}],$$

where $c_4, c_5 > 0$ are suitably large constants. After that we make a pruning step which depends on a parameter $\tau \in \mathbb{N}$ chosen in Theorem 1 below: We choose for each $k \in \{1, \dots, K_n\}$ and for each attention head in each matrix in each row $\tau \in \mathbb{N}$ of its weights randomly by independent uniform distributions and set all weights not chosen to zero. Similarly, we choose for each $k \in \{1, \dots, K_n\}$ and for each matrix $W_{k,r,1}$ in each row and for each matrix $W_{k,r,2}$ in each column $\tau$ of its weights randomly by uniform distributions and set all weights not chosen to zero. Furthermore we set all entries in $W_{query,k,r,1}$ and $W_{key,k,r,1}$, all entries in the first $d+l+1$ columns of $W_{k,r,2}$, and all entries in the last two rows of $W_{query,k,r,s}$ and $W_{key,k,r,s}$ in columns greater than $d+l+1$ to zero.

## 2.3 Learning of the weights of the transformer network

The aim in choosing the weights $\mathbf{w} = (w_k, \mathbf{W}_k, \mathbf{V}_k)_{k=1,\dots,K_n}$ of our transformer network is the minimization of the empirical logistic loss. Let

$$\varphi(z) = \log(1 + \exp(-z))$$

be the logistic loss (or cross entropy loss). Then the empirical logistic loss of $f_{\mathbf{w}} = f_{(w_k)_{k=1,\dots,K_n}, (\mathbf{W}_k)_{k=1,\dots,K_n}, (\mathbf{V}_k)_{k=1,\dots,K_n}}$ is defined by

$$F_n(\mathbf{w}) = \frac{1}{n} \sum_{i=1}^{n} \varphi(Y_i \cdot f_{\mathbf{w}}(X_i)). \tag{18}$$

We use gradient descent together with a projection step in order to minimize (18). Let $A$ be the set of all $(w_k)_{k=1,\dots,K_n}$ which satisfy (5) and let $B$ be the set of all $(\mathbf{W}_k, \mathbf{V}_k)_{k=1,\dots,K_n}$ which have nonzero components only in components which have not been set to zero in the pruning step of the initialization of the weights and which satisfy

$$\|(\mathbf{W}_k, \mathbf{V}_k)_{k=1,\dots,K_n} - (\mathbf{W}_k^{(0)}, \mathbf{V}_k^{(0)})_{k=1,\dots,K_n}\| \le c_6, \tag{19}$$

where $c_6 > 0$ is a constant which will be chosen sufficiently small in Theorem 1 below. Let $\lambda_n > 0$ be the stepsize of the gradient descent and let $\mathbf{w}^{(0)} = (\mathbf{w}_k^{(0)}, \mathbf{W}_k^{(0)}, \mathbf{V}_k^{(0)})_{k=1,\dots,K_n}$ be defined as in Subsection 2.2. Then we define $\mathbf{w}^{(t)} = (\mathbf{w}_k^{(t)}, \mathbf{W}_k^{(t)}, \mathbf{V}_k^{(t)})_{k=1,\dots,K_n}$ recursively by setting

$$\left(w_k^{(t+1)}\right)_{k=1,\dots,K_n} = Proj_A\left(\left(w_k^{(t)} - \lambda_n \cdot \frac{\partial F_n(\mathbf{w}^{(t)})}{\partial w_k}\right)_{k=1,\dots,K_n}\right)$$

and

$$\begin{aligned}
&(\mathbf{W}_k^{(t+1)}, \mathbf{V}_k^{(t+1)})_{k=1,\dots,K_n} \\
&= Proj_B\left(\left((\mathbf{W}_k^{(t)}, \mathbf{V}_k^{(t)}) - \lambda_n \cdot \nabla_{(\mathbf{W}_k,\mathbf{V}_k)}(F_n(\mathbf{w}^{(t)}))\right)_{k=1,\dots,K_n}\right)
\end{aligned}$$

($t = 0, \dots, t_n - 1$), where $t_n \in \mathbb{N}$ is the number of gradient descent steps which will be chosen in Theorem 1 below.

### 2.4 Definition of the estimate

We define our estimate as the plug-in classifier corresponding to the over-parametrized Transformer network network with weight vector $\mathbf{w}^{(\hat{t})}$ where $\hat{t} \in \{0, 1, \dots, t_n\}$ is the index for which the empirical logistic loss is minimal during the training, i.e., we set

$$\hat{t} = \arg \min_{t \in \{0,1,\dots,t_n\}} F_n(\mathbf{w}^{(t)}), \tag{20}$$

$$f_n(x) = f_{\mathbf{w}^{(\hat{t})}}(x) \tag{21}$$

and

$$\eta_n(x) = sgn(f_n(x)). \tag{22}$$

## 3 Main result

Our main result is the following bound on the difference of the misclassification probability of our estimate and the minimal misclassification probability.

**Theorem 1** *Let $A \geq 1$. Let $(X, Y)$, $(X_1, Y_1)$, ..., $(X_n, Y_n)$ be independent and identically distributed $[-A, A]^{d \cdot l} \times \{-1, 1\}$–valued random variables, and let $m(x) = \mathbf{P}\{Y = 1 | X = x\}$ be the corresponding a posteriori probability. Let $\mathcal{P}$ be a finite subset of $[1, \infty) \times \mathbb{N}$ and assume that $m$ satisfies a hierarchical composition model with some finite level and smoothness and order constraint $\mathcal{P}$ and that all functions $g : \mathbb{R}^K \to \mathbb{R}$ in this hierarchical composition model are Lipschitz continuous and satisfy*

$$\|g\|_{C^q(\mathbb{R}^K)} \leq c_7 < \infty,$$

*where $p = q + s$ with $s \in (0, 1]$ and $q \in \mathbb{N}_0$ (here $(p, K) \in \mathcal{P}$ is the smoothness and order corresponding to $g$ in the hierarchical composition model). Let $K_n \in \mathbb{N}$ be such that*

$$\frac{K_n}{e^{(\log n)^3 \cdot \sqrt{n}}} \to \infty \quad (n \to \infty). \tag{23}$$

*Set $\beta_n = c_3 \cdot \log n$,*

$$h = \left\lceil \max_{(p,K) \in \mathcal{P}} n^{\frac{K}{2p+K}} \right\rceil, \quad d_{ff} = 2 \cdot h + 2, \quad I = \lceil \log n \rceil, \quad J_n = \lceil c_8 \cdot n^{1/3} \rceil, \quad t_n = n \cdot K_n$$

*and*

$$\lambda_n = \frac{1}{t_n},$$

*choose* $\tau \in \{l+1, l+2, , \dots, l+d+1\}$ *and choose* $N \in \mathbb{N}$ *sufficiently large,* $c_6 > 0$ *sufficiently small,* $c_4, c_5 > 0$ *sufficiently large,* $d_{key} \geq 4$*, and define the estimate* $\eta_n$ *as in Section 2.*

**a)** *We have for* $n$ *sufficiently large*

$$\mathbf{P}\{\eta_n(X) \neq Y\} - \min_{\eta:\mathbb{R}^{d\cdot l}\to\{-1,1\}} \mathbf{P}\{\eta(X) \neq Y\} \leq c_9 \cdot (\log n)^3 \cdot \max_{(p,K)\in\mathcal{P}} n^{-\min\left\{\frac{p}{2\cdot(2p+K)}, \frac{1}{6}\right\}}.$$

**b)** *If, in addition,*

$$\mathbf{P}\left\{\max\left\{\frac{\mathbf{P}\{Y=1|X\}}{1-\mathbf{P}\{Y=1|X\}}, \frac{1-\mathbf{P}\{Y=1|X\}}{\mathbf{P}\{Y=1|X\}}\right\} > n^{1/3}\right\} \geq 1 - \frac{1}{n^{1/3}} \quad (n \in \mathbb{N}) \tag{24}$$

*holds, then we have for* $n$ *sufficiently large*

$$\mathbf{P}\{\eta_n(X) \neq Y\} - \min_{\eta:\mathbb{R}^{d\cdot l}\to\{-1,1\}} \mathbf{P}\{\eta(X) \neq Y\} \leq c_{10} \cdot (\log n)^3 \cdot \max_{(p,K)\in\mathcal{P}} n^{-\min\left\{\frac{p}{(2p+K)}, \frac{1}{3}\right\}}.$$

**Remark 1** *The upper bound in parts a) and b) of Theorem 1 do not depend on the dimension* $d \cdot l$ *of* $X$*, hence the Transformer encoder estimate is able to circumvent the curse of dimensionality in case that the a posteriori probability satisfies a suitable hierarchical composition model.*

**Remark 2** *In the definition of the estimate we use twice a projection step in the definition of the gradient descent. Here the projection on the outer weights* $(w_k)_{k=1,\dots,K_n}$ *is our main tool which enables us to show that the over-parametrization of the estimate does not hurt the generalization. The second projection is used to ensure that the change of the inner weights during gradient descent does not hurt the approximation properties of the estimate. For neural networks with smooth activation function it is possible to show that such a projection step is automatically satisfied during gradient descent steps for suitable chosen stepsizes and number of gradient descent steps, cf. Lemma 1 in Drews and Kohler (2023).*

**Remark 3** *The proof of Theorem 1 implies that the result also holds for an estimate where gradient descent is only applied to the outer weights* $(w_k)_{k=1,\dots,K_n}$ *and for all other weights their initial randomly chosen values are not changed. Consequently, our estimate is based on representation guessing and not representation learning.*

**Remark 4** *By assumption (23) the number of parameters of our estimate grows exponential in the sample size, so as in many modern applications of deep learning our estimate uses a massive overfitting.*

## 4 A general result

Let $W \in \mathbb{N}$ and let $\mathbf{\Theta} \subseteq \mathbb{R}^W$ be a closed and convex set of parameter values (weights) for a deep network of a given topology. In the sequel we assume that our aim is to learn the parameter $\vartheta \in \mathbf{\Theta}$ (vector of weights) for a deep network

$$f_\vartheta : \mathbb{R}^{d \cdot l} \to \mathbb{R}$$

from the data $\mathcal{D}_n$ such that

$$sgn(f_\vartheta(x))$$

is a good classifier. We do this by considering linear combinations

$$f_{(\mathbf{w},\vartheta)}(x) = \sum_{k=1}^{K_n} w_k \cdot T_{\beta_n}(f_{\vartheta_k}(x)) \tag{25}$$

of truncated versions of estimates $f_{\vartheta_k}(x)$ $(k = 1, \dots, K_n)$, where $\mathbf{w} = (w_k)_{k=1,\dots,K_n}$ satisfies

$$w_k \geq 0 \quad (k = 1, \dots, K_n) \quad \text{and} \quad \sum_{k=1}^{K_n} w_k \leq 1 \tag{26}$$

and $\vartheta = (\vartheta_1, \dots, \vartheta_{K_n}) \in \mathbf{\Theta}^{K_n}$. Observe that by choosing $w_1 = 1$ and $w_k = 0$ for $k > 1$ we get

$$f_{(\mathbf{w},\vartheta)}(x) = T_{\beta_n}(f_{\vartheta_1}(x))$$

and in this way we can construct an estimate which satisfies

$$sgn(f_{(\mathbf{w},\vartheta)}(x)) = sgn(f_{\vartheta_1}(x))$$

for any $\vartheta_1 \in \mathbf{\Theta}$. And by choosing $K_n$ very large our estimate will be over-parametrized in the sense that the number of parameters of the estimate is much larger than the sample size.

Let

$$\varphi(z) = \log(1 + \exp(-z))$$

be the logistic loss (or cross entropy loss) and let $m(x) = \mathbf{P}\{Y = 1 | X = x\}$. Then

$$f_{\varphi^*}(x) = \begin{cases} \infty & \text{if } m(x) = 1, \\ \log \frac{m(x)}{1-m(x)} & \text{if } 0 < m(x) < 1, \\ -\infty & \text{if } m(x) = 0 \end{cases}$$

minimizes the expected logistic loss, i.e.,

$$\begin{aligned} &\mathbf{E}\{\varphi(Y \cdot f_{\varphi^*}(X))\} \\ &= \mathbf{E}\left\{m(X) \cdot \log(1 + \exp(-f_{\varphi^*}(X))) + (1 - m(X)) \cdot \log(1 + \exp(f_{\varphi^*}(X)))\right\} \\ &= \min_{f:\mathbb{R}^{d \cdot l} \to \bar{\mathbb{R}}} \mathbf{E}\{\varphi(Y \cdot f(X))\} \end{aligned}$$

holds. Because of

$$sgn(f_{\varphi^*}(x)) = \begin{cases} 1 & \text{if } m(x) > \frac{1}{2}, \\ -1 & \text{if } m(x) < \frac{1}{2}, \end{cases}$$

this implies

$$\mathbf{P}\{sgn(f_{\varphi^*}(X)) \neq Y\} = \min_{f:\mathbb{R}^{d \cdot l} \to \{-1,1\}} \mathbf{P}\{f(X) \neq Y\},$$

i.e., we can compute the optimal predictor of $Y$ given $X$ by minimizing the expected logistic loss.

Our aim in choosing $(\mathbf{w}, \vartheta)$ is the minimization of the empirical logistic loss

$$F_n((\mathbf{w}, \vartheta)) = \frac{1}{n} \sum_{i=1}^{n} \varphi(Y_i \cdot f_{(\mathbf{w},\vartheta)}(X_i)).$$

In order to achieve this, we start with a random initialization of $(\mathbf{w}, \vartheta)$: We choose

$$\vartheta_1^{(0)}, \dots, \vartheta_{K_n}^{(0)} \tag{27}$$

randomly from some set $\mathbf{\Theta}^0 \subseteq \mathbf{\Theta}$ such that the random variables in (27) are independent and also independent from $(X, Y)$, $(X_1, Y_1)$, ..., $(X_n, Y_n)$, and we set

$$w_k^{(0)} = 0 \quad (k = 1, \dots, K_n).$$

Then we perform $t_n \in \mathbb{N}$ gradient descent steps starting with

$$\vartheta^{(0)} = (\vartheta_1^{(0)}, \dots, \vartheta_{K_n}^{(0)}) \quad \text{and} \quad \mathbf{w}^{(0)} = (w_1^{(0)}, \dots, w_{K_n}^{(0)}).$$

To do this, we choose a stepsize $\lambda_n > 0$ and set

$$\begin{aligned} \mathbf{w}^{(t+1)} &= Proj_A\left(\mathbf{w}^{(t)} - \lambda_n \cdot \nabla_{\mathbf{w}} F_n((\mathbf{w}^{(t)}, \vartheta^{(t)}))\right), \\ \vartheta^{(t+1)} &= Proj_B\left(\vartheta^{(t)} - \lambda_n \cdot \nabla_{\vartheta} F_n((\mathbf{w}^{(t)}, \vartheta^{(t)}))\right) \end{aligned}$$

for $t = 1, \dots, t_n$. Here $A$ is the set of all $\mathbf{w}$ which satisfy (26), and

$$B = \left\{\vartheta \in \mathbf{\Theta}^{K_n} : \|\vartheta - \vartheta^{(0)}\| \leq c_6\right\},$$

where $c_6 > 0$ is a constant, and $Proj_A$ and $Proj_B$ is the $L_2$ projection on the closed and convex sets $A$ and $B$. (Here closeness and convexity of $B$ is implied by the closeness and convexity of $\mathbf{\Theta}$.) Our estimate is then defined by

$$\hat{t} = \arg \min_{t \in \{0,1,\dots,t_n\}} F_n((\mathbf{w}^{(t)}, \vartheta^{(t)})) \tag{28}$$

and

$$f_n(x) = f_{(\mathbf{w}^{(\hat{t})}, \vartheta^{(\hat{t})})}(x). \tag{29}$$

**Theorem 2** *Let $(X,Y)$, $(X_1,Y_1)$, ..., $(X_n,Y_n)$ be independent and identically distributed random variables with values in $\mathbb{R}^{d\cdot l} \times \{-1,1\}$. Let $t_n, N_n, I_n \in \mathbb{N}$, set*

$$\lambda_n = \frac{1}{t_n}, \quad K_n = N_n \cdot I_n,$$

*choose $c_6 > 0$, and define the estimate $f_n$ as above.*

*Let $\mathbf{\Theta}^* \subset \mathbf{\Theta}^0$ and set*

$$\bar{\mathbf{\Theta}} = \left\{ \vartheta \in \mathbf{\Theta} \ : \ \inf_{\tilde{\vartheta} \in \mathbf{\Theta}^0} \|\vartheta - \tilde{\vartheta}\| \leq c_6 \right\}.$$

*Let $C_n, D_n \geq 0$ and assume*

$$\|f_\vartheta - f_{\vartheta^*}\|_{\infty, supp(X)} \leq C_n \cdot \|\vartheta - \vartheta^*\| \tag{30}$$

*for all $\vartheta^* \in \mathbf{\Theta}^*$ and all $\vartheta \in \{\bar{\vartheta} \in \mathbf{\Theta} \ : \ \|\bar{\vartheta} - \vartheta^*\| \leq c_6\}$,*

$$\epsilon_n = \mathbf{P}\left\{ \vartheta_1^{(0)} \in \mathbf{\Theta}^* \right\} > 0, \tag{31}$$

$$N_n \cdot (1-\epsilon_n)^{I_n} \leq \frac{1}{n} \tag{32}$$

*and*

$$\|\nabla_{\mathbf{w}} F_n((\mathbf{w}, \vartheta))\| \leq D_n \quad \text{for all } \mathbf{w} \in A, \vartheta \in \bar{\mathbf{\Theta}}. \tag{33}$$

*Then we have*

$$\begin{aligned} &\mathbf{E}\left\{\varphi(Y \cdot f_n(X))\right\} - \min_{f:\mathbb{R}^{d\cdot l} \to \bar{\mathbb{R}}} \mathbf{E}\left\{\varphi(Y \cdot f(X))\right\} \\ &\leq c_{11} \cdot \Bigg( \frac{\log n}{n} + \mathbf{E}\left\{ \sup_{\vartheta \in \bar{\mathbf{\Theta}}} \left| \frac{1}{n} \sum_{i=1}^n \epsilon_i \cdot T_{\beta_n}(f_\vartheta(X_i)) \right| \right\} + \frac{C_n + 1}{\sqrt{N_n}} + \frac{D_n^2}{t_n} \\ &\qquad + \sup_{\vartheta \in \mathbf{\Theta}^*} \mathbf{E}\left\{\varphi(Y \cdot f_\vartheta(X))\right\} - \min_{f:\mathbb{R}^{d\cdot l} \to \bar{\mathbb{R}}} \mathbf{E}\left\{\varphi(Y \cdot f(X))\right\} \Bigg), \end{aligned}$$

*where $\epsilon_1$, ..., $\epsilon_n$ are independent and uniformly distributed on $\{-1,1\}$ (so-called Rademacher random variables) and independent from $X_1, \dots, X_n$.*

**Remark 5** *In Theorem 2 the Rademacher complexity*

$$\mathbf{E}\left\{ \sup_{\vartheta \in \bar{\mathbf{\Theta}}} \left| \frac{1}{n} \sum_{i=1}^n \epsilon_i \cdot T_{\beta_n}(f_\vartheta(X_i)) \right| \right\}$$

*is used to control the generalization error of the estimate,*

$$\sup_{\vartheta \in \mathbf{\Theta}^*} \mathbf{E}\left\{\varphi(Y \cdot f_\vartheta(X))\right\} - \min_{f:\mathbb{R}^{d\cdot l} \to \bar{\mathbb{R}}} \mathbf{E}\left\{\varphi(Y \cdot f(X))\right\},$$

*which describes the worst error occuring in the set $\Theta^*$ of "good" parameter values, is used to measure the approximation error, and*

$$\frac{C_n+1}{\sqrt{N_n}} + \frac{D_n^2}{t_n}$$

*is used to bound the error occuring due to gradient descent.*

**Proof of Theorem 2.** Let $E_n$ be the event that there exist pairwise distinct $j_1,\dots,j_{N_n} \in \{1,\dots,K_n\}$ such that

$$\vartheta_{j_i}^{(0)} \in \mathbf{\Theta}^*$$

holds for all $i=1,\dots,N_n$. If $E_n$ holds set

$$w_{j_i}^* = \frac{1}{N_n} \quad (i=1,\dots,N_n) \quad \text{and} \quad w_k^*=0 \quad (k \in \{1,\dots,K_n\} \setminus \{j_1,\dots,j_{N_n}\})$$

and $\mathbf{w}^* = (w_k^*)_{k=1,\dots,K_n}$, otherwise set $\mathbf{w}^*=0$.

We will use the following error decomposition:

$$\begin{aligned}
&\mathbf{E}\left\{\varphi(Y \cdot f_n(X))\right\} - \min_{f:\mathbb{R}^{d\cdot l}\to\bar{\mathbb{R}}} \mathbf{E}\left\{\varphi(Y\cdot f(X))\right\} \\
&= \mathbf{E}\left\{\varphi(Y\cdot f_n(X)) \cdot 1_{E_n^c}\right\} \\
&\quad + \mathbf{E}\left\{\left(\mathbf{E}\left\{\varphi(Y\cdot f_n(X)) \middle| \vartheta^{(0)}, \mathcal{D}_n\right\} - \frac{1}{n}\sum_{i=1}^n \varphi(Y_i \cdot f_n(X_i))\right)\cdot 1_{E_n}\right\} \\
&\quad + \mathbf{E}\left\{\frac{1}{n}\sum_{i=1}^n \varphi(Y_i\cdot f_n(X_i))\cdot 1_{E_n}\right\} - \min_{f:\mathbb{R}^{d\cdot l}\to\bar{\mathbb{R}}} \mathbf{E}\left\{\varphi(Y\cdot f(X))\right\} \\
&=: T_{1,n} + T_{2,n} + T_{3,n}.
\end{aligned}$$

In the *first step of the proof* we show

$$\mathbf{P}\{E_n^c\} \le \frac{1}{n}. \tag{34}$$

To do this we consider a sequential choice of the initial weights $\vartheta_1^{(0)}, \dots, \vartheta_{K_n}^{(0)}$. By definition of $\epsilon_n$ we know that the probability that none of $\vartheta_1^{(0)}, \dots, \vartheta_{I_n}^{(0)}$ is contained in $\mathbf{\Theta}^*$ is given by

$$(1-\epsilon_n)^{I_n}.$$

This implies that the probability that there exists $l \in \{1,\dots,N_n\}$ such that none of $\vartheta_{(l-1)*I_n+1}^{(0)}, \dots, \vartheta_{l\cdot I_n}^{(0)}$ is contained in $\mathbf{\Theta}^*$ is upper bounded by

$$N_n \cdot (1-\epsilon_n)^{I_n}.$$

Using (32) we can conclude

$$\mathbf{P}\{E_n^c\} \le N_n\cdot(1-\epsilon_n)^{I_n} \le \frac{1}{n}.$$

In the *second step of the proof* we show

$$T_{1,n} \leq c_{12} \cdot \frac{(\log n)}{n}.$$

To do this, we observe that for $|z| \leq \beta_n$ we have

$$\varphi(z) = \log(1+\exp(-z)) \leq (\log 4) \cdot I_{\{z>-1\}} + \log(2 \cdot \exp(-z)) \cdot I_{\{z \leq -1\}} \leq 3 + |z| \leq c_{13} \cdot \log n,$$

from which we can conclude by the first step of the proof

$$T_{1,n} \leq c_{13} \cdot (\log n) \cdot \mathbf{P}\{E_n^c\} \leq c_{13} \cdot \frac{(\log n)}{n}.$$

Let $\mathcal{F}$ be the set of all $f_{(\mathbf{w},\vartheta)}$ where $\mathbf{w} \in A$ and $\vartheta \in \bar{\mathbf{\Theta}}^{K_n}$. In the *third step of the proof* we show

$$T_{2,n} \leq \mathbf{E}\left\{ \mathbf{E}\left\{ \sup_{f \in \mathcal{F}} \left( \mathbf{E}\{\varphi(f(X) \cdot Y)\} - \frac{1}{n} \sum_{i=1}^n \varphi(f(X_i) \cdot Y_i) \right) \right\} \cdot 1_{E_n} \right\}.$$

This follows from

$$\begin{aligned}
T_{2,n} &= \mathbf{E}\left\{ \mathbf{E}\left\{ \left( \mathbf{E}\left\{ \varphi(Y \cdot f_n(X)) \middle| \vartheta^{(0)}, \mathcal{D}_n \right\} - \frac{1}{n} \sum_{i=1}^n \varphi(Y_i \cdot f_n(X_i)) \right) \middle| \vartheta^{(0)} \right\} \cdot 1_{E_n} \right\} \\
&\leq \mathbf{E}\left\{ \mathbf{E}\left\{ \sup_{f \in \mathcal{F}} \left( \mathbf{E}\{\varphi(f(X) \cdot Y)\} - \frac{1}{n} \sum_{i=1}^n \varphi(f(X_i) \cdot Y_i) \right) \middle| \vartheta^{(0)} \right\} \cdot 1_{E_n} \right\} \\
&= \mathbf{E}\left\{ \mathbf{E}\left\{ \sup_{f \in \mathcal{F}} \left( \mathbf{E}\{\varphi(f(X) \cdot Y)\} - \frac{1}{n} \sum_{i=1}^n \varphi(f(X_i) \cdot Y_i) \right) \right\} \cdot 1_{E_n} \right\}.
\end{aligned}$$

Here the first inequality followed from $\mathbf{w}^{(t)} \in A$ and $\vartheta^{(t)} \in \bar{\mathbf{\Theta}}^{K_n}$ $(t \in \{0, 1, \dots, t_n\})$.

In the *fourth step of the proof* we show

$$\begin{aligned}
&\mathbf{E}\left\{ \sup_{f \in \mathcal{F}} \left( \mathbf{E}\{\varphi(f(X) \cdot Y)\} - \frac{1}{n} \sum_{i=1}^n \varphi(f(X_i) \cdot Y_i) \right) \right\} \\
&\leq 2 \cdot \mathbf{E}\left\{ \sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^n \epsilon_i \cdot f(X_i) \right\}. \qquad (35)
\end{aligned}$$

Choose random variables $(X_1', Y_1')$, ..., $(X_n', Y_n')$ such that

$$(X_1, Y_1), \dots, (X_n, Y_n), \epsilon_1, \dots, \epsilon_n, (X_1', Y_1'), \dots, (X_n', Y_n')$$

are independent and such that

$$(X_1, Y_1), \dots, (X_n, Y_n), (X_1', Y_1'), \dots, (X_n', Y_n')$$

are identically distributed and set $(X,Y)_1^n = ((X_1,Y_1),\dots,(X_n,Y_n))$. We have

$$\begin{aligned}
&\mathbf{E}\left\{\sup_{f\in\mathcal{F}}\left(\mathbf{E}\{\varphi(f(X)\cdot Y)\} - \frac{1}{n}\sum_{i=1}^n \varphi(f(X_i)\cdot Y_i)\right)\right\}\\
&= \mathbf{E}\left\{\sup_{f\in\mathcal{F}}\left(\mathbf{E}\{\frac{1}{n}\sum_{i=1}^n \varphi(f(X_i')\cdot Y_i')|(X,Y)_1^n\} - \frac{1}{n}\sum_{i=1}^n \varphi(f(X_i)\cdot Y_i)\right)\right\}\\
&\le \mathbf{E}\left\{\mathbf{E}\left\{\sup_{f\in\mathcal{F}}\left(\frac{1}{n}\sum_{i=1}^n \varphi(f(X_i')\cdot Y_i') - \frac{1}{n}\sum_{i=1}^n \varphi(f(X_i)\cdot Y_i)\right)|(X,Y)_1^n\right\}\right\}\\
&= \mathbf{E}\left\{\sup_{f\in\mathcal{F}}\left(\frac{1}{n}\sum_{i=1}^n \varphi(f(X_i')\cdot Y_i') - \frac{1}{n}\sum_{i=1}^n \varphi(f(X_i)\cdot Y_i)\right)\right\}.
\end{aligned}$$

Since the joint distribution of $(X_1,Y_1),\dots,(X_n,Y_n),(X_1',Y_1'),\dots,(X_n',Y_n')$ does not change if we (randomly) interchange $(X_i,Y_i)$ and $(X_i',Y_i')$, the last term is equal to

$$\begin{aligned}
&\mathbf{E}\left\{\sup_{f\in\mathcal{F}}\left(\frac{1}{n}\sum_{i=1}^n \epsilon_i\cdot\big(\varphi(f(X_i')\cdot Y_i') - \varphi(f(X_i)\cdot Y_i)\big)\right)\right\}\\
&\le \mathbf{E}\left\{\sup_{f\in\mathcal{F}}\left(\frac{1}{n}\sum_{i=1}^n \epsilon_i\cdot\varphi(f(X_i')\cdot Y_i')\right)\right\} + \mathbf{E}\left\{\sup_{f\in\mathcal{F}}\left(\frac{1}{n}\sum_{i=1}^n (-\epsilon_i)\cdot\varphi(f(X_i)\cdot Y_i)\right)\right\}\\
&= 2\cdot\mathbf{E}\left\{\sup_{f\in\mathcal{F}}\left(\frac{1}{n}\sum_{i=1}^n \epsilon_i\cdot\varphi(f(X_i)\cdot Y_i)\right)\right\}.
\end{aligned}$$

Next we use a contraction-style argument. Because of the independence of the random variables we can compute the expectation by first computing the expectation with respect to $\epsilon_1$ and then computing the expectation with respect to all other random variables. Consequently, the last term above is equal to

$$\begin{aligned}
&2\cdot\mathbf{E}\Bigg\{\frac{1}{2}\cdot\sup_{f\in\mathcal{F}}\left(\frac{1}{n}\sum_{i=2}^n \epsilon_i\cdot\varphi(f(X_i)\cdot Y_i) + \frac{1}{n}\cdot\varphi(f(X_1)\cdot Y_1)\right)\\
&\qquad + \frac{1}{2}\cdot\sup_{g\in\mathcal{F}}\left(\frac{1}{n}\sum_{i=2}^n \epsilon_i\cdot\varphi(g(X_i)\cdot Y_i) - \frac{1}{n}\cdot\varphi(g(X_1)\cdot Y_1)\right)\Bigg\}\\
&= \mathbf{E}\Bigg\{\sup_{f,g\in\mathcal{F}}\Bigg(\frac{1}{n}\sum_{i=2}^n \epsilon_i\cdot\varphi(f(X_i)\cdot Y_i) + \frac{1}{n}\sum_{i=2}^n \epsilon_i\cdot\varphi(g(X_i)\cdot Y_i)\\
&\qquad + \frac{1}{n}\cdot\varphi(f(X_1)\cdot Y_1) - \frac{1}{n}\cdot\varphi(g(X_1)\cdot Y_1)\Bigg)\Bigg\}.
\end{aligned}$$

Because of

$$\varphi'(z) = \frac{1}{1+\exp(-z)}\cdot\exp(-z)\cdot(-1) \in [-1,0],$$

$\varphi$ is Lipschitz continuous with Lipschitz constant 1 which implies

$$\begin{aligned}\frac{1}{n}\cdot\varphi(f(X_1)\cdot Y_1)-\frac{1}{n}\cdot\varphi(g(X_1)\cdot Y_1) &\leq \frac{1}{n}\cdot|f(X_1)\cdot Y_1-g(X_1)\cdot Y_1| \\ &\leq \frac{1}{n}\cdot|f(X_1)-g(X_1)|.\end{aligned}$$

Hence the last expectation above is upper bounded by

$$\mathbf{E}\left\{\sup_{f,g\in\mathcal{F}}\left(\frac{1}{n}\sum_{i=2}^{n}\epsilon_i\cdot\varphi(f(X_i)\cdot Y_i)+\frac{1}{n}\sum_{i=2}^{n}\epsilon_i\cdot\varphi(g(X_i)\cdot Y_i)+\frac{1}{n}\cdot|f(X_1)-g(X_1)|\right)\right\}.$$

For fixed $(X_1,Y_1)$, ..., $(X_n,Y_n)$, $\epsilon_2$, ..., $\epsilon_n$ the term

$$\frac{1}{n}\sum_{i=2}^{n}\epsilon_i\cdot\varphi(f(X_i)\cdot Y_i)+\frac{1}{n}\sum_{i=2}^{n}\epsilon_i\cdot\varphi(g(X_i)\cdot Y_i)+\frac{1}{n}\cdot|f(X_1)-g(X_1)|$$

is symmetric in $f$ and $g$. Therefore we can assume w.l.o.g. that $f(X_1)\geq g(X_1)$ holds which implies that we have

$$\begin{aligned}&\sup_{f,g\in\mathcal{F}}\left(\frac{1}{n}\sum_{i=2}^{n}\epsilon_i\cdot\varphi(f(X_i)\cdot Y_i)+\frac{1}{n}\sum_{i=2}^{n}\epsilon_i\cdot\varphi(g(X_i)\cdot Y_i)+\frac{1}{n}\cdot|f(X_1)-g(X_1)|\right)\\ &=\sup_{f,g\in\mathcal{F}}\left(\frac{1}{n}\sum_{i=2}^{n}\epsilon_i\cdot\varphi(f(X_i)\cdot Y_i)+\frac{1}{n}\sum_{i=2}^{n}\epsilon_i\cdot\varphi(g(X_i)\cdot Y_i)+\frac{1}{n}\cdot(f(X_1)-g(X_1))\right).\end{aligned}$$

In the same way we see that the term above is also equal to

$$\sup_{f,g\in\mathcal{F}}\left(\frac{1}{n}\sum_{i=2}^{n}\epsilon_i\cdot\varphi(f(X_i)\cdot Y_i)+\frac{1}{n}\sum_{i=2}^{n}\epsilon_i\cdot\varphi(g(X_i)\cdot Y_i)-\frac{1}{n}\cdot(f(X_1)-g(X_1))\right),$$

and we get

$$\begin{aligned}&\mathbf{E}\Bigg\{\sup_{f,g\in\mathcal{F}}\Bigg(\frac{1}{n}\sum_{i=2}^{n}\epsilon_i\cdot\varphi(f(X_i)\cdot Y_i)+\frac{1}{n}\sum_{i=2}^{n}\epsilon_i\cdot\varphi(g(X_i)\cdot Y_i)\\ &\qquad+\frac{1}{n}\cdot|f(X_1)-g(X_1)|\Bigg)\Bigg\}\\ &=\mathbf{E}\Bigg\{\frac{1}{2}\cdot\sup_{f,g\in\mathcal{F}}\Bigg(\frac{1}{n}\sum_{i=2}^{n}\epsilon_i\cdot\varphi(f(X_i)\cdot Y_i)+\frac{1}{n}\sum_{i=2}^{n}\epsilon_i\cdot\varphi(g(X_i)\cdot Y_i)\\ &\qquad+\frac{1}{n}\cdot(f(X_1)-g(X_1))\Bigg)\\ &\quad+\frac{1}{2}\cdot\sup_{f,g\in\mathcal{F}}\Bigg(\frac{1}{n}\sum_{i=2}^{n}\epsilon_i\cdot\varphi(f(X_i)\cdot Y_i)+\frac{1}{n}\sum_{i=2}^{n}\epsilon_i\cdot\varphi(g(X_i)\cdot Y_i)\end{aligned}$$

$$
\begin{aligned}
&\qquad\qquad -\frac{1}{n}\cdot (f(X_1)-g(X_1))\Bigg)\Bigg\}\\
&= \mathbf{E}\Bigg\{\sup_{f,g\in\mathcal{F}}\Bigg(\frac{1}{n}\sum_{i=2}^{n}\epsilon_i\cdot\varphi(f(X_i)\cdot Y_i)+\frac{1}{n}\sum_{i=2}^{n}\epsilon_i\cdot\varphi(g(X_i)\cdot Y_i)\\
&\qquad\qquad +\frac{1}{n}\cdot\epsilon_1\cdot(f(X_1)-g(X_1))\Bigg)\Bigg\}\\
&\le \mathbf{E}\Bigg\{\sup_{f,g\in\mathcal{F}}\Bigg(\frac{1}{n}\sum_{i=2}^{n}\epsilon_i\cdot\varphi(f(X_i)\cdot Y_i)+\frac{1}{n}\cdot\epsilon_1\cdot f(X_1)\Bigg)\Bigg\}\\
&\qquad + \sup_{f,g\in\mathcal{F}}\Bigg(\frac{1}{n}\sum_{i=2}^{n}\epsilon_i\cdot\varphi(g(X_i)\cdot Y_i)+\frac{1}{n}\cdot(-\epsilon_1)\cdot g(X_1)\Bigg)\Bigg\}\\
&= 2\cdot\mathbf{E}\Bigg\{\sup_{f\in\mathcal{F}}\Bigg(\frac{1}{n}\sum_{i=2}^{n}\epsilon_i\cdot\varphi(f(X_i)\cdot Y_i)+\frac{1}{n}\cdot\epsilon_1\cdot f(X_1)\Bigg)\Bigg\},
\end{aligned}
$$

where we have used that $-\epsilon_1$ has the same distribution as $\epsilon_1$.

Arguing in the same way for $i=2,\dots,n$ we get

$$
\begin{aligned}
&2\cdot\mathbf{E}\left\{\sup_{f\in\mathcal{F}}\left(\frac{1}{n}\sum_{i=1}^{n}\epsilon_i\cdot\varphi(f(X_i)\cdot Y_i)\right)\right\}\\
&\le 2\cdot\mathbf{E}\Bigg\{\sup_{f\in\mathcal{F}}\Bigg(\frac{1}{n}\sum_{i=2}^{n}\epsilon_i\cdot\varphi(f(X_i)\cdot Y_i)+\frac{1}{n}\cdot\epsilon_1\cdot f(X_1)\Bigg)\Bigg\}\\
&\le 2\cdot\mathbf{E}\Bigg\{\sup_{f\in\mathcal{F}}\Bigg(\frac{1}{n}\sum_{i=3}^{n}\epsilon_i\cdot\varphi(f(X_i)\cdot Y_i)+\frac{1}{n}\cdot(\epsilon_1\cdot f(X_1)+\epsilon_2\cdot f(X_2))\Bigg)\Bigg\}\\
&\le \dots\\
&\le 2\cdot\mathbf{E}\left\{\sup_{f\in\mathcal{F}}\frac{1}{n}\cdot\sum_{i=1}^{n}\epsilon_i\cdot f(X_i)\right\},
\end{aligned}
$$

which finishes the fourth step of the proof.

In the *fifth step of the proof* we show

$$
\mathbf{E}\left\{\sup_{f\in\mathcal{F}}\frac{1}{n}\sum_{i=1}^{n}\epsilon_i\cdot f(X_i)\right\}\cdot 1_{E_n}\le \mathbf{E}\left\{\sup_{\vartheta\in\bar{\mathbf{\Theta}}}\left|\frac{1}{n}\sum_{i=1}^{n}\epsilon_i\cdot(T_{\beta_n}(f_\vartheta(X_i))\right|\right\}.
$$

Let $\mathcal{W}$ be the set of all weight vectors $\mathbf{w}=((w_k)_{k=1,\dots,K_n},(\vartheta_k)_{k=1,\dots,K_n})$ which satisfy $\vartheta=(\vartheta_k)_{k=1,\dots,K_n}\in\bar{\mathbf{\Theta}}^{K_n}$ and (26). Because of $f=0$ is contained in $\mathcal{F}$ it implies

$$
\sup_{f\in\mathcal{F}}\frac{1}{n}\sum_{i=1}^{n}\epsilon_i\cdot f(X_i)\ge 0
$$

from which we can conclude

$$
\begin{aligned}
& \mathbf{E}\left\{\sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^n \epsilon_i \cdot f(X_i)\right\} \cdot 1_{E_n} \\
&\leq \mathbf{E}\left\{\sup_{\mathbf{w}\in\mathcal{W}} \frac{1}{n}\sum_{i=1}^n \epsilon_i \cdot \sum_{j=1}^{K_n} w_j \cdot (T_{\beta_n} f_{\vartheta_j}(X_i)))\right\} \\
&= \mathbf{E}\left\{\sup_{\mathbf{w}\in\mathcal{W}} \sum_{j=1}^{K_n} w_j \cdot \frac{1}{n}\sum_{i=1}^n \epsilon_i \cdot (T_{\beta_n} f_{\vartheta_j}(X_i)))\right\} \\
&\leq \mathbf{E}\left\{\sup_{\mathbf{w}\in\mathcal{W}} \sum_{j=1}^{K_n} |w_j| \cdot \left|\frac{1}{n}\sum_{i=1}^n \epsilon_i \cdot (T_{\beta_n} f_{\vartheta_j}(X_i)))\right|\right\} \\
&\leq \mathbf{E}\left\{\sup_{\mathbf{w}\in\mathcal{W}} \sum_{j=1}^{K_n} |w_j| \cdot \sup_{\vartheta\in\bar{\mathbf{\Theta}}^{K_n}, k\in\{1,\dots,K_n\}} \left|\frac{1}{n}\sum_{i=1}^n \epsilon_i \cdot (T_{\beta_n} f_{\vartheta_k}(X_i)))\right|\right\} \\
&\leq 1 \cdot \mathbf{E}\left\{\sup_{\vartheta\in\bar{\mathbf{\Theta}}^{K_n}, k\in\{1,\dots,K_n\}} \left|\frac{1}{n}\sum_{i=1}^n \epsilon_i \cdot (T_{\beta_n} f_{\vartheta_k}(X_i)))\right|\right\} \\
&= \mathbf{E}\left\{\sup_{\vartheta\in\bar{\mathbf{\Theta}}^{K_n}} \left|\frac{1}{n}\sum_{i=1}^n \epsilon_i \cdot (T_{\beta_n} f_{\vartheta_1}(X_i)))\right|\right\} \\
&= \mathbf{E}\left\{\sup_{\vartheta\in\bar{\mathbf{\Theta}}} \left|\frac{1}{n}\sum_{i=1}^n \epsilon_i \cdot (T_{\beta_n}(f_{\vartheta}(X_i))\right|\right\},
\end{aligned}
$$

where the last inequality followed from

$$
\{T_{\beta_n} f_{\vartheta_k} \,:\, \vartheta \in \bar{\mathbf{\Theta}}^{K_n}, k \in \{1,\dots,K_n\}\} = \{T_{\beta_n} f_{\vartheta_1} \,:\, \vartheta \in \bar{\mathbf{\Theta}}\}.
$$

In the *sixth step of the proof* we show

$$
\begin{aligned}
T_{3,n} \quad \leq \quad c_{14} \cdot \Bigg( & \frac{C_n+1}{\sqrt{N_n}} + \frac{D_n^2}{t_n} + \sup_{\vartheta\in\mathbf{\Theta}^*} \mathbf{E}\left\{\varphi(Y \cdot f_{\vartheta}(X))\right\} \\
& - \min_{f:\mathbb{R}^{d\cdot l}\to\bar{\mathbb{R}}} \mathbf{E}\left\{\varphi(Y \cdot f(X))\right\} \Bigg).
\end{aligned}
$$

Application of standard techniques concerning the analysis of gradient descent in case of convex function (cf., Lemma 1) yields

$$
\begin{aligned}
& \frac{1}{n}\sum_{i=1}^n \varphi(Y_i \cdot f_n(X_i)) \cdot 1_{E_n} \\
&= \min_{t=0,\dots,t_n} F_n((\mathbf{w}^{(t)}, \vartheta^{(t)})) \cdot 1_{E_n} \\
&\leq F_n((\mathbf{w}^*, \vartheta^{(0)})) \cdot 1_{E_n} + \frac{1}{t_n} \cdot \sum_{t=1}^{t_n} |F_n((\mathbf{w}^*, \vartheta^{(t)})) - F_n((\mathbf{w}^*, \vartheta^{(0)}))| \cdot 1_{E_n}
\end{aligned}
$$

$$+\frac{\|\mathbf{w}^*\|^2}{2} + \frac{D_n^2}{2 \cdot t_n}.$$

By the definition of $\mathbf{w}^*$ we know

$$\frac{\|\mathbf{w}^*\|^2}{2} \leq \frac{1}{2 \cdot N_n}.$$

The logistic loss is convex (since $\varphi''(z) \geq 0$ for all $z \in \mathbb{R}$) from which we can conclude

$$\begin{aligned}
&\mathbf{E}\left\{F_n((\mathbf{w}^*, \vartheta^{(0)})) \cdot 1_{E_n}\right\} \\
&= \mathbf{E}\left\{\mathbf{E}\left\{F_n((\mathbf{w}^*, \vartheta^{(0)}))\big|\vartheta^{(0)}\right\} \cdot 1_{E_n}\right\} \\
&= \mathbf{E}\left\{\mathbf{E}\left\{\frac{1}{n}\sum_{i=1}^{n}\varphi(\frac{1}{N_n}\sum_{k=1}^{N_n} Y_i \cdot f_{\vartheta_{j_k}^{(0)}}(X_i))\big|\vartheta^{(0)}\right\} \cdot 1_{E_n}\right\} \\
&\leq \mathbf{E}\left\{\frac{1}{N_n}\sum_{k=1}^{N_n}\mathbf{E}\left\{\frac{1}{n}\sum_{i=1}^{n}\varphi(Y_i \cdot f_{\vartheta_{j_k}^{(0)}}(X_i))\big|\vartheta^{(0)}\right\} \cdot 1_{E_n}\right\} \\
&\leq \sup_{\vartheta \in \Theta^*} \mathbf{E}\left\{\varphi(Y \cdot f_\vartheta(X)\right\}.
\end{aligned}$$

Finally we conclude from the fact that $\varphi$ is Lipschitz continuous with Lipschitz constant 1, the Cauchy-Schwarz inequality and assumption (30) that we have

$$\begin{aligned}
&\frac{1}{t_n} \cdot \sum_{t=1}^{t_n} |F_n((\mathbf{w}^*, \vartheta^{(t)})) - F_n((\mathbf{w}^*, \vartheta^{(0)}))| \cdot 1_{E_n} \\
&= \frac{1}{t_n} \cdot \sum_{t=1}^{t_n} \left|\frac{1}{n}\sum_{i=1}^{n}\Big(\varphi(Y_i \cdot f_{(\mathbf{w}^*, \vartheta^{(t)})}(X_i)) - \varphi(Y_i \cdot f_{(\mathbf{w}^*, \vartheta^{(0)})}(X_i))\Big)\right| \cdot 1_{E_n} \\
&\leq \max_{t=1,\dots,t_n} \max_{i=1,\dots,n} |f_{(\mathbf{w}^*, \vartheta^{(t)})}(X_i) - f_{(\mathbf{w}^*, \vartheta^{(0)})}(X_i)| \cdot 1_{E_n} \\
&\leq \max_{t=1,\dots,t_n} \max_{i=1,\dots,n} \sqrt{\sum_{k=1}^{K_n} |w_k^*|^2} \cdot \sqrt{\sum_{i=1}^{N_n} |f_{\vartheta_{j_i}^{(t)}}(X_i) - f_{\vartheta_{j_i}^{(0)}}(X_i)|^2 \cdot 1_{E_n}} \\
&\leq \frac{1}{\sqrt{N_n}} \cdot \sqrt{\sum_{i=1}^{N_n} C_n^2 \cdot \|\vartheta_{j_i}^{(t)} - \vartheta_{j_i}^{(0)}\|^2} = \frac{1}{\sqrt{N_n}} \cdot C_n \cdot \sqrt{\sum_{i=1}^{N_n} \|\vartheta_{j_i}^{(t)} - \vartheta_{j_i}^{(0)}\|^2} \\
&\leq \frac{1}{\sqrt{N_n}} \cdot C_n \cdot \|\vartheta^{(t)} - \vartheta^{(0)}\| \leq c_6 \cdot \frac{C_n}{\sqrt{N_n}}.
\end{aligned}$$

Here (30) is applicable because the definition of the estimate implies

$$\|\vartheta_{j_i}^{(t)} - \vartheta_{j_i}^{(0)}\| \quad \leq \quad \sqrt{\sum_{s=1}^{N_n} \|\vartheta_{j_s}^{(t)} - \vartheta_{j_s}^{(0)}\|^2} \leq \sqrt{\|\vartheta^{(t)} - \vartheta^{(0)}\|^2} \leq c_6.$$

Gathering the above results completes the proof. □

## 5 Proof of Theorem 1

In the sequel we show

$$\begin{aligned}
&\mathbf{E}\left\{\varphi(Y \cdot sgn(\hat{f}_n(X)))\right\} - \mathbf{E}\left\{\varphi(Y \cdot f_\varphi^*(X))\right\} \\
&\leq c_{85} \cdot (\log n)^6 \cdot \max_{(p,K)\in\mathcal{P}} n^{-\min\left\{\frac{p}{(2p+K)}, \frac{1}{3}\right\}}. \qquad (36)
\end{aligned}$$

This implies the assertion, because by Lemma 2 a) we conclude from (36)

$$\begin{aligned}
&\mathbf{P}\left\{Y \neq sgn(f_n(X))|\right\} - \mathbf{P}\left\{Y \neq \eta^*(X)\right\} \\
&\leq \mathbf{E}\left\{\frac{1}{\sqrt{2}} \cdot \left(\mathbf{E}\left\{\varphi(Y \cdot f_n(X))|\mathcal{D}_n\right\} - \mathbf{E}\left\{\varphi(Y \cdot f_{\varphi^*}(X))\right\}\right)^{1/2}\right\} \\
&\leq \frac{1}{\sqrt{2}} \cdot \sqrt{\mathbf{E}\left\{\varphi(Y \cdot f_n(X))\right\} - \mathbf{E}\left\{\varphi(Y \cdot f_{\varphi^*}(X))\right\}} \\
&\leq c_{86} \cdot (\log n)^3 \cdot \max_{(p,K)\in\mathcal{P}} n^{-\min\left\{\frac{p}{2\cdot(2p+K)}, \frac{1}{6}\right\}}
\end{aligned}$$

And from Lemma 2 b), (24) and Lemma 2 c) we conclude from (36)

$$\begin{aligned}
&\mathbf{P}\left\{Y \neq sgn(f_n(X))|\right\} - \mathbf{P}\left\{Y \neq \eta^*(X)\right\} \\
&\leq 2 \cdot \left(\mathbf{E}\left\{\varphi(Y \cdot f_n(X))\right\} - \mathbf{E}\left\{\varphi(Y \cdot f_{\varphi^*}(X))\right\}\right) + 4 \cdot \frac{c_{87} \cdot \log n}{n^{1/3}} \\
&\leq c_{88} \cdot (\log n)^3 \cdot \max_{(p,K)\in\mathcal{P}} n^{-\min\left\{\frac{p}{2p+K}, \frac{1}{3}\right\}}.
\end{aligned}$$

Here we have used the fact that

$$\max\left\{\frac{\mathbf{P}\{Y=1|X\}}{1-\mathbf{P}\{Y=1|X\}}, \frac{1-\mathbf{P}\{Y=1|X\}}{\mathbf{P}\{Y=1|X\}}\right\} > n^{1/3}$$

is equivalent to

$$|f_{\varphi^*}(X)| = \left|\log\frac{\mathbf{P}\{Y=1|X\}}{1-\mathbf{P}\{Y=1|X\}}\right| > \frac{1}{3} \cdot \log n.$$

So it suffices to prove (36), which we do in the sequel by applying Theorem 2.

In the *first step of the proof* we define $\mathbf{\Theta}$, $\mathbf{\Theta}^0$ and $\mathbf{\Theta}^*$.

Let $\mathbf{\Theta} = \mathbf{\Theta}_0$ be the set of all pairs $(\mathbf{W}, \mathbf{V})$ of weight matrices of the transformer networks $f_{(\mathbf{W}_k,\mathbf{V}_k)}$ introduced in Subsection 2.1. In the supplement we will introduce Transformer networks with good approximation properties, and we use here these Transformer networks for the definition of $\mathbf{\Theta}^*$: Let $\mathbf{\Theta}^*$ be the set of all weight matrices $(\mathbf{W}, \mathbf{V})$ where $\mathbf{W}$ is from the weight matrices introduced in Theorem 3 in supremum norm not further away than

$$\epsilon = \frac{1}{c_{89} \cdot n^{c_{90}}},$$

and where $\mathbf{V}$ is from the weight matrix introduced in Lemma 12 in supremum norm not further away than

$$\bar{\epsilon} = \frac{1}{c_{91} \cdot n^{c_{92}}}.$$

In the *second step of the proof* we show that

$$C_n = c_{93} \cdot n^{c_{94}}$$

satisfies (30). We will show this in the Supplement in Lemma 13, which is applicable provided we choose $c_6 \leq 1/(2 \cdot c_{62})$.

In the *third step of the proof* we show that $\epsilon_n$ defined by (31) satisfies

$$\epsilon_n \geq \frac{1}{e^{(\log n)^2 \cdot \sqrt{n}}}.$$

The event $\{\theta_1^{(0)} \in \mathbf{\Theta}^*\}$ occurs if the pruning step selects the right subset of size $L_n \leq c_{95} \cdot \sqrt{n}$ out of all subsets of size $L_n$ of the possible set of parameters, which has size less equal than $c_{96} \cdot n$, and if the uniform distributions (on intervals of length $2 \cdot c_4 \cdot n^{c_5}$) choose each of these $L_n$ parameters correctly from an interval of size $1/(c_{97} \cdot n^{c_{98}})$. This implies for large $n$

$$\mathbf{P}\{\theta_1^{(0)} \in \mathbf{\Theta}^*\} \geq \frac{1}{(c_{95} \cdot n)^{c_{96} \cdot \sqrt{n}}} \cdot \left( \frac{1}{2 \cdot c_4 \cdot n^{c_5} \cdot c_{97} \cdot n^{c_{98}}} \right)^{\sqrt{n}} \geq \frac{1}{e^{(\log n)^2 \cdot \sqrt{n}}}.$$

In the *fourth step of the proof* we show that

$$N_n = n^{c_{99}}, \quad I_n = \lceil (\log n)^2 \cdot e^{(\log n)^2 \cdot \sqrt{n}} \rceil$$

satisfies (32) for $n$ large.

This follows from

$$\begin{aligned} N_n \cdot (1 - \epsilon_n)^{I_n} &\leq n^{c_{99}} \cdot \left( 1 - \frac{1}{e^{(\log n)^2 \cdot \sqrt{n}}} \right)^{\lceil (\log n)^2 \cdot e^{(\log n)^2 \cdot \sqrt{n}} \rceil} \\ &\leq \frac{1}{n} \end{aligned}$$

for $n$ large.

In the *fifth step of the proof* we show that

$$D_n = \sqrt{K_n} \cdot \beta_n$$

satisfies (33).

We will show this in Lemma 3 in the Supplement.

In the *sixth step of the proof* we show

$$\mathbf{E}\left\{ \left| \sup_{\vartheta \in \bar{\mathbf{\Theta}}} \frac{1}{n} \sum_{i=1}^{n} \epsilon_i \cdot T_{\beta_n}(f_\vartheta(X_i)) \right| \right\} \leq c_{100} \cdot (\log n)^3 \cdot \left( \max_{(p,K) \in \mathcal{P}} n^{-\frac{p}{2p+K}} + n^{-\frac{1}{3}} \right).$$

To see this, we use standard techniques from empirical process theory which are summarized in Lemma 4 in the Supplement. From this we conclude

$$\begin{aligned}
&\mathbf{E}\left\{\left|\sup_{\vartheta\in\bar{\mathbf{\Theta}}} \frac{1}{n}\sum_{i=1}^n \epsilon_i \cdot T_{\beta_n}(f_\vartheta(X_i))\right|\right\} \\
&\leq c_{101} \cdot \frac{\sqrt{\max\{h \cdot I, d_{ff}, J_n\}} \cdot (\log n)^2}{\sqrt{n}} \\
&\leq c_{102} \cdot \frac{(\sqrt{\log n} \cdot \max_{(p,K)\in\mathcal{P}} n^{K/(2\cdot(2\cdot p+K))} + n^{\frac{1}{6}}) \cdot (\log n)^2}{\sqrt{n}} \\
&\leq c_{103} \cdot (\log n)^3 \cdot \left(\max_{(p,K)\in\mathcal{P}} n^{-\frac{p}{2p+K}} + n^{-\frac{1}{3}}\right).
\end{aligned}$$

In the *seventh of the proof* we show

$$\sup_{\vartheta\in\mathbf{\Theta}^*} \mathbf{E}\left\{\varphi(Y \cdot f_\vartheta(X))\right\} - \min_{f:\mathbb{R}^d\to\bar{\mathbb{R}}} \mathbf{E}\left\{\varphi(Y \cdot f(X))\right\} \leq c_{104} \cdot \frac{\log n}{n^{1/3}} + c_{105} \cdot \max_{j,i} h^{-p_j^{(i)}/K_j^{(i)}}.$$

We have

$$\begin{aligned}
&\sup_{\vartheta\in\mathbf{\Theta}^*} \mathbf{E}\left\{\varphi(Y \cdot f_\vartheta(X))\right\} - \min_{f:\mathbb{R}^d\to\bar{\mathbb{R}}} \mathbf{E}\left\{\varphi(Y \cdot f(X))\right\} \\
&= \sup_{\vartheta\in\mathbf{\Theta}^*} \mathbf{E}\left\{\varphi(Y \cdot f_\vartheta(X)) - \varphi(Y \cdot f_{\varphi^*}(X))\right\} \\
&= \sup_{\vartheta\in\mathbf{\Theta}^*} \mathbf{E}\Bigg\{ 1_{\{Y=1\}} \cdot (\varphi(f_\vartheta(X)) - \varphi(f_{\varphi^*}(X))) \\
&\qquad\qquad + 1_{\{Y=-1\}} \cdot (\varphi(-f_\vartheta(X)) - \varphi(-f_{\varphi^*}(X))) \Bigg\} \\
&= \sup_{\vartheta\in\mathbf{\Theta}^*} \mathbf{E}\Bigg\{ m(X) \cdot (\varphi(f_\vartheta(X)) - \varphi(f_{\varphi^*}(X))) \\
&\qquad\qquad + (1-m(X)) \cdot (\varphi(-f_\vartheta(X)) - \varphi(-f_{\varphi^*}(X))) \Bigg\} \\
&\leq \sup_{\vartheta\in\mathbf{\Theta}^*} \sup_{x\in\mathbb{R}^{d\cdot l}} \Bigg( |m(x)| \cdot |\varphi(f_\vartheta(x)) - \varphi(f_{\varphi^*}(x))| \\
&\qquad\qquad + |1-m(x)| \cdot |\varphi(-f_\vartheta(x)) - \varphi(-f_{\varphi^*}(x))| \Bigg).
\end{aligned}$$

Application of the approximation results for Transformer networks derived in the Supplement, i.e., application of Lemma 12 and Theorem 3 (which is applicable because of the pruning step in the gradient descent introduced in Section 2, which implies in particular that some components of $z_{k,r}$ do not change during the computation) yields the assertion.

In the *eighth step of the proof* we complete the proof by showing (36).

Thanks to the results of the steps 1 through 5 we know that the assumptions of Theorem 2 are satisfied. Application of Theorem 2 yields

$$
\begin{aligned}
&\mathbf{E}\left\{\varphi(Y \cdot sgn(f_n(X)))\right\} - \mathbf{E}\left\{\varphi(Y \cdot f_{\varphi^*}(X))\right\} \\
&\leq c_{106} \cdot \left( \frac{\log n}{n} + \mathbf{E}\left\{ \left| \sup_{\vartheta \in \bar{\mathbf{\Theta}}} \frac{1}{n} \sum_{i=1}^{n} \epsilon_i \cdot T_{\beta_n}(f_\vartheta(X_i)) \right| \right\} + \frac{C_n + 1}{\sqrt{N_n}} + \frac{D_n^2}{t_n} \right. \\
&\qquad \left. + \sup_{\vartheta \in \mathbf{\Theta}^*} \mathbf{E}\left\{\varphi(Y \cdot f_\vartheta(X))\right\} - \min_{f:\mathbb{R}^{d \cdot l} \to \bar{\mathbb{R}}} \mathbf{E}\left\{\varphi(Y \cdot f(X))\right\} \right).
\end{aligned}
$$

Plugging in the results of steps 6 and 7 and the values of $C_n$, $N_n$ and $D_n$ derived in steps 2, 4 and 5 yields the assertion. □

## References


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

[2] Andoni, A., Panigraphy, R., Valiant, G., and Zhang, L. (2014). Learning polynomials with neural networks. In *International Conference on Machine Learning*, pages 1908–1916.

[3] Barron, A. R. (1994). Approximation and estimation bounds for artificial neural networks. *Machine Learning* **14**, pp. 115-133.

[4] Bartlett, P., Harvey, N., Liaw, C., and Mehrabian, A. (2019). Nearly-tight VC-dimension bounds for piecewise linear neural networks. *Journal of Machine Learning Research* **20**, pp. 1-17.

[5] Bartlett, P. L., Montanari, A., and Rakhlin, A. (2021). Deep learning: a statistical viewpoint. Preprint, *arXiv: 2103.09177.*

[6] Bauer, B., and Kohler, M. (2019). On deep learning as a remedy for the curse of dimensionality in nonparametric regression. *Annals of Statistics* **4**, pp. 2261–2285.

[7] Birman, M. S., and Solomjak, M. Z. (1967). Piece-wise polynomial approximations of functions in the classes $W_p^\alpha$. *Mathematics of the USSR Sbornik* **73**, pp. 295-317.

[8] Braun, A., Kohler, M., Langer, S., and Walk, H. (2023). Convergence rates for shallow neural networks learned by gradient descent. Accepted for publication in *Bernoulli.* Preprint, *arXiv: 2107.09550.*

[9] Chizat, L., and Bach, F. (2018). On the global convergence of gradient descent for over-parameterized models using optimal transport. Preprint, *arXiv: 1805.09545.*

[10] Daniely, A. (2017). SGD learns the conjugate kernel class of the network. In *Advances in Neural Information Processing Systems*, pp. 2422–2430.

[11] Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv: 1810.04805.*

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

[13] Drews, S., and Kohler, M. (2022). On the universal consistency of an over-parametrized deep neural network estimate learned by gradient descent. Preprint.

[14] Drews, S., and Kohler, M. (2023). Analysis of the expected $L_2$ error of an over-parametrized deep neural network estimate learned by gradient descent without regularization. Preprint.

[15] Du, S., Lee, J., Li, H., Wang, L., und Zhai, X. (2019). Gradient descent finds global minima of deep neural networks. International Conference on Machine Learning, Preprint, *arXiv: 1811.03804.*

[16] Golowich, N., Rakhlin, A., and Shamir, O. (2019). Size-Independent sample complexity of neural networks. Preprint, *arXiv: 1712.06541.*

[17] Gonon, L. (2021). Random feature networks learn Black-Scholes type PDEs without curse of dimensionality. Preprint, *arXiv: 2106.08900.*

[18] Gurevych, I., Kohler, M., and Şahin, G. (2022). On the rate of convergence of a classifier based on a Transformer encoder. *IEEE Transactions Information Theory*, **68**, pp. 8139-8155. *arXiv: 2111.14574.*

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

[20] Hanin, B., and Nica, M. (2019). Finite depth and width corrections to the neural tangent kernel. *arXiv: 1909.05989.*

[21] Huang, G. B., Chen, L., and Siew, C.-K. (2006). Universal approximation using incremental constructive feedforward networks with random hidden nodes. *IEEE Transactions on Neural Networks* **17**, pp. 879-892.

[22] Imaizumi, M., and Fukamizu, K. (2018). Deep neural networks learn non-smooth functions effectively. *Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics (AISTATS 2019)*, Naha, Okinawa, Japan.

[23] Jacot, A., Gabriel, F., and Hongler, C. (2020). Neural Tangent Kernel: Convergence and Generalization in Neural Networks. *arXiv: 1806.07572v4.*

[24] 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.

[25] Kim, Y., Ohn, I.S, and Kim, D. (2019). Fast convergence rates of deep neural networks for classification. *arXiv: 1812.03599.*

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

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

[28] Kohler, M., and Krzyżak, A. (2022). Analysis of the rate of convergence of an over-parametrized deep neural network estimate learned by gradient descent. Preprint, *arXiv: 2210.01443.*

[29] Kohler, M., and Krzyżak, A. (2023). On the rate of convergence of an over-parametrized deep neural network regression estimate with ReLU activation function learned by gradient descent. Preprint.

[30] Kohler, M., and Langer, S. (2020a). Discussion of "Nonparametric regression using deep neural networks with ReLU activation function". *Annals of Statistics* **48**, pp. 1906-1910.

[31] Kohler, M., and Langer, S. (2020b). Statistical theory for image classification using deep convolutional neural networks with cross-entropy loss. Preprint, *arXiv:* 2011.13602.

[32] Kohler, M., and Langer, S. (2021). On the rate of convergence of fully connected deep neural network regression estimates using ReLU activation functions. *Annals of Statistics* **49**, pp. 2231-2249. Preprint, *arXiv: 1908.11133.*

[33] Langer, S. (2021). Approximating smooth functions by deep neural networks with sigmoid activation function. *Journal of Multivariate Analysis* **182**.

[34] Li, G., Gu, Y., and Ding, J. (2021). The rate of convergence of variation-constrained deep neural networks. Preprint, *arXiv: 2106.12068.*

[35] Liang, T., Rakhlin, A., and Sridharan, K. (2015). Learning with square loss: localization through offset Rademacher complexity. Preprint, *arXiv: 1502.06134.*

[36] Lin, S., and Zhang, J. (2019). Generalization bounds for convolutional neural networks. Preprint, *arXiv: 1910.01487.*

[37] Lu, J., Shen, Z., Yang, H., and Zhang, S. (2020). Deep network approximation for smooth functions. *arxiv: 2001.03040.*

[38] Mei, S., Montanari, A., and Nguyen, P.-M. (2018). A mean field view of the landscape of two-layer neural networks. In *Proceedings of the National Academy of Sciences*, **115**, pp. E7665-E7671.

[39] Nguyen, P.-M., and Pham, H. T. (2020). A rigorous framework for the mean field limit of multilayer neural networks. Preprint, *arXiv: 2001.1144.*

[40] Nitanda, A., and Suzuki, T. (2021). Optimal rates for averaged stochastic gradient descent under neural tangent kernel regime. *arXiv: 2006.12297.*

[41] Rahimi, A., and Recht, B. (2008a). Random features for large-scale kernel machines. In *Advances in Neural Information Procesing Systems*, pp. 1177-1184.

[42] Rahimi, A., and Recht, B. (2008b). Uniform approximation of function with random bases. In *2008 46th Annual Allerton Conference on Communication, Control, and Computing*, pp. 555-561, IEEE.

[43] Rahimi, A., and Recht, B. (2009). Weighted sums of random kitchen sinks: Replacing minimization with randomization in learning. In D. Koller, D. Schuurman, Y. Bengio, and L. Bottou, editors, *Advances in Neural Information Processing Systems*, Curran Associates, Inc. **21**, pp. 1313-1320.

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

[45] Suzuki, T. (2018). Adaptivity of deep ReLU network for learning in Besov and mixed smooth Besov spaces: optimal rate and curse of dimensionality. Preprint, *arXiv: 1810.08033.*

[46] Suzuki, T., and Nitanda, A. (2019). Deep learning is adaptive to intrinsic dimensionality of model smoothness in anisotropic Besov space. Preprint, *arXiv: 1910.12799.*

[47] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A., Kaiser, L., and Polosukhin,I. (2017). Attention is all you need. *arXiv: 1706.03762.*

[48] Wang, M., and Ma, C. (2022). Generalization error bounds for deep neural network trained by SGD. Preprint, *arXiv: 2206.03299v1.*

[49] Yarotsky, D. (2018). Optimal approximation of continuous functions by very deep ReLU networks. Preprint, *arXiv: 1802.03620.*

[50] Yarotsky, D., and Zhevnerchuk, A. (2019). The phase diagram of approximation rates for deep neural networks. Preprint, *arXiv: 1906.09477.*

[51] Yehudai, G., and Shamir, O. (2022). On the power and limitations of random features for understanding neural networks. Preprint, *arXiv: 1904.00687.*

[52] Zhang, T. (2004). Statistical behavior and consistency of classification methods based on convex risk minimization. *Annals of Statistics*, **32**, pp. 56 - 134.

[53] Zou, D., Cao, Y., Zhou, D., und Gu, Q. (2018). Stochastic gradient descent optimizes over-parameterized deep ReLU networks. Preprint, *arXiv: 1811.08888.*

# SUPPLEMENTARY MATERIAL

## .1 A result for gradient descent

**Lemma 1** *Let $d_1, d_2 \in \mathbb{N}$, let $D_n \geq 0$, let $A \subset \mathbb{R}^{d_1}$ and $B \subseteq \mathbb{R}^{d_2}$ be closed and convex, and let $F : \mathbb{R}^{d_1} \times \mathbb{R}^{d_2} \to \mathbb{R}_+$ be a function such that*

$$u \mapsto F(u,v) \quad \textit{is differentiable and convex for all } v \in \mathbb{R}^{d_2}$$

*and*

$$\|(\nabla_u F)(u,v)\| \leq D_n \tag{37}$$

*for all $(u,v) \in A \times B$. Choose $(u_0, v_0) \in A \times B$, let $v_1, \dots, v_{t_n} \in B$ and set*

$$u_{t+1} = Proj_A \left( u_t - \lambda \cdot (\nabla_u F)(u_t, v_t) \right) \quad (t = 0, \dots, t_n - 1),$$

*where*

$$\lambda = \frac{1}{t_n}.$$

*Let $u^* \in A$. Then it holds:*

$$\min_{t=0,\dots,t_n} F(u_t, v_t) \leq F(u^*, v_0) + \frac{1}{t_n} \sum_{t=1}^{t_n} |F(u^*, v_t) - F(u^*, v_0)| + \frac{\|u^* - u_0\|^2}{2} + \frac{D_n^2}{2 \cdot t_n}.$$

**Proof.** The result follows from the proof of Lemma 1 in Kohler and Krzyżak (2023). For the sake of completeness we give nevertheless a complete proof here.

In the *first step of the proof* we show

$$\frac{1}{t_n} \sum_{t=0}^{t_n-1} F(u_t, v_t) \leq \frac{1}{t_n} \sum_{t=0}^{t_n-1} F(u^*, v_t) + \frac{\|u^* - u_0\|^2}{2} + \frac{1}{2 \cdot t_n^2} \sum_{t=0}^{t_n-1} \|(\nabla_u F)(u_t, v_t)\|^2. \tag{38}$$

By convexity of $u \mapsto F(u, v_t)$ and because of $u^* \in A$ we have

$$\begin{aligned}
& F(u_t, v_t) - F(u^*, v_t) \\
& \leq < (\nabla_u F)(u_t, v_t), u_t - u^* > \\
& = \frac{1}{2 \cdot \lambda} \cdot 2 \cdot < \lambda \cdot (\nabla_u F)(u_t, v_t), u_t - u^* > \\
& = \frac{1}{2 \cdot \lambda} \cdot \left( -\|u_t - u^* - \lambda \cdot (\nabla_u F)(u_t, v_t)\|^2 + \|u_t - u^*\|^2 + \|\lambda \cdot (\nabla_u F)(u_t, v_t)\|^2 \right) \\
& \leq \frac{1}{2 \cdot \lambda} \cdot \left( -\|Proj_A(u_t - \lambda \cdot (\nabla_u F)(u_t, v_t)) - u^*\|^2 + \|u_t - u^*\|^2 + \lambda^2 \cdot \|(\nabla_u F)(u_t, v_t)\|^2 \right) \\
& = \frac{1}{2 \cdot \lambda} \cdot \left( \|u_t - u^*\|^2 - \|u_{t+1} - u^*\|^2 + \lambda^2 \cdot \|(\nabla_u F)(u_t, v_t)\|^2 \right).
\end{aligned}$$

This implies

$$\frac{1}{t_n} \sum_{t=0}^{t_n-1} F(u_t, v_t) - \frac{1}{t_n} \sum_{t=0}^{t_n-1} F(u^*, v_t)$$

$$
\begin{aligned}
&= \frac{1}{t_n} \sum_{t=0}^{t_n-1} \left(F(u_t, v_t) - F(u^*, v_t)\right) \\
&\leq \frac{1}{t_n} \sum_{t=0}^{t_n-1} \frac{1}{2 \cdot \lambda} \cdot \left(\|u_t - u^*\|^2 - \|u_{t+1} - u^*\|^2\right) + \frac{1}{t_n} \sum_{t=0}^{t_n-1} \frac{\lambda}{2} \cdot \|(\nabla_u F)(u_t, v_t)\|^2 \\
&= \frac{1}{2} \cdot \sum_{t=0}^{t_n-1} \left(\|u_t - u^*\|^2 - \|u_{t+1} - u^*\|^2\right) + \frac{1}{2 \cdot t_n^2} \sum_{t=0}^{t_n-1} \|(\nabla_u F)(u_t, v_t)\|^2 \\
&\leq \frac{\|u_0 - u^*\|^2}{2} + \frac{1}{2 \cdot t_n^2} \sum_{t=0}^{t_n-1} \|(\nabla_u F)(u_t, v_t)\|^2.
\end{aligned}
$$

In the *second step of the proof* we show the assertion.
Using the result of step 1 we get

$$
\begin{aligned}
&\min_{t=0,\dots,t_n} F(u_t, v_t) \\
&\leq \frac{1}{t_n} \sum_{t=0}^{t_n-1} F(u_t, v_t) \\
&\leq \frac{1}{t_n} \sum_{t=0}^{t_n-1} F(u^*, v_t) + \frac{\|u^* - u_0\|^2}{2} + \frac{1}{2 \cdot t_n^2} \sum_{t=0}^{t_n-1} \|(\nabla_u F)(u_t, v_t)\|^2 \\
&\leq F(u^*, v_0) + \frac{1}{t_n} \sum_{t=0}^{t_n-1} |F(u^*, v_t) - F(u^*, v_0)| + \frac{\|u^* - u_0\|^2}{2} \\
&\quad + \frac{1}{2 \cdot t_n^2} \sum_{t=0}^{t_n-1} \|(\nabla_u F)(u_t, v_t)\|^2.
\end{aligned}
$$

By (37) we get

$$
\frac{1}{2 \cdot t_n^2} \sum_{t=0}^{t_n-1} \|(\nabla_u F)(u_t, v_t)\|^2 \leq \frac{1}{2 \cdot t_n^2} \sum_{t=0}^{t_n-1} D_n^2 = \frac{D_n^2}{2 \cdot t_n}.
$$

Summarizing the above results, the proof is complete.

□

## .2 An auxiliary result

**Lemma 2** *Let $\varphi$ be the logistic loss. Let $(X, Y), (X_1, Y_1), \dots, (X_n, Y_n)$ and $\eta^*$, $\mathcal{D}_n, f_n$ and $\eta_n$ as in Sections 1 and 2, and set*

$$
f_{\varphi^*} = \arg \min_{f:\mathbb{R}^{d \cdot l} \to \bar{\mathbb{R}}} \mathbf{E}\left\{\varphi(Y \cdot f(X))\right\}.
$$

**a)** *Then*

$$
\mathbf{P}\left\{Y \neq \eta_n(X) | \mathcal{D}_n\right\} - \mathbf{P}\left\{Y \neq \eta^*(X)\right\}
$$

$$\leq \frac{1}{\sqrt{2}} \cdot \left(\mathbf{E}\left\{\varphi(Y \cdot f_n(X)) | \mathcal{D}_n\right\} - \mathbf{E}\left\{\varphi(Y \cdot f_{\varphi^*}(X))\right\}\right)^{1/2}$$

*holds.*
**b)** *Then*

$$\begin{aligned}&\mathbf{P}\left\{Y \neq \eta_n(X) | \mathcal{D}_n\right\} - \mathbf{P}\left\{Y \neq \eta^*(X)\right\} \\ &\leq 2 \cdot \left(\mathbf{E}\left\{\varphi(Y \cdot f_n(X)) | \mathcal{D}_n\right\} - \mathbf{E}\left\{\varphi(Y \cdot f_{\varphi^*}(X))\right\}\right) + 4 \cdot \mathbf{E}\left\{\varphi(Y \cdot f_{\varphi^*}(X))\right\}.\end{aligned}$$

*holds.*
**c)** *Assume that*

$$\mathbf{P}\left\{|f_{\varphi^*}(X)| > \tilde{F}_n\right\} \geq 1 - e^{-\tilde{F}_n}$$

*for a given sequence* $\{\tilde{F}_n\}_{n \in \mathbb{N}}$ *with* $\tilde{F}_n \to \infty$. *Then*

$$\mathbf{E}\left\{\varphi(Y \cdot f_{\varphi^*}(X))\right\} \leq c_{15} \cdot \tilde{F}_n \cdot e^{-\tilde{F}_n}$$

*holds.*

**Proof. a)** This result follows from Theorem 2.1 in Zhang (2004), where we choose $s = 2$ and $c = 2^{-1/2}$.
**b)** This result follows from Lemma 1 b) in Kohler and Langer (2020b).
**c)** This result follows from Lemma 3 in Kim, Ohn and Kim (2019). □

## .3 A bound on the gradient

In the proof of Theorem 1 we will apply Theorem 2. For this we need the following bound on the gradient (with respect to the outer weights) of $F_n$.

**Lemma 3** *Let* $F_n$ *be defined by (18). Then we have*

$$\|\nabla_{(w_k)_{k=1,\dots,K_n}} F_n(\mathbf{w})\| \leq \sqrt{K_n} \cdot \beta_n.$$

**Proof.** For $k \in \{1, \dots, K_n\}$ we have

$$\frac{\partial F_n(\mathbf{w})}{\partial w_k} = \frac{1}{n} \sum_{i=1}^{n} \varphi'(Y_i \cdot f_{\mathbf{w}}(X_i)) \cdot Y_i \cdot T_{\beta_n}(f_{\mathbf{w}_k, \mathbf{v}_k}(X_i)).$$

Because of $|\varphi'(z)| \leq 1$ we can conclude

$$\left|\frac{\partial F_n(\mathbf{w})}{\partial w_k}\right| \leq \beta_n$$

and

$$\|\nabla_{(w_k)_{k=1,\dots,K_n}} F_n(\mathbf{w})\|^2 = \sum_{k=1}^{K_n} \left|\frac{\partial F_n(\mathbf{w})}{\partial w_k}\right|^2 \leq K_n \cdot \beta_n^2.$$

□

## .4 Generalization error

**Lemma 4** *Let $d_{model} = h \cdot I$ and let $\mathcal{F}$ be the set of all functions*

$$(x_1, \dots, x_l) \mapsto z_{1,N,1}^{(d+l+2)},$$

*where $z_{1,N,1}$ is defined in Section 2 depending on*

$$(W_{query,1,r,s}, W_{key,1,r,s}, W_{value,1,r,s})_{r \in \{1,\dots,N\}, s \in \{1,\dots,h\}} \tag{39}$$

*and on*

$$(W_{1,r,1}, b_{1,r,1}, W_{1,r,2}, b_{1,r,2})_{r \in \{1,\dots,N\}} \tag{40}$$

*and where the total number of nonzero components in each row in all the matrices in (39) is bounded by $\tau \in \mathbb{N}$ and where all matrices $W_{1,r,1}$ and $W_{1,r,2}$ in (40) have the property that in each row in $W_{1,r,1}$ and in each column in $W_{1,r,2}$ there are at most $\tau$ nonzero entries. Let $\mathcal{G}$ be the set of all (shallow) feedforward neural networks $g : \mathbb{R} \to \mathbb{R}$ with one hidden layer and $J_n$ hidden neurons and ReLU activation function. Assume*

$$\max\{N, d_{key}, d_v, l\} \leq c_{16} \quad \text{and} \quad \max\{J_n, h, I, d_{ff}\} \leq c_{17} \cdot n^{c_{17}}.$$

*Let $A \geq 1$, let $X_1, \dots, X_n$ be independent and identically distributed $[-A, A]^{d \cdot l}$-valued random vectors and let $\epsilon_1, \dots, \epsilon_n$ be independent Rademacher random variables, which are independent from $X_1, \dots, X_n$. Then we have*

$$\mathbf{E}\left\{ \left| \sup_{f \in \mathcal{G} \circ \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \epsilon_i \cdot T_{\beta_n}(f(X_i)) \right| \right\} \leq c_{18} \cdot \frac{\sqrt{\max\{h \cdot I, d_{ff}, J_n\}} \cdot (\log n)^2}{\sqrt{n}}.$$

In order to prove Lemma 4 we need the following bound on the covering number.

**Lemma 5** *Define $\mathcal{F}$ and $\mathcal{G}$ as in Lemma 4. Let $\beta \geq 0$ and let $T_\beta \mathcal{G} \circ \mathcal{F}$ be the set of all functions $g \circ f$ truncated on height $\beta$ and $-\beta$ where $g \in \mathcal{G}$ and $f \in \mathcal{F}$. Then we have for any $0 < \epsilon < \beta/2$*

$$\sup_{z_1^n \in (\mathbb{R}^{d \cdot l})^n} \log \mathcal{M}_1(\epsilon, T_\beta \mathcal{G} \circ \mathcal{F}, z_1^n)$$
$$\leq c_{19} \cdot \tau \cdot \max\{h \cdot I, d_{ff}, J_n\} \cdot N^3 \cdot \log(\max\{J_n, N, h, d_{ff}, I, d_{key}, d_v, l, 2\}) \log\left(\frac{\beta}{\epsilon}\right).$$

In order to prove Lemma 5 we will first show the following bound on the VC-dimension of subsets of $\mathcal{F}$, where the nonzero components appear only at fixed positions.

**Lemma 6** *Let $\mathcal{F}$ be the set of all functions*

$$(x_1, \dots, x_l) \mapsto z_{1,N,1}^{(d+l+2)},$$

*where $z_{1,N,1}$ is defined in Section 2 depending on (39) and (40) and where in all matrics in (39) there are in each row at most $\tau$ fixed components where the entries are allowed* *to be nonzero, and where all matrices $W_{1,r,1}$ and $W_{1,r,2}$ in (40) have the property that in each row in $W_{1,r,1}$ and in each column in $W_{1,r,2}$ there are at most $\tau$ fixed components (depending on $r$) where the entries are allowed to be nonzero. Let $\mathcal{G}$ be defined as in Lemma 4. Then we have*

$$V_{(\mathcal{G}\circ\mathcal{F})^+} \leq c_{20} \cdot \tau \cdot \max\{h \cdot I, d_{ff}, J_n\} \cdot N^3 \cdot \log(\max\{J_n, N, h, d_{ff}, d_{key}, d_v, l, 2\}).$$

The proof of Lemma 6 is a modification of the proof of Lemma 9 in Gurevych, Kohler and Sarin (2022), which in turn is based on the proof of Theorem 6 in Bartlett et al. (1999). In the proof of Lemma 6 we will need the following two auxiliary results.

**Lemma 7** *Suppose $W \leq m$ and let $f_1, ..., f_m$ be polynomials of degree at most $D$ in $W$ variables. Define*

$$K := |\{(sgn(f_1(\mathbf{a})), \dots, sgn(f_m(\mathbf{a}))) : \mathbf{a} \in \mathbb{R}^W\}|.$$

*Then we have*

$$K \leq 2 \cdot \left(\frac{2 \cdot e \cdot m \cdot D}{W}\right)^W.$$

**Proof.** See Theorem 8.3 in Anthony and Bartlett (1999). □

**Lemma 8** *Suppose that $2^m \leq 2^L \cdot (m \cdot R/w)^w$ for some $R \geq 16$ and $m \geq w \geq L \geq 0$. Then,*

$$m \leq L + w \cdot \log_2(2 \cdot R \cdot \log_2(R)).$$

**Proof.** See Lemma 16 in Bartlett et al. (2019). □

**Proof of Lemma 6**. Let $\mathcal{H}$ be the set of all functions $h$ defined by

$$h : \mathbb{R}^{d \cdot l} \times \mathbb{R} \to \mathbb{R}, \quad h(x, y) = g(x) - y$$

for some $g \in \mathcal{G} \circ \mathcal{F}$. Let $(x_1, y_1), \dots (x_m, y_m) \in \mathbb{R}^{d \cdot l} \times \mathbb{R}$ be such that

$$|\{(sgn(h(x_1, y_1)), \dots, sgn(h(x_m, y_m))) \, : \, h \in \mathcal{H}\}| = 2^m. \tag{41}$$

It suffices to show

$$m \leq c_{20} \cdot \tau \cdot \max\{h \cdot I, d_{ff}, J_n\} \cdot N^3 \cdot \log(\max\{J_n, N, h, d_{ff}, d_{key}, d_v, l, 2\}). \tag{42}$$

To show this we partition $\mathcal{G} \circ \mathcal{F}$ in subsets such that for each subset all

$$g \circ f(x_i) \quad (i = 1, \dots, m)$$

are polynomials of some fixed degree and use Lemma 7 in order to derive an upper bound on the left-hand side of (41). This upper bound will depend polynomially on $m$ which will enable us to conclude (42) by an application of Lemma 8.

Let

$$\bar{\theta} = \Big( (W_{query,r,s}, W_{key,r,s}, W_{value,r,s})_{r \in \{1,\dots,N\}, s \in \{1,\dots,h\}},$$

$$(W_{r,1}, b_{r,1}, W_{r,2}, b_{r,2})_{r \in \{1,\dots,N\}}, (v_r^{(1)})_{r \in \{1,\dots,J_n\}}, (v_{r,s}^{(0)})_{r \in \{1,\dots,J_n\}, s \in \{0,1\}} \Big)$$

be the parameters which determine a function in $\mathcal{G} \circ \mathcal{F}$. By assumption, each function in $\mathcal{G} \circ \mathcal{F}$ can be also described by such a parameter vector. Here only

$$\bar{L}_n = N \cdot h \cdot (2 \cdot d_{key} + d_v) \cdot \tau + N \cdot (d_{ff} \cdot (\tau + 1) + h \cdot I \cdot (\tau + 1)) + 3 \cdot J_n$$

components of the matrices and vectors occuring in the parameter vector are allowed to be nonzero and the positions where these nonzero parameters can occur are fixed. Denote the vector in $\mathbb{R}^{\bar{L}_n}$ which contains all values of these possible nonzero parameters by $\theta$. Then we can write

$$\mathcal{G} \circ \mathcal{F} = \{ g(\cdot, \theta) : \mathbb{R}^{d \cdot l} \to \mathbb{R} \, : \, \theta \in \mathbb{R}^{\bar{L}_n} \}.$$

In the sequel we construct a partition $\mathcal{P}_{N+1}$ of $\mathbb{R}^{\bar{L}_n}$ such that for all $S \in \mathcal{P}_{N+1}$ we have that

$$g(x_1, \theta), \dots, g(x_m, \theta)$$

(considered as functions of $\theta$) are polynomials of degree at most $8^N + 2$ for $\theta \in S$.

In order to construct this partition we construct first recursively partitions $\mathcal{P}_0$, ..., $\mathcal{P}_N$ of $\mathbb{R}^{\bar{L}_n}$ such that for each $r \in \{1, \dots, N\}$ and all $S \in \mathcal{P}_r$ all components in

$$z_r = z_r(x) \quad (x \in \{x_1, \dots, x_m\})$$

(considered as a function of $\theta$) are polynomials of degree at most $8^r$ in $\theta$ for $\theta \in S$.

Since all components of $z_0$ are constant as functions of $\theta$ this holds for $r = 0$ if we set $\mathcal{P}_0 = \{\mathbb{R}^{\bar{L}_n}\}$.

Let $r \in \{1, \dots, N\}$ and assume that for all $S \in \mathcal{P}_{r-1}$ all components in

$$z_{r-1}(x) \quad (x \in \{x_1, \dots, x_m\})$$

(considered as a function of $\theta$) are polynomials of degree at most $8^{r-1}$ in $\theta$ for $\theta \in S$. Then all components in

$$q_{r-1,s,i}(x), k_{r-1,s,i}(x) \quad \text{and} \quad v_{r-1,s,i}(x) \quad (x \in \{x_1, \dots, x_m\})$$

are on each set $S \in \mathcal{P}_{r-1}$ polynomials of degree at most $8^{r-1} + 1$. Consequently, for each $S \in \mathcal{P}_{r-1}$ each value

$$< q_{r-1,s,i}(x), k_{r-1,s,j}(x) > \quad (x \in \{x_1, \dots, x_m\})$$

is (considered as a function of $\theta$) a polynomial of degree at most $2 \cdot 8^{r-1} + 2$ for $\theta \in S$. Application of Lemma 7 yields that

$$< q_{r-1,s,i}(x), k_{r-1,s,j_1}(x) > - < q_{r-1,s,i}(x), k_{r-1,s,j_2}(x) >$$

$(s \in \{1, \dots, h\}, i, j_1, j_2 \in \{1, \dots, l\}, x \in \{x_1, \dots, x_m\})$ has at most

$$\Delta = 2 \cdot \left( \frac{2 \cdot e \cdot h \cdot l^3 \cdot m \cdot (2 \cdot 8^{r-1} + 2)}{\bar{L}_n} \right)^{\bar{L}_n}$$

different sign patterns. If we partition each set in $\mathcal{P}_{r-1}$ according to these sign patterns in $\Delta$ subsets, then on each set in the new partition all components in

$$v_{r-1,s,\hat{j}_{r-1,s,i}}(x) \cdot < q_{r-1,s,i}(x), k_{r-1,s,\hat{j}_{r-1,s,i}}(x) > \quad (x \in \{x_1, \dots, x_m\})$$

are polynomials of degree at most $3 \cdot 8^{r-1} + 3$ (since on each such set

$$< q_{r-1,s,i}(x), k_{r-1,s,\hat{j}_{r-1,s,i}}(x) >$$

is equal to one of the $< q_{r-1,s,i}(x), k_{r-1,s,j}(x) >$). On each set within this partition every component of the $\mathbb{R}^{d_{ff}}$-valued vectors

$$W_{r,1} \cdot y_{r,s}(x) + b_{r,1} \quad (s = 1, \dots, h, x \in \{x_1, \dots, x_m\})$$

is (considered as a function of $\theta$) a polynomial of degree at most $3 \cdot 8^{r-1} + 4$.

By another application of Lemma 7 we can refine each set in this partition into

$$2 \cdot \left( \frac{2 \cdot e \cdot h \cdot d_{ff} \cdot m \cdot (3 \cdot 8^{r-1} + 4)}{\bar{L}_n} \right)^{\bar{L}_n}$$

sets such that all components in

$$W_{r,1} \cdot y_{r,s}(x) + b_{r,1} \quad (x \in \{x_1, \dots, x_m\}) \tag{43}$$

have the same sign patterns within the refined partition. We call this refined partition $\mathcal{P}_r$. Since on each set of $\mathcal{P}_r$ the sign of all components in (43) does not change we can conclude that all components in

$$\sigma(W_{r,1} \cdot y_{r,s}(x) + b_{r,1}) \quad (x \in \{x_1, \dots, x_m\}) \tag{44}$$

are either equal to zero or they are equal to a polynomial of degree at most $3 \cdot 8^{r-1} + 4$. Consequently we have that on each set in $\mathcal{P}_r$ all components of

$$z_r(x) \quad (x \in \{x_1, \dots, x_m\})$$

are equal to a polynomial of degree at most $3 \cdot 8^{r-1} + 5 \leq 8^r$.

The partition $\mathcal{P}_N$ satisfies

$$|\mathcal{P}_N| = \prod_{r=1}^{N} \frac{|\mathcal{P}_r|}{|\mathcal{P}_{r-1}|} \leq \prod_{r=1}^{N} 2 \cdot \left( \frac{2 \cdot e \cdot h \cdot l^3 \cdot m \cdot 8^r}{\bar{L}_n} \right)^{\bar{L}_n} \cdot 2 \cdot \left( \frac{2 \cdot e \cdot h \cdot d_{ff} \cdot m \cdot 8^r}{\bar{L}_n} \right)^{\bar{L}_n}.$$

Next we construct a partition $\mathcal{P}_{N+1}$ of $\mathbb{R}^{\bar{L}_n}$ such that for all $S \in \mathcal{P}_{N+1}$

$$g(z_{N,1}^{(d+l+2)}(x)) = \sum_{j=1}^{J_n} v_j^{(1)} \cdot \sigma \left( v_{j,1}^{(0)} \cdot z_{N,1}^{(d+l+2)}(x) + v_{j,0}^{(0)} \right) \quad (x \in \{x_1, \dots, x_m\})$$

(considered as a function of $\theta$) is a polynomial of degree at most $8^N + 2$ for $\vartheta \in S$.

For all $S \in \mathcal{P}_N$ all components in

$$\left( v_{j,1}^{(0)} \cdot z_{N,1}^{(d+l+2)}(x) + v_{j,0}^{(0)} \right)_{j=1,\dots,J_n} \quad (x \in \{x_1, \dots, x_m\})$$

(considered as a function of $\theta$) are polynomials of degree at most $8^N + 1$. Application of Lemma 7 implies

$$v_{j,1}^{(0)} \cdot z_{N,1}^{(d+l+2)}(x) + v_{j,0}^{(0)} \quad (j \in \{1, \dots, J_n\}, x \in \{x_1, \dots, x_m\})$$

has at most

$$\Delta = 2 \cdot \left( \frac{2 \cdot e \cdot J_n \cdot m \cdot (8^N + 1)}{\bar{L}_n} \right)^{\bar{L}_n}$$

different sign patterns. If we partition in each set in $\mathcal{P}_N$ according to these sign patterns in $\Delta$ subsets, then on each set in the new partition $\mathcal{P}_{N+1}$ all components in

$$\sigma \left( v_{j,1}^{(0)} \cdot z_{N,1}^{(d+l+2)}(x) + v_{j,0}^{(0)} \right) \quad (j \in \{1, \dots, J_n\}, x \in \{x_1, \dots, x_m\})$$

are polynomials of degree at most $8^N + 1$ (since after the application of $\sigma(z) = \max\{z, 0\}$ the component is either equal to zero on the set or equal to the argument of $\sigma$). Consequently on each set in $\mathcal{P}_{N+1}$

$$g(x, \theta) = \sum_{j=1}^{J_n} v_j^{(1)}(x) \cdot \sigma \left( v_{j,1}^{(0)} \cdot (z_{N,1}^{(d+l+2)} + v_{j,0}^{(0)} \right) \quad (x \in \{x_1, \dots, x_m\})$$

(considered as a function of $\theta$) is a polynomial of degree $8^N + 2$.

The partition $\mathcal{P}_{N+1}$ satisfies

$$\begin{aligned} |\mathcal{P}_{N+1}| &= \frac{|\mathcal{P}_{N+1}|}{|\mathcal{P}_N|} \cdot |\mathcal{P}_N| \\ &\leq 2 \cdot \left( \frac{2 \cdot e \cdot J_n \cdot m \cdot (8^N + 1)}{\bar{L}_n} \right)^{\bar{L}_n} \cdot \left( \prod_{r=1}^{N} 2 \cdot \left( \frac{2 \cdot e \cdot h \cdot l^3 \cdot m \cdot 8^r}{\bar{L}_n} \right)^{\bar{L}_n} \right. \end{aligned}$$

$$\cdot 2 \cdot \left( \frac{2 \cdot e \cdot h \cdot d_{ff} \cdot m \cdot 8^r}{\bar{L}_n} \right)^{\bar{L}_n} \Bigg)$$

and has the property that for all $S \in \mathcal{P}_{N+1}$ and for all $(x, y) \in \{(x_1, y_1), \ldots, (x_m, y_m)\}$

$$g(x) = g(x, \theta) \quad \text{and} \quad h(x, y) = g(x, \theta) - y$$

(considered as a function of $\theta$) are polynomials of degree at most $8^N + 2$ in $\theta$ for $\theta \in S$.

Using

$$\begin{aligned}
&|\{(sgn(h(x_1, y_1)), \ldots, sgn(h(x_m, y_m))) \, : \, h \in \mathcal{H}\}| \\
&\leq \sum_{S \in \mathcal{P}_{N+1}} |\{(sgn(g(x_1, \theta) - y_1), \ldots, sgn(g(x_m, \theta) - y_m)) \, : \, \theta \in S\}|
\end{aligned}$$

we can apply one more time Lemma 7 to conclude

$$\begin{aligned}
&2^m \\
&= |\{(sgn(h(x_1, y_1)), \ldots, sgn(h(x_m, y_m))) \, : \, h \in \mathcal{H}\}| \\
&\leq |\mathcal{P}_{N+1}| \cdot 2 \cdot \left( \frac{2 \cdot e \cdot m \cdot (8^N + 2)}{\bar{L}_n} \right)^{\bar{L}_n} \\
&\leq 2 \cdot \left( \frac{2 \cdot e \cdot m \cdot (8^N + 2)}{\bar{L}_n} \right)^{\bar{L}_n} \cdot \left( \prod_{r=1}^{N} 2 \cdot \left( \frac{2 \cdot e \cdot h \cdot l^3 \cdot m \cdot 8^r}{\bar{L}_n} \right)^{\bar{L}_n} \right. \\
&\qquad \left. \cdot 2 \cdot \left( \frac{2 \cdot e \cdot h \cdot d_{ff} \cdot m \cdot 8^r}{\bar{L}_n} \right)^{\bar{L}_n} \right) \cdot 2 \cdot \left( \frac{2 \cdot e \cdot J_n \cdot m \cdot (8^N + 1)}{\bar{L}_n} \right)^{\bar{L}_n} \\
&\leq 2^{2 \cdot N + 2} \cdot \\
&\left( \frac{m \cdot 2e \cdot (2N + 2) \cdot \max\{J_n, h\} \cdot (\max\{l, d_{ff}\})^3 \cdot (8^N + 2)}{(2N + 2) \cdot \bar{L}_n} \right)^{(2N+2) \cdot \bar{L}_n}.
\end{aligned}$$

Assume $m \geq (2N+2) \cdot (N \cdot h \cdot (2 \cdot d_{key} + d_v) \cdot \tau + N \cdot (d_{ff} \cdot (\tau+1) + h \cdot I \cdot (\tau+1)) + 3 \cdot J_n)$. Application of Lemma 8 with $L = 2 \cdot N + 2$, $R = 2e \cdot (2N+2) \cdot \max\{J_n, h\} \cdot (\max\{l, d_{ff}\})^3 \cdot (8^N + 2)$ and $w = (2N + 2) \cdot \bar{L}_n$ yields

$$\begin{aligned}
m &\leq (2 \cdot N + 2) + (2N + 2) \cdot \bar{L}_n \cdot \log_2(2 \cdot R \cdot \log_2(R)) \\
&\leq c_{21} \cdot \tau \cdot \max\{h \cdot I, d_{ff}, J_n\} \cdot N^3 \cdot \log(\max\{J_n, N, h, d_{ff}, l, 2\}),
\end{aligned}$$

which implies (42). □

**Proof of Lemma 5.** The functions in the function set $T_\beta \mathcal{G} \circ \mathcal{F}$ depend on at most

$$\lceil c_{22} \cdot (N \cdot h^2 + J_n) \cdot I \cdot \max\{d_k, d_{ff}, d_v\} \rceil$$

many parameters, and of these parameters at most

$$\bar{L}_n = N \cdot h \cdot (2 \cdot d_{key} + d_v) \cdot \tau + N \cdot (d_{ff} \cdot (\tau + 1) + h \cdot I \cdot (\tau + 1)) + 3 \cdot J_n$$

are allowed to be nonzero. We have

$$\binom{\lceil c_{22} \cdot (N \cdot h^2 + J_n) \cdot I \cdot \max\{d_{key}, d_{ff}, d_v\} \rceil}{\bar{L}_n}$$
$$\leq \left(\lceil c_{22} \cdot (N \cdot h^2 + J_n) \cdot I \cdot \max\{d_{key}, d_{ff}, d_v\} \rceil\right)^{\bar{L}_n}$$

many possibilities to choose these positions. If we fix these positions, we get one function space $\mathcal{G} \circ \mathcal{F}$ for which we can bound its VC dimension by Lemma 6. Using Lemma 6, $V_{(T_\beta \mathcal{G} \circ \mathcal{F})^+} \leq V_{(\mathcal{G} \circ \mathcal{F})^+}$, and Theorem 9.4 in Györfi et al. (2002) we get

$$\begin{aligned} \mathcal{M}_1\left(\epsilon, T_\beta \mathcal{G} \circ \mathcal{F}, \mathbf{x}_1^n\right) &\leq 3 \cdot \left(\frac{4e \cdot \beta}{\epsilon} \cdot \log \frac{6e \cdot \beta}{\epsilon}\right)^{V_{(T_\beta \mathcal{G} \circ \mathcal{F})^+}} \\ &\leq 3 \cdot \left(\frac{6e \cdot \beta}{\epsilon}\right)^{2 \cdot c_{20} \cdot \tau \cdot \max\{h \cdot I, d_{ff}, J_n\} \cdot N^3 \cdot \log(\max\{J_n, N, h, d_{ff}, d_{key}, d_v, l, 2\})}. \end{aligned}$$

From this we conclude

$$\begin{aligned} &\sup_{z_1^n \in (\mathbb{R}^{d \cdot l})^n)} \log \mathcal{M}_1(\epsilon, T_\beta \mathcal{G} \circ \mathcal{F}, z_1^n) \\ &\leq \bar{L}_n \cdot \log\left(\lceil c_{22} \cdot (N \cdot h^2 + J_n) \cdot I \cdot \max\{d_{key}, d_{ff}, d_v\} \rceil\right) \\ &\quad + 2 \cdot c_{20} \cdot \tau \cdot \max\{h \cdot I, d_{ff}, J_n\} \cdot N^3 \cdot \log(\max\{J_n, N, h, d_{ff}, d_{key}, d_v, l, 2\}) \cdot \log\left(\frac{\beta}{\epsilon}\right) \\ &\leq c_{23} \cdot \tau \cdot \max\{h \cdot I, d_{ff}, J_n\} \cdot N^3 \cdot \log(\max\{J_n, N, h, I, d_{ff}, d_{key}, d_v, 2\}) \cdot \log\left(\frac{\beta}{\epsilon}\right). \end{aligned}$$

□

**Proof of Lemma 4.** For $\delta_n > 0$ we have

$$\begin{aligned} &\mathbf{E}\left\{\sup_{f \in \mathcal{G} \circ \mathcal{F}} \left|\frac{1}{n} \sum_{i=1}^n \epsilon_i \cdot (T_{\beta_n}(f(X_i))\right|\right\} \\ &= \int_0^\infty \mathbf{P}\left\{\sup_{f \in \mathcal{G} \circ \mathcal{F}} \left|\frac{1}{n} \sum_{i=1}^n \epsilon_i \cdot (T_{\beta_n}(f(X_i))\right| > t\right\} dt \\ &\leq \delta_n + \int_{\delta_n}^\infty \mathbf{P}\left\{\sup_{f \in \mathcal{G} \circ \mathcal{F}} \left|\frac{1}{n} \sum_{i=1}^n \epsilon_i \cdot (T_{\beta_n}(f(X_i))\right| > t\right\} dt. \end{aligned}$$

Using a standard covering argument from empirical process theory we see that for any $\beta_n \geq t \geq \delta_n$ we have

$$\begin{aligned} &\mathbf{P}\left\{\sup_{f \in \mathcal{G} \circ \mathcal{F}} \left|\frac{1}{n} \sum_{i=1}^n \epsilon_i \cdot (T_{\beta_n}(f(X_i))\right| > t\right\} \\ &\leq \sup_{x_1^n \in (\mathbb{R}^{d \cdot l})^n} \mathcal{M}_1\left(\frac{\delta_n}{2}, \{T_{\beta_n} f : f \in \mathcal{G} \circ \mathcal{F}\}, x_1^n\right) \end{aligned}$$

$$\cdot \sup_{f \in \mathcal{G} \circ \mathcal{F}} \mathbf{P} \left\{ \left| \frac{1}{n} \sum_{i=1}^{n} \epsilon_i \cdot (T_{\beta_n}(f(X_i)) \right| > \frac{t}{2} \right\}.$$

By Lemma 5 we know

$$\sup_{x_1^n \in (\mathbb{R}^{d \cdot l})^n} \mathcal{M}_1 \left( \frac{\delta_n}{2}, \{T_{\beta_n} f : f \in \mathcal{G} \circ \mathcal{F}\}, x_1^n \right) \leq c_{24} \cdot \left( \frac{c_{25} \cdot \beta_n}{\delta_n} \right)^{c_{26} \cdot \max\{h \cdot I \cdot d_{ff}, J_n\} \cdot \log(n)}.$$

By the inequality of Hoeffding (cf., e.g., Lemma A.3 in Györfi et al. (2002)) and

$$|T_{\beta_n}(f(x))| \leq \beta_n \quad (x \in \mathbb{R}^d)$$

we have for any $f \in \mathcal{G} \circ \mathcal{F}$

$$\mathbf{P} \left\{ \left| \frac{1}{n} \sum_{i=1}^{n} \epsilon_i \cdot T_{\beta_n}(f(X_i)) \right| > t \right\} \leq 2 \cdot \exp \left( - \frac{2 \cdot n \cdot t^2}{4 \cdot \beta_n^2} \right).$$

Hence we get

$$\begin{aligned}
&\mathbf{E} \left\{ \sup_{f \in \mathcal{G} \circ \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^{n} \epsilon_i \cdot (T_{\beta_n}(f(X_i)) \right| \right\} \\
&\leq \delta_n + \int_{\delta_n}^{\beta_n} c_{24} \cdot \left( \frac{c_{25} \cdot \beta_n}{\delta_n} \right)^{c_{26} \cdot \max\{h \cdot I \cdot d_{ff}, J_n\} \cdot \log(n)} \cdot 2 \cdot \exp \left( - \frac{2 \cdot n \cdot t^2}{4 \cdot \beta_n^2} \right) dt \\
&\leq \delta_n + \int_{\delta_n}^{\beta_n} c_{24} \cdot \left( \frac{c_{25} \cdot \beta_n}{\delta_n} \right)^{c_{26} \cdot \max\{h \cdot I \cdot d_{ff}, J_n\} \cdot \log(n)} \cdot 2 \cdot \exp \left( - \frac{n \cdot \delta_n \cdot t}{2 \cdot \beta_n^2} \right) dt \\
&\leq \delta_n + c_{24} \cdot \left( \frac{c_{25} \cdot \beta_n}{\delta_n} \right)^{c_{26} \cdot \max\{h \cdot I \cdot d_{ff}, J_n\} \cdot \log(n)} \frac{4 \cdot \beta_n^2}{n \cdot \delta_n} \cdot \exp \left( - \frac{n \cdot \delta_n^2}{2 \cdot \beta_n^2} \right).
\end{aligned}$$

With

$$\delta_n = \sqrt{\max\{h \cdot I \cdot d_{ff}, J_n\}} \cdot \log n \cdot \sqrt{\frac{2 \cdot \beta_n^2}{n}}$$

we get the assertion. □

## .5 Approximation error

**Lemma 9** *Let $\tau \in \{l+1, l+2, \dots, l+d+1\}$. Let $l, h, I \in \mathbb{N}$ and set $d_{model} = h \cdot I$. Let $d_{key} \geq 3$. Set $d_v = d_{model}/h = I$. Let $s_0 \in \{1, \dots, h\}$, $s_1, s_2 \in \{1, \dots, d_{model}\}$, $j \in \{1, \dots, l\}$, $k \in \{1, \dots, l\} \setminus \{j\}$, $s_3 \in \{(s_0 - 1) \cdot d_v + 1, \dots, s_0 \cdot d_v\}$, $\beta \in \mathbb{R}$, $\delta \geq 0$ and*

$$B \geq 168 \cdot d_{key} \cdot \tau^2 \cdot l \cdot (|\beta|+1) \cdot \|z_0\|_\infty^2 \cdot \max\{\delta^2, 1\}, \quad 0 \leq \epsilon \leq \min \left\{ 1, \frac{1}{36 \cdot \tau \cdot \|z_0\|_\infty^2} \right\}. \quad (45)$$

*Then there exist*

$$\mathbf{W}_{query,0,s_0}, \quad \mathbf{W}_{key,0,s_0} \quad \text{and} \quad \mathbf{W}_{value,0,s_0}$$

*such that in each row of the above matrices at most $\tau$ of its entries are not equal to zero, such that in the last two rows of $\mathbf{W}_{query,0,s_0}$ and $\mathbf{W}_{key,0,s_0}$ all entries in any column greater than $d+1+l$ are zero, such that all entries are bounded in absolute value by $2 \cdot B$, and such that we have for all $z_{0,r}, \tilde{z}_{0,r} \in \mathbb{R}^{d_{model}}$ satisfying*

$$z_{0,r}^{(s)} = \tilde{z}_{0,r}^{(s)} = \begin{cases} x_r^{(s)} & \text{if } s \in \{1, \dots, d\}, \\ 1 & \text{if } s = d+1, \\ \delta_{s-d-1,r} & \text{if } s \in \{d+2, \dots, d+1+l\} \end{cases} \tag{46}$$

$(r \in \{1, \dots, l\})$ *and*

$$\|\tilde{z}_{0,r} - z_{0,r}\|_\infty \le \delta,$$

$(r \in \{1, \dots, l\})$ *and all*

$$\tilde{\mathbf{W}}_{query,0,s}, \quad \tilde{\mathbf{W}}_{key,0,s} \quad \text{and} \quad \tilde{\mathbf{W}}_{value,0,s}$$

$(s \in \{1, \dots, h\})$ *which satisfy*

$$\|\tilde{\mathbf{W}}_{query,0,s_0} - \mathbf{W}_{query,0,s_0}\|_\infty \le \epsilon,$$

$$\|\tilde{\mathbf{W}}_{key,0,s_0} - \mathbf{W}_{key,0,s_0}\|_\infty \le \epsilon,$$

$$\|\tilde{\mathbf{W}}_{value,0,s_0} - \mathbf{W}_{value,0,s_0}\|_\infty \le \epsilon,$$

*and where in the last two rows of $\tilde{\mathbf{W}}_{query,0,s_0}$ and $\tilde{\mathbf{W}}_{key,0,s_0}$ all entries in any column greater than $d+l+1$ are zero and where in $\tilde{\mathbf{W}}_{query,0,s_0} - \mathbf{W}_{query,0,s_0}$, $\tilde{\mathbf{W}}_{key,0,s_0} - \mathbf{W}_{key,0,s_0}$ and $\tilde{\mathbf{W}}_{value,0,s_0} - \mathbf{W}_{value,0,s_0}$ in each row at most $\tau$ entries are nonzero, that the following holds:*

*If we set for $s \in \{1, \dots, h\}$, $i \in \{1, \dots, l\}$*

$$q_{0,s,i} = \tilde{W}_{query,0,s} \cdot \tilde{z}_{0,i}, \quad k_{0,s,i} = \tilde{W}_{key,0,s} \cdot \tilde{z}_{0,i} \quad \text{and} \quad v_{0,s,i} = \tilde{W}_{value,0,s} \cdot \tilde{z}_{0,i},$$

$$\hat{j}_{s,i} = \arg \max_{j \in \{1,\dots,l\}} < q_{0,s,i}, k_{0,s,j} >,$$

$$\bar{y}_{0,s,i} = v_{0,s,\hat{j}_{s,i}} \cdot < q_{0,s,i}, k_{0,s,\hat{j}_{s,i}} >,$$

$$\bar{y}_{0,i} = (\bar{y}_{0,1,i}, \dots, \bar{y}_{0,h,i})$$

*and*

$$\tilde{y}_{0,i} = \tilde{z}_{0,i} + \bar{y}_{0,i}$$

*then we have:*

$$\hat{j}_{s_0,1} = j, \quad \hat{j}_{s_0,r} = k \quad \text{if } r > 1, \tag{47}$$

$$\begin{aligned} &|\tilde{y}_{0,1}^{(s_3)} - (z_{0,1}^{(s_3)} + z_{0,1}^{(s_1)} \cdot (z_{0,j}^{(s_2)} + \beta))| \\ &\le 136 \cdot d_{key} \cdot \tau^3 \cdot l \cdot (|\beta| + 1) \cdot \|z_0\|_\infty^3 \cdot B \cdot \max\{\delta^3, 1\} \cdot \epsilon \\ &\quad +25 \cdot \tau \cdot (|\beta| + 1) \cdot \|z_0\|_\infty \cdot \max\{\delta, 1\} \cdot \delta \end{aligned} \tag{48}$$

*and*

$$|\tilde{y}_{0,r}^{(s)}| \leq 136 \cdot d_{key} \cdot \tau^3 \cdot l \cdot (|\beta|+1) \cdot \|z_0\|_\infty^3 \cdot B \cdot \max\{\delta^3, 1\} \cdot \epsilon$$
$$+25 \cdot \tau \cdot (|\beta|+1) \cdot \|z_0\|_\infty \cdot \max\{\delta, 1\} \cdot \delta \qquad (49)$$

*whenever* $r > 1$ *or* $s \in \{(s_0-1) \cdot d_v + 1, \ldots, s_0 \cdot d_v\} \setminus \{s_3\}$.

**Proof of Lemma 9.** W.l.o.g. we assume $d_{key} = 3$.

In the *first step of the proof* we define $\mathbf{W}_{query,0,s_0}$, $\mathbf{W}_{key,0,s_0}$ and $\mathbf{W}_{value,0,s_0}$ and present some of their properties.

Set

$$W_{query,0,s_0} = \begin{pmatrix} 0 & \ldots & 0 & 0 & 0 & 0 & 0 & \ldots & 0 & 0 & \ldots & 0 & 1 & 0 & \ldots & 0 \\ 0 & \ldots & 0 & -B & 0 & 0 & 0 & \ldots & 0 & 0 & \ldots & 0 & 0 & 0 & \ldots & 0 \\ 0 & \ldots & 0 & 0 & 0 & 1 & 1 & \ldots & 1 & 0 & \ldots & 0 & 0 & 0 & \ldots & 0 \end{pmatrix}$$

where all columns are zero except columns number $d+1$, $d+3$, $d+4$, $\ldots$, $d+1+l$ and $s_1$,

$$W_{key,0,s_0} =$$
$$\begin{pmatrix} 0 & \ldots & 0 & \beta & 0 & \ldots & 0 & 0 & 0 & \ldots & 0 & 0 & 0 & \ldots & 0 & 0 & \ldots & 0 & 1 & 0 & \ldots & 0 \\ 0 & \ldots & 0 & 0 & 1 & \ldots & 1 & 1 & 1 & \ldots & 1 & 0 & 1 & \ldots & 1 & 0 & \ldots & 0 & 0 & 0 & \ldots & 0 \\ 0 & \ldots & 0 & 0 & 0 & \ldots & 0 & 2 \cdot B & 0 & \ldots & 0 & 0 & 0 & \ldots & 0 & 0 & \ldots & 0 & 0 & 0 & \ldots & 0 \end{pmatrix}$$

where in the first row only the entries in columns $d+1$ and $s_2$ are nonzero, where in the second row only the entries in columns $d+2$, $d+3$, $\ldots$, $d+1+j-1$, $d+1+j+1$, $d+1+j+2$, $\ldots$, $d+1+l$ are nonzero, and where in the third row only the entry in column $d+1+k$ is nonzero, and

$$W_{value,0,s_0} = \begin{pmatrix} 0 & \ldots & 0 & 0 & 0 & \ldots & 0 \\ \vdots & \ldots & \vdots & \vdots & \vdots & \ldots & \vdots \\ 0 & \ldots & 0 & 0 & 0 & \ldots & 0 \\ 0 & \ldots & 0 & 1 & 0 & \ldots & 0 \\ 0 & \ldots & 0 & 0 & 0 & \ldots & 0 \\ \vdots & \ldots & \vdots & \vdots & \vdots & \ldots & \vdots \\ 0 & \ldots & 0 & 0 & 0 & \ldots & 0 \end{pmatrix}$$

where all rows and all columns are zero except column number $d+1+j$ and row number $s_3 - (s_0-1) \cdot d_v$.

Then we have

$$W_{query,0,s_0} \cdot z_{0,r_1} = \begin{pmatrix} z_{0,r_1}^{(s_1)} \\ -B \\ \sum_{i=2}^{l} \delta_{r_1,i} \end{pmatrix}, \quad W_{key,0,s_0} \cdot z_{0,r_2} = \begin{pmatrix} \beta + z_{0,r_2}^{(s_3)} \\ \sum_{i \in \{1,\ldots,l\} \setminus \{j\}} \delta_{r_2,i} \\ \delta_{r_2,k} \cdot 2 \cdot B \end{pmatrix}$$

and

$$W_{value,0,s_0} \cdot z_{0,r_2} = \delta_{r_2,j} \cdot \mathbf{e}_{s_3-(s_0-1) \cdot d_v}$$

where $\mathbf{e}_r$ denotes the $r$-th unit vector in $\mathbb{R}^{d_v}$. Hence

$$< W_{query,0,s_0} z_{0,r_1}, W_{key,0,s_0} z_{0,r_2} >$$
$$= z_{0,r_1}^{(s_1)} \cdot (\beta + z_{0,r_2}^{(s_3)}) - B \cdot \sum_{i \in \{1,\dots,l\} \setminus \{j\}} \delta_{r_2,i} + \sum_{i=2}^{l} \delta_{r_1,i} \cdot \delta_{r_2,k} \cdot 2 \cdot B,$$

which implies

$$< W_{query,0,s_0} z_{0,1}, W_{key,0,s_0} z_{0,j} > \quad = \quad z_{0,1}^{(s_1)} \cdot (\beta + z_{0,j}^{(s_2)}), \tag{50}$$

for $r_2 \neq j$

$$< W_{query,0,s_0} z_{0,1}, W_{key,0,s_0} z_{0,r_2} > \quad = \quad z_{0,1}^{(s_1)} \cdot (\beta + z_{0,r_2}^{(s_2)}) - B,$$

for $r_1 > 1$

$$< W_{query,0,s_0} z_{0,r_1}, W_{key,0,s_0} z_{0,k} > \quad = \quad z_{0,r_1}^{(s_1)} \cdot (\beta + z_{0,k}^{(s_2)}) + B,$$

and for $r_1 > 1$, $r_2 \in \{1, \dots, l\} \setminus \{k\}$

$$< W_{query,0,s_0} z_{0,r_1}, W_{key,0,s_0} z_{0,r_2} > \quad = \quad z_{0,r_1}^{(s_1)} \cdot (\beta + z_{0,r_2}^{(s_2)}) - B \cdot (1 - \delta_{r_2,j}).$$

Because of

$$B > 4 \cdot \max_{r_1,r_2} |z_{0,r_1}^{(s_1)} \cdot (\beta + z_{0,r_2}^{(s_2)})|$$

we conclude

$$< W_{query,0,s_0} z_{0,1}, W_{key,0,s_0} z_{0,j} >$$
$$> \frac{B}{2} + \max_{r_2 \neq j} < W_{query,0,s_0} z_{0,1}, W_{key,0,s_0} z_{0,r_2} > \tag{51}$$

and for $r_1 > 1$

$$< W_{query,0,s_0} z_{0,r_1}, W_{key,0,s_0} z_{0,k} >$$
$$> \frac{B}{2} + \max_{r_2 \neq k} < W_{query,0,s_0} z_{0,r_1}, W_{key,0,s_0} z_{0,r_2} > . \tag{52}$$

Furthermore we have

$$W_{value,0,s_0} z_{0,r} = \delta_{r,j} \cdot \mathbf{e}_{s_3-(s_0-1)\cdot d_v}. \tag{53}$$

In the *second step of the proof* we bound the difference between

$$< \tilde{W}_{query,0,s_0} \tilde{z}_{0,r_1}, \tilde{W}_{key,0,s_0} \tilde{z}_{0,r_2} > \quad \text{and} \quad < W_{query,0,s_0} z_{0,r_1}, W_{key,0,s_0} z_{0,r_2} > .$$

We have

$$| < \tilde{W}_{query,0,s_0} \tilde{z}_{0,r_1}, \tilde{W}_{key,0,s_0} \tilde{z}_{0,r_2} > - < W_{query,0,s_0} z_{0,r_1}, W_{key,0,s_0} z_{0,r_2} > |$$

$$
\begin{aligned}
&= | < (\tilde{W}_{query,0,s_0} - W_{query,0,s_0})\tilde{z}_{0,r_1} + W_{query,0,s_0}(\tilde{z}_{0,r_1} - z_{0,r_1}) + W_{query,0,s_0} z_{0,r_1}, \\
&\qquad\qquad (\tilde{W}_{key,0,s_0} - W_{key,0,s_0})\tilde{z}_{0,r_2} + W_{key,0,s_0}(\tilde{z}_{0,r_2} - z_{0,r_2}) + W_{key,0,s_0} z_{0,r_2} > \\
&\quad - < W_{query,0,s_0} z_{0,r_1}, W_{key,0,s_0} z_{0,r_2} > | \\
&\leq | < (\tilde{W}_{query,0,s_0} - W_{query,0,s_0})\tilde{z}_{0,r_1}), (\tilde{W}_{key,0,s_0} - W_{key,0,s_0})\tilde{z}_{0,r_2} > | \\
&\quad + | < (\tilde{W}_{query,0,s_0} - W_{query,0,s_0})\tilde{z}_{0,r_1}), W_{key,0,s_0}(\tilde{z}_{0,r_2} - z_{0,r_2}) > | \\
&\quad + | < (\tilde{W}_{query,0,s_0} - W_{query,0,s_0})\tilde{z}_{0,r_1}, W_{key,0,s_0} z_{0,r_2} > | \\
&\quad + | < W_{query,0,s_0}(\tilde{z}_{0,r_1} - z_{0,r_1}), (\tilde{W}_{key,0,s_0} - W_{key,0,s_0})\tilde{z}_{0,r_2} > | \\
&\quad + | < W_{query,0,s_0}(\tilde{z}_{0,r_1} - z_{0,r_1}), W_{key,0,s_0}(\tilde{z}_{0,r_2} - z_{0,r_2}) > | \\
&\quad + | < W_{query,0,s_0}(\tilde{z}_{0,r_1} - z_{0,r_1}), W_{key,0,s_0} z_{0,r_2} > | \\
&\quad + | < W_{query,0,s_0} z_{0,r_1}, (\tilde{W}_{key,0,s_0} - W_{key,0,s_0})\tilde{z}_{0,r_2} > | \\
&\quad + | < W_{query,0,s_0} z_{0,r_1}, W_{key,0,s_0}(\tilde{z}_{0,r_2} - z_{0,r_2}) > | \\
&=: \sum_{i=1}^{8} T_i.
\end{aligned}
$$

We have

$$T_1 \leq d_{key} \cdot \tau^2 \cdot (\|z_0\|_\infty + \delta)^2 \cdot \epsilon^2,$$

$$T_2 \leq \tau \cdot (\|z_0\|_\infty + \delta) \cdot \epsilon \cdot \delta$$

(where we have used the fact that $\tilde{z}_{0,r_2} - z_{0,r_2}$ is zero in components less than $d + l + 2$ and consequently only the first component of $W_{key,0,s_0}(\tilde{z}_{0,r_2} - z_{0,r_2})$ is nonzero),

$$T_3 \leq \tau \cdot (\|z_0\|_\infty + \delta) \cdot \epsilon \cdot (|\beta| + 1) \cdot \|z_0\|_\infty + \tau \cdot \|z_0\|_\infty \cdot \epsilon \cdot l \cdot \|z_0\|_\infty + \tau \cdot \|z_0\|_\infty \cdot \epsilon \cdot 2 \cdot B$$

(that is the consequence of the fact that the last two components of $(\tilde{W}_{query,0,s_0} - W_{query,0,s_0})\tilde{z}_{0,r_1}$ depend on $z_{0,r_1}$ and not on $\tilde{z}_{0,r_1}$, which follows from the assumption that in the last two rows of $\tilde{W}_{query,0,s_0}$ and $\tilde{W}_{key,0,s_0}$ all entries in columns greater than $d + l + 1$ are zero),

$$T_4 \leq \delta \cdot \tau \cdot (\|z_0\|_\infty + \delta) \cdot \epsilon$$

(where we have used the fact that only the first component of $W_{query,0,s_0}(\tilde{z}_{0,r_1} - z_{0,r_1})$ is nonzero)

$$T_5 \leq \delta \cdot \delta,$$

$$T_6 \leq \delta \cdot (|\beta| + 1) \cdot \|z_0\|_\infty,$$

$$T_7 \leq \|z_0\|_\infty \cdot \tau \cdot \epsilon \cdot (\|z_0\|_\infty + \delta) + (B + l) \cdot \|z_0\|_\infty \cdot \tau \cdot \epsilon \cdot \|z_0\|_\infty$$

(where we have used the fact that the last two components of $(\tilde{W}_{key,0,s_0} - W_{key,0,s_0})\tilde{z}_{0,r_2}$ depend on $z_{0,r_2}$ and not on $\tilde{z}_{0,r_2}$, which follows as above from the assumption that in the last two rows of $\tilde{W}_{query,0,s_0}$ and $\tilde{W}_{key,0,s_0}$ all entries in columns greater than $d+l+1$ are zero), and

$$T_8 \leq \|z_0\|_\infty \cdot (|\beta| + 1) \cdot \delta$$

(where we have used the fact that $\tilde{z}_{0,r_2} - z_{0,r_2}$ is zero in components greater $d+l+1$ and consequently only the first component of $W_{key,0,s_0}(\tilde{z}_{0,r_2} - z_{0,r_2})$ is nonzero). This proves

$$\begin{aligned}
& | < \tilde{W}_{query,0,s_0} \tilde{z}_{0,r_1}, \tilde{W}_{key,0,s_0} \tilde{z}_{0,r_2} > - < W_{query,0,s_0} z_{0,r_1}, W_{key,0,s_0} z_{0,r_2} > | \\
& \leq 14 \cdot d_{key} \cdot \tau^2 \cdot l \cdot (|\beta| + 1) \cdot \|z_0\|_\infty^2 \cdot \max\{\delta^2, 1\} \cdot \epsilon \\
& \quad +3 \cdot (|\beta| + 1) \cdot \|z_0\|_\infty \cdot \max\{\delta, 1\} \cdot \delta \\
& \quad +3 \cdot B \cdot \epsilon \cdot \tau \cdot \|z_0\|_\infty^2. \qquad (54)
\end{aligned}$$

Since we have

$$\epsilon \leq \min\left\{1, \frac{1}{36 \cdot \tau \cdot \|z_0\|_\infty^2}\right\}, \quad B > 36 \cdot (|\beta| + 1) \cdot \|z_0\|_\infty^2 \cdot \max\{\delta, 1\} \cdot \epsilon$$

and

$$B > 168 \cdot d_{key} \cdot \tau^2 \cdot l \cdot (|\beta| + 1) \cdot \|z_0\|^2 \cdot \max\{\delta^2, 1\} \cdot \epsilon$$

the right-hand side of (54) is less than $B/4$.

In the *third step of the proof* we show (47).

To do this, we conclude from step 2 that we have

$$\begin{aligned}
& < \tilde{W}_{query,0,s_0} \tilde{z}_{0,r_1}, \tilde{W}_{key,0,s_0} \tilde{z}_{0,r_2} > - \max_{r_3 \neq r_2} < \tilde{W}_{query,0,s_0} \tilde{z}_{0,r_1}, \tilde{W}_{key,0,s_0} \tilde{z}_{0,r_3} > \\
& > \quad < W_{query,0,s_0} z_{0,r_1}, W_{key,0,s_0} z_{0,r_2} > - \max_{r_3 \neq r_2} < W_{query,0,s_0} z_{0,r_1}, W_{key,0,s_0} z_{0,r_3} > - \frac{B}{2}.
\end{aligned}$$

The assertion follows from (51) and (52).

In the *fourth step of the proof* we show the assertion. Because of (47), (50) and (53) it suffices to show

$$\begin{aligned}
& \Bigg\| \tilde{W}_{value,0,s_0} \tilde{z}_{0,r_2} \cdot < \tilde{W}_{query,0,s_0} \tilde{z}_{0,r_1}, \tilde{W}_{key,0,s_0} \tilde{z}_{0,r_2} > \\
& \qquad\qquad - W_{value,0,s_0} z_{0,r_2} < W_{query,0,s_0} z_{0,r_1}, W_{key,0,s_0} z_{0,r_2} > \Bigg\|_\infty \\
& \leq 144 \cdot d_{key} \cdot \tau^3 \cdot l \cdot (|\beta| + 1) \cdot \|z_0\|_\infty^3 \cdot B \cdot \max\{\delta^3, 1\} \cdot \epsilon \\
& \quad +24 \cdot \tau \cdot (|\beta| + 1) \cdot \|z_0\|_\infty \cdot \max\{\delta, 1\} \cdot \delta.
\end{aligned}$$

We have

$$\begin{aligned}
& \Bigg\| \tilde{W}_{value,0,s_0} \tilde{z}_{0,r_2} \cdot < \tilde{W}_{query,0,s_0} \tilde{z}_{0,r_1}, \tilde{W}_{key,0,s_0} \tilde{z}_{0,r_2} > \\
& \qquad\qquad - W_{value,0,s_0} z_{0,r_2} < W_{query,0,s_0} z_{0,r_1}, W_{key,0,s_0} z_{0,r_2} > \Bigg\|_\infty \\
& \leq \Bigg\| \tilde{W}_{value,0,s_0} \tilde{z}_{0,r_2} \cdot \Big( < \tilde{W}_{query,0,s_0} \tilde{z}_{0,r_1}, \tilde{W}_{key,0,s_0} \tilde{z}_{0,r_2} >
\end{aligned}$$

$$- < W_{query,0,s_0} z_{0,r_1}, W_{key,0,s_0} z_{0,r_2} > \Big) \Bigg\|_\infty$$

$$+ \Bigg\| (\tilde{W}_{value,0,s_0} \tilde{z}_{0,r_2} - W_{value,0,s_0} z_{0,r_2}) \cdot < W_{query,0,s_0} z_{0,r_1}, W_{key,0,s_0} z_{0,r_2} > \Bigg\|_\infty$$

$$\leq \|\tilde{W}_{value,0,s_0} \tilde{z}_{0,r_2}\|_\infty$$

$$\cdot | < \tilde{W}_{query,0,s_0} \tilde{z}_{0,r_1}, \tilde{W}_{key,0,s_0} \tilde{z}_{0,r_2} > - < W_{query,0,s_0} z_{0,r_1}, W_{key,0,s_0} z_{0,r_2} > |$$

$$+ \|\tilde{W}_{value,0,s_0} \tilde{z}_{0,r_2} - W_{value,0,s_0} z_{0,r_2}\|_\infty \cdot | < W_{query,0,s_0} z_{0,r_1}, W_{key,0,s_0} z_{0,r_2} > |$$

$$\leq (\tau + 1) \cdot (1 + \epsilon) \cdot (\|z\|_\infty + \delta) \cdot (14 \cdot d_{key} \cdot \tau^2 \cdot l \cdot (|\beta| + 1) \cdot \|z_0\|_\infty^2 \cdot \max\{\delta^2, 1\} \cdot \epsilon$$

$$+3 \cdot (|\beta| + 1) \cdot \|z_0\|_\infty \cdot \max\{\delta, 1\} \cdot \delta + 3 \cdot B \cdot \epsilon \cdot \tau \cdot \|z_0\|_\infty^2)$$

$$+ \|\tilde{W}_{value,0,s_0} \tilde{z}_{0,r_2} - W_{value,0,s_0} z_{0,r_2}\|_\infty \cdot (\|z_0\|_\infty \cdot (|\beta| + \|z_0\|_\infty) + 2 \cdot B).$$

With

$$\|\tilde{W}_{value,0,s_0} \tilde{z}_{0,r_2} - W_{value,0,s_0} z_{0,r_2}\|_\infty$$
$$\leq \|(\tilde{W}_{value,0,s_0} - W_{value,0,s_0}) \tilde{z}_{0,r_2}\|_\infty + \|W_{value,0,s_0} (\tilde{z}_{0,r_2} - z_{0,r_2})\|_\infty$$
$$\leq \tau \cdot \epsilon \cdot (\|z_0\|_\infty + \delta) + 0$$

(where we have used that $\tilde{z}_{0,r_2} - z_{0,r_2}$ is zero in components less than $d + l + 2$) we get the assertion. □

**Lemma 10** *Let $\epsilon \in [0, 1)$, let $\delta \geq 0$ and let $\alpha \in \mathbb{R}$. Let $d_{ff}, d_{model} \in \mathbb{N}$ and assume $d_{ff} \geq 4$. Let $j_1, j_2 \in \{1, \ldots, d_{model}\}$ with $j_1 \neq j_2$. Then there exist*

$$W_{r,1} \in \mathbb{R}^{d_{ff} \times d_{model}}, b_{r,1} \in \mathbb{R}^{d_{ff}}, W_{r,2} \in \mathbb{R}^{d_{model} \times d_{ff}}, b_{r,2} \in \mathbb{R}^{d_{model}},$$

*where in $W_{r,1}$ in each row and in $W_{r,2}$ in each column at most 2 components are not equal to zero and where all entries are bounded in absolute value by $\max\{|\alpha|, 1\}$, such that for all $\tilde{W}_{r,1} \in \mathbb{R}^{d_{ff} \times d_{model}}$, $\tilde{b}_{r,1} \in \mathbb{R}^{d_{ff}}$, $\tilde{W}_{r,2} \in \mathbb{R}^{d_{model} \times d_{ff}}$, $\tilde{b}_{r,2} \in \mathbb{R}^{d_{model}}$ with*

$$\|W_{r,1} - \tilde{W}_{r,1}\|_\infty < \epsilon, \quad \|b_{r,1} - \tilde{b}_{r,1}\|_\infty < \epsilon, \quad \|W_{r,2} - \tilde{W}_{r,2}\|_\infty < \epsilon, \quad \|b_{r,2} - \tilde{b}_{r,2}\|_\infty < \epsilon,$$

*and all $y_{r,i}, \tilde{y}_{r,i} \in \mathbb{R}^{d_{model}}$ $(i \in \{1, \ldots, l\})$ with*

$$\|y_{r,i} - \tilde{y}_{r,i}\|_\infty < \delta \quad (i \in \{1, \ldots, l\})$$

*and*

$$\tilde{z}_{r,s} = \tilde{y}_{r,s} + \tilde{W}_{r,2} \cdot \sigma\left(\tilde{W}_{r,1} \cdot \tilde{y}_{r,s} + \tilde{b}_{r,1}\right) + \tilde{b}_{r,2} \quad (s \in \{1, \ldots, l\})$$

*we have:*

$$\left|\tilde{z}_{r,s}^{(j_1)} - \alpha \cdot \max\{y_{r,s}^{(j_2)}, 0\}\right| \leq 5 \cdot d_{ff} \cdot \max\{|\alpha|, 1\} \cdot (\|y_{r,s}\|_\infty + 1) \cdot d_{model} \cdot (\delta + \epsilon),$$

$$|\tilde{z}_{r,s}^{(j_2)}| \leq 5 \cdot d_{ff} \cdot \max\{|\alpha|, 1\} \cdot (\|y_{r,s}\|_\infty + 1) \cdot d_{model} \cdot (\delta + \epsilon)$$

*and*

$$|\tilde{z}_{r,s}^{(j)} - y_{r,s}^{(j)}| \leq 5 \cdot d_{ff} \cdot \max\{|\alpha|, 1\} \cdot (\|y_{r,s}\|_\infty + 1) \cdot d_{model} \cdot (\delta + \epsilon)$$

*whenever* $j \in \{1, \dots, d_{model}\} \setminus \{j_1, j_2\}$.

*Furthermore, the assertion of the lemma holds also if we replace* $\alpha \cdot \max\{y_{r,s}^{(j_2)}, 0\}$ *by* $\alpha \cdot y_{r,s}^{(j_2)}$.

**Proof.** Set

$$z_{r,s} = y_{r,s} + W_{r,2} \cdot \sigma\left(W_{r,1} \cdot y_{r,s} + b_{r,1}\right) + b_{r,2} \quad (s \in \{1, \dots, l\}).$$

In the *first step of the proof* we show that we can choose

$$W_{r,1} \in \mathbb{R}^{d_{ff} \times d_{model}}, b_{r,1} \in \mathbb{R}^{d_{ff}}, W_{r,2} \in \mathbb{R}^{d_{model} \times d_{ff}}, b_{r,2} \in \mathbb{R}^{d_{model}},$$

such that at most 9 components are not equal to zero and such that

$$z_{r,s}^{(j_1)} = \alpha \cdot \max\{y_{r,s}^{(j_2)}, 0\}, \quad z_{r,s}^{(j_2)} = 0 \quad \text{and} \quad z_{r,s}^{(j)} = y_{r,s}^{(j)}$$

hold whenever $j \in \{1, \dots, d_{model}\} \setminus \{j_1, j_2\}$.

W.l.o.g. we assume $d_{ff} = 4$ and $j_1 < j_2$. We choose $b_{r,1} = 0$, $b_{r,2} = 0$,

$$W_{r,1} = \begin{pmatrix} 0 & \dots & 0 & 1 & 0 & \dots & 0 & 0 & 0 & \dots & 0 \\ 0 & \dots & 0 & -1 & 0 & \dots & 0 & 0 & 0 & \dots & 0 \\ 0 & \dots & 0 & 0 & 0 & \dots & 0 & 1 & 0 & \dots & 0 \\ 0 & \dots & 0 & 0 & 0 & \dots & 0 & -1 & 0 & \dots & 0 \end{pmatrix},$$

where all columns except columns number $j_1$ and $j_2$ are zero, and

$$W_{r,2} = \begin{pmatrix} 0 & 0 & 0 & 0 \\ \vdots & & & \vdots \\ 0 & 0 & 0 & 0 \\ -1 & 1 & \alpha & 0 \\ 0 & 0 & 0 & 0 \\ \vdots & & & \vdots \\ 0 & 0 & 0 & 0 \\ 0 & 0 & -1 & 1 \\ 0 & 0 & 0 & 0 \\ \vdots & & & \vdots \\ 0 & 0 & 0 & 0 \end{pmatrix},$$

where all rows except row number $j_1$ and $j_2$ are zero. Then we have

$$W_{2,r} \cdot \sigma\left(W_{1,r} \cdot y_{r,s} + b_{1,r}\right) + b_{2,r}$$

$$= \begin{pmatrix} 0 \\ \vdots \\ 0 \\ \alpha \cdot \sigma(y_{r,s}^{(j_2)}) - (\sigma(y_{r,s}^{(j_1)}) - \sigma(-y_{r,s}^{(j_1)})) \\ 0 \\ \vdots \\ 0 \\ -(\sigma(y_{r,s}^{(j_2)}) - \sigma(-y_{r,s}^{(j_2)})) \\ 0 \\ \vdots \\ 0 \end{pmatrix}.$$

Because of

$$\sigma(u) - \sigma(-u) = u$$

for $u \in \mathbb{R}$ this implies the assertion of the first step.

In the *second step of the proof* we show

$$|\tilde{z}_{r,s}^{(j)} - z_{r,s}^{(j)}| \le 5 \cdot d_{ff} \cdot \max\{|\alpha|, 1\} \cdot (\|y_{r,s}\|_\infty + 1) \cdot d_{model} \cdot (\delta + \epsilon).$$

We have

$$\begin{aligned} &\|\tilde{W}_{r,1} \cdot \tilde{y}_{r,s} + \tilde{b}_{r,1} - (W_{r,1} \cdot y_{r,s} + b_{r,1})\|_\infty \\ &\le \|(\tilde{W}_{r,1} - W_{r,1}) \cdot \tilde{y}_{r,s}\|_\infty + \|W_{r,1} \cdot (\tilde{y}_{r,s} - y_{r,s})\|_\infty + \epsilon \\ &\le d_{model} \cdot \epsilon \cdot (\|y_{r,s}\|_\infty + \delta) + \delta + \epsilon \\ &\le \delta + d_{model} \cdot \epsilon \cdot (\|y_{r,s}\|_\infty + 1 + \delta), \end{aligned}$$

which implies

$$\begin{aligned} &\|\sigma(\tilde{W}_{r,1} \cdot \tilde{y}_{r,s} + \tilde{b}_{r,1}) - \sigma(W_{r,1} \cdot y_{r,s} + b_{r,1})\|_\infty \\ &\le \delta + d_{model} \cdot \epsilon \cdot (\|y_{r,s}\|_\infty + 1 + \delta) \end{aligned}$$

and

$$\begin{aligned} &\|\sigma(\tilde{W}_{r,1} \cdot \tilde{y}_{r,s} + \tilde{b}_{r,1})\|_\infty \\ &\le \delta + d_{model} \cdot \epsilon \cdot (\|y_{r,s}\|_\infty + 1 + \delta) + \|\sigma(W_{r,1} \cdot y_{r,s} + b_{r,1})\|_\infty \\ &\le \delta + d_{model} \cdot \epsilon \cdot (\|y_{r,s}\|_\infty + 1 + \delta) + \|y_{r,s}\|_\infty. \end{aligned}$$

From this we conclude

$$\begin{aligned} &|\tilde{z}_{r,s}^{(j)} - z_{r,s}^{(j)}| \\ &\le \left\| \tilde{y}_{r,s} + \tilde{W}_{r,2} \cdot \sigma\left(\tilde{W}_{r,1} \cdot \tilde{y}_{r,s} + \tilde{b}_{r,1}\right) + \tilde{b}_{r,2} \right. \end{aligned}$$

$$
\begin{aligned}
&\qquad\qquad - \left(y_{r,s} + W_{r,2} \cdot \sigma\left(W_{r,1} \cdot y_{r,s} + b_{r,1}\right) + b_{r,2}\right) \Big\|_\infty \\
&\leq \delta + \left\| \tilde{W}_{r,2} \cdot \sigma\left(\tilde{W}_{r,1} \cdot \tilde{y}_{r,s} + \tilde{b}_{r,1}\right) - W_{r,2} \cdot \sigma\left(W_{r,1} \cdot y_{r,s} + b_{r,1}\right) \right\|_\infty + \epsilon \\
&\leq \delta + \left\| (\tilde{W}_{r,2} - W_{r,2}) \cdot \sigma\left(\tilde{W}_{r,1} \cdot \tilde{y}_{r,s} + \tilde{b}_{r,1}\right) \right\|_\infty \\
&\quad + \left\| W_{r,2} \cdot \left( \sigma\left(\tilde{W}_{r,1} \cdot \tilde{y}_{r,s} + \tilde{b}_{r,1}\right) - \sigma\left(W_{r,1} \cdot y_{r,s} + b_{r,1}\right) \right) \right\|_\infty + \epsilon \\
&\leq \delta + d_{ff} \cdot \left\| \tilde{W}_{r,2} - W_{r,2} \right\|_\infty \cdot \left\| \sigma\left(\tilde{W}_{r,1} \cdot \tilde{y}_{r,s} + \tilde{b}_{r,1}\right) \right\|_\infty \\
&\quad + d_{ff} \cdot \|W_{r,2}\|_\infty \cdot \left\| \sigma\left(\tilde{W}_{r,1} \cdot \tilde{y}_{r,s} + \tilde{b}_{r,1}\right) - \sigma\left(W_{r,1} \cdot y_{r,s} + b_{r,1}\right) \right\|_\infty + \epsilon \\
&\leq \delta + d_{ff} \cdot \epsilon \cdot (\delta + d_{model} \cdot \epsilon \cdot (\|y_{r,s}\|_\infty + 1 + \delta) + \|y_{r,s}\|_\infty) \\
&\quad + d_{ff} \cdot \max\{|\alpha|, 1\} \cdot (\delta + d_{model} \cdot \epsilon \cdot (\|y_{r,s}\|_\infty + 1 + \delta)) + \epsilon \\
&\leq 5 \cdot d_{ff} \cdot \max\{|\alpha|, 1\} \cdot (\|y_{r,s}\|_\infty + 1) \cdot d_{model} \cdot (\delta + \epsilon).
\end{aligned}
$$

By slightly changing the matrix $W_{r,2}$ we conclude that the assertion also holds upon replacing $\alpha \cdot \max\{y_{r,s}^{(j_2)}, 0\}$ by $\alpha \cdot y_{r,s}^{(j_2)}$. □

**Lemma 11** *Let $I \geq d + l + 4$, let $\tau \in \{l + 1, l + 2, \ldots, l + d + 1\}$. Let $c \geq A \geq 1$, let $K \in \mathbb{N}$, let $u_k \in [-A, A]$ $(k = 1, \ldots, K - 1)$, set*

$$
B_j(x) = x^j \quad \text{for } j = 0, 1, \ldots, M
$$

*and set*

$$
B_j(x) = (x - u_{j-M})_+^M \quad \text{for } j = M + 1, M + 2, \ldots, M + K - 1.
$$

*Let $i \in \{d + l + 4, \ldots, I\}$. Let $h \in \mathbb{N}$ with $1 \leq h \leq n$ and for $s \in \{2, \ldots, h\}$ let $j_{s,1}, \ldots, j_{s,d} \in \{0, 1, \ldots, M + K - 1\}$ and $\alpha_s \in \mathbb{R}$. Let $d_{key} \geq 3$ and set $d_{ff} = 2 \cdot h + 2$. Then there exists a transformer encoder consisting of $M \cdot (d \cdot l) + 1$ pairs of layers, where the first layer is a multi-head attention layer with $h$ attention units and the second layer is a pointwise feedforward neural network, and where all matrices in the attention heads have in each row at most $\tau$ nonzero entries, and where all matrices $W_{r,1}$ and $W_{r,2}$ in the pointwise feedforward neural networks have the property that in each row in $W_{r,1}$ and in each column in $W_{r,2}$ at most $\tau$ of the entries are nonzero and where all matrices and vectors depend only on $(u_k)_k$ and $\alpha_s$ $(s \in \{1, \ldots, k\})$ and all entries are bounded in absolute value by*

$$
c_{27} \cdot n^{12^{M \cdot (d \cdot l)}} \cdot (d_{model})^{12^{M \cdot (d \cdot l)}} \cdot c^{2 \cdot 12^{M \cdot (d \cdot l)}},
$$

*and where the matrices $\tilde{W}_{query,0,s}$, $\tilde{W}_{key,0,s}$ and $\tilde{W}_{value,0,s}$ satisfy the assumptions of Lemma 9, which has the following property: Any transformer network which has the same structure as the above Transformer network and whose weights are in supremum norm no further than $\epsilon$ away from the weights of the above network for some*

$$
0 \leq \epsilon \leq \min\left\{1, \frac{1}{36 \cdot \tau \cdot (2 \cdot A)^{M \cdot d \cdot l}}\right\}
$$

*has the property that if it gets as input $\tilde{z}_0$ which satisfies (46) and*

$$\|\tilde{z}_0 - z_0\|_\infty \leq c \cdot \epsilon$$

*for some $z_0 \in [-A, A]^{l \cdot d_{model}}$ defined as in Subsection 2.1 (which encodes in particular $x = (x_1^T, \ldots, x_l^T) \in \mathbb{R}^{d \cdot l}$) it produces as output $\tilde{z}_{M \cdot d+1}$, which satisfies for $n$ sufficiently large*

$$|\tilde{z}_{M \cdot d,1}^{(i)} - \sum_{s=2}^{h} \alpha_s \prod_{k=1}^{d \cdot l} B_{j_{s,k}}(x^{(k)})| \leq c_{28} \cdot n^{6^{M \cdot (d \cdot l)}+1} \cdot (d_{model})^{6^{M \cdot (d \cdot l)}+1} \cdot c^{8^{M \cdot (d \cdot l)+2}+2 \cdot (M \cdot d \cdot l+1)} \cdot \epsilon$$

*and*

$$|\tilde{z}_{M \cdot d,j}^{(l)} - z_{M \cdot d,j}^{(l)}| \leq c_{28} \cdot n^{6^{M \cdot (d \cdot l)}+1} \cdot (d_{model})^{6^{M \cdot (d \cdot l)}+1} \cdot c^{8^{M \cdot (d \cdot l)+2} \cdot (M \cdot d \cdot l+1)} \cdot \epsilon$$

*whenever $j > 1$ or*

$$l \in \{1, \ldots, d_{model}\} \setminus \{i, I+d+l+2, I+d+l+3, 2 \cdot I+d+l+2, \\ 2 \cdot I+d+l+3, \ldots, (h-1) \cdot I+d+l+2, (h-1) \cdot I+d+l+3\}.$$

**Proof.** In the *first step of the proof* we show that the $h-1$ products of the B-splines in the sum above can be computed in the first $M \cdot (d \cdot l)$ pairs of attention heads and pointwise feedforward network.

The basic idea is as follows. Each attention head of the network works only on one of the parts $2, \ldots, h$ of length $I$ of the input. It uses the fact that each $B_j(x)$ can be written as

$$B_j(x) = \prod_{k=1}^{M} B_{j,k}(x)$$

where $B_{j,k}(x)$ is one of the functions

$$x \mapsto 1, \quad x \mapsto x \quad \text{and} \quad x \mapsto (x - u_r)_+.$$

Using Lemma 9 (with a suitable value for $B$, which will be chosen in the third step of the proof) and Lemma 10 we can combine an attention layer and a pointwise feedforward layer such that the following holds: They get as input an approximation $\tilde{z}_{0,j}$ of $z_{0,j}$ where $z_{0,j}$ is given as in Lemma 9 and where the component $(s-1) \cdot I+d+l+2$ of $z_{0,j}$ is zero, and they modify the components $(s-1) \cdot I+d+l+2$ and $(s-1) \cdot I+d+l+3$ of $z_{0,j}$. More precisely, they combine the attention head of Lemma 9 and the pointwise feedforward neural network of Lemma 10 such that they produce an output $\tilde{y}_j$ where $\tilde{y}_1^{((s-1) \cdot I+d+l+3)}$ is an approximation of the product of an approximation of either $z_{0,1}^{(d+1)} = 1$ or $z_{0,1}^{((s-1) \cdot I+d+l+3)}$ and one of the functions

$$x^{(k)} \mapsto 1, \quad x^{(k)} \mapsto x^{(k)} \quad \text{and} \quad x^{(k)} \mapsto (x^{(k)} - u_r)_+ \quad (r \in \{1, \ldots, K-1\})$$

and where $\tilde{y}_j^{(r)}$ is approximately equal to $z_{0,j}^{(r)}$ otherwise ($r \in \{(s-1) \cdot I + 1, \dots, s \cdot I\} \setminus \{(s-1) \cdot I + d + l + 2, (s-1) \cdot I + d + l + 3\}$). Using this $M \cdot (d \cdot l)$ times we get an approximation of

$$\alpha_s \prod_{k=1}^{d \cdot l} B_{j_{s,k}}(x^{(k)}) \tag{55}$$

in $z_{M \cdot d,1}^{((s-1) \cdot I + (d+l+3))}$ for $s = 2, \dots, h$.

In the *second step of the proof* we show how one pair of attention head and pointwise feedforward neural network can be used to compute the sum of the values in (55). To do this, we choose $W_{value,M \cdot (d \cdot l)+1} = 0$ (which results in $y_{M \cdot (d \cdot l)+1} = z_{M \cdot (d \cdot l)}$ and $\tilde{y}_{M \cdot (d \cdot l)+1} \approx \tilde{z}_{M \cdot (d \cdot l)}$) and choose $W_{1,M \cdot (d \cdot l)+1}$, $b_{1,M \cdot (d \cdot l)+1}$, $W_{2,M \cdot (d \cdot l)+1}$, $b_{2,M \cdot (d \cdot l)+1}$, such that

$$W_{2,M \cdot (d \cdot l)+1} \cdot \sigma\left(W_{1,M \cdot (d \cdot l)+1} \cdot y_{M \cdot (d \cdot l)+1} + b_{1,M \cdot (d \cdot l)+1}\right) + b_{2,M \cdot (d \cdot l)+1}$$

$$= \begin{pmatrix} 0 \\ \vdots \\ 0 \\ -(\sigma(y_{M \cdot d+1,1}^{(i)}) - \sigma(-y_{M \cdot d+1,1}^{(i)})) + \sigma(\sum_{s=2}^h y_{M \cdot d+1,1}^{(s-1) \cdot I + (d+l+3)}) - \sigma(-\sum_{s=2}^h y_{M \cdot d+1,1}^{(s-1) \cdot I + (d+l+3)}) \\ 0 \\ \vdots \\ 0 \end{pmatrix}$$

holds, where the nonzero entry is in row number $i$.

In the *third step of the proof* we analyze the error occurring in the above approximation. Here we describe in particular how the values of $B$ in the application of Lemma 9 need to be chosen. In the applications of Lemma 9 we will have

$$\|z_0\|_\infty \leq (2 \cdot A)^{M \cdot d \cdot l} \quad \text{and} \quad |\beta| \leq A.$$

In the first layer we set

$$B = B_1 = c_{29} \cdot c^{2 \cdot M \cdot d \cdot l + 3},$$

which implies

$$\begin{aligned} &168 \cdot d_{key} \cdot \tau^2 \cdot l \cdot (|\beta| + 1) \cdot \|z_0\|_\infty^2 \cdot \max\{(c \cdot \epsilon)^2, 1\} \leq c_{30} \cdot A^{2 \cdot M \cdot d \cdot l + 1} \cdot \max\{(c \cdot \epsilon)^2, 1\} \\ &\leq c_{30} \cdot c^{2 \cdot M \cdot d \cdot l + 3} \cdot \max\{\epsilon^2, 1\} \leq B_1. \end{aligned}$$

From this we can conclude by Lemma 9 that the output of the first attention head has an error not exceeding

$$\begin{aligned} &136 \cdot d_{key} \cdot \tau^3 \cdot l \cdot (|\beta| + 1) \cdot \|z_0\|_\infty^3 \cdot B_1 \cdot \max\{(c \cdot \epsilon)^3, 1\} \cdot \epsilon \\ &\quad +25 \cdot (|\beta| + 1) \cdot \|z_0\|_\infty \cdot \max\{c \cdot \epsilon, 1\} \cdot (c \cdot \epsilon) \\ &\leq c_{31} \cdot c^{5 \cdot M \cdot d \cdot l + 7} \cdot \epsilon. \end{aligned}$$

Application of Lemma 10 (where the input is bounded in absolute value by $c_{32} \cdot A^{M \cdot d \cdot l} \leq c_{32} \cdot c^{M \cdot d \cdot l}$) yields that after the pointwise feedforward neural network the error is maximal

$$\delta_1 = c_{33} \cdot n \cdot d_{model} \cdot c^{6 \cdot M \cdot d \cdot l + 7} \cdot \epsilon.$$

In the second level we set

$$B = B_2 = c_{34} \cdot n^2 \cdot d_{model}^2 \cdot c^{14 \cdot M \cdot d \cdot l + 14}$$

which implies

$$\begin{aligned} & 168 \cdot d_{key} \cdot \tau^2 \cdot l \cdot (|\beta| + 1) \cdot \|z_0\|_\infty^2 \cdot \max\{\delta_1^2, 1\} \\ & \leq c_{35} \cdot c^{2 \cdot M \cdot d \cdot l} \max\{(n \cdot d_{model} \cdot c^{6 \cdot M \cdot d \cdot l + 7} \cdot \epsilon)^2, 1\} \\ & \leq c_{35} \cdot n^2 \cdot d_{model}^2 \cdot c^{14 \cdot M \cdot d \cdot l + 14} \cdot \max\{\epsilon^2, 1\} \leq B_2. \end{aligned}$$

From this we can conclude by Lemma 9 that the output of the second attention head has error not exceeding

$$\begin{aligned} & 136 \cdot d_{key} \cdot \tau^3 \cdot l \cdot (|\beta| + 1) \cdot \|z_0\|_\infty^3 \cdot B_2 \cdot \max\{(c_{33} \cdot n \cdot d_{model} \cdot c^{14 \cdot M \cdot d \cdot l + 14} \cdot \epsilon)^3, 1\} \cdot \epsilon \\ & \quad + 25 \cdot (|\beta| + 1) \cdot \|z_0\|_\infty \cdot \max\{c_{33} \cdot n \cdot d_{model} \cdot c^{14 \cdot M \cdot d \cdot l + 14} \cdot \epsilon, 1\} \cdot (c_{33} \cdot n \cdot d_{model} \\ & \quad \cdot c^{14 \cdot M \cdot d \cdot l + 14} \cdot \epsilon) \\ & \leq c_{36} \cdot n^5 \cdot (d_{model})^5 \cdot c^{59 \cdot M \cdot d \cdot l + 57} \cdot \epsilon. \end{aligned}$$

Application of Lemma 10 (where the input is bounded in absolute value by $c_{37} \cdot c^{M \cdot d \cdot l}$) yields that after the second pointwise feedforward neural network the error is bounded above by

$$c_{38} \cdot n \cdot c^{M \cdot d \cdot l} \cdot d_{model} \cdot n^5 \cdot d_{model} \cdot c^{59 \cdot M \cdot d \cdot l + 57} \cdot \epsilon \leq c_{38} \cdot n^{6^1} \cdot (d_{model})^{6^1} \cdot c^{8^2 \cdot (M \cdot d \cdot l + 1)} \cdot \epsilon =: \delta_2.$$

Arguing recursively with value

$$B_r = c_{39,r} \cdot n^{2 \cdot 6^{r-1}} \cdot (d_{model})^{2 \cdot 6^{r-1}} \cdot c^{4 \cdot 8^r \cdot (M \cdot d \cdot l + 1)}$$

on level $r \in \{1, \ldots, M \cdot (d \cdot l)\}$ we see that after level $r$ the error of the output is at most

$$\delta_r = c_{40,r} \cdot n^{6^r} \cdot (d_{model})^{6^r} \cdot c^{8^{r+1} \cdot (M \cdot d \cdot l + 1)} \cdot \epsilon.$$

The last pair of attention head and pointwise feedforward neural network in level $M \cdot (d \cdot l) + 1$, where the entries in all matrices are bounded by a constant and all entries in all matrices in the attention head are all close to zero, increases this error at most by a factor

$$\|z_{M \cdot (d \cdot l)}\|_\infty \cdot c_{41} \cdot h \cdot d_{model}$$

(cf., Step 2 in the proof of Lemma 10), which implies that the error of the output of our transformer network is bounded by

$$c_{42} \cdot n^{6^{M \cdot d \cdot l} + 1} \cdot (d_{model})^{6^{M \cdot d \cdot l} + 1} \cdot c^{8^{M \cdot d \cdot l + 2} \cdot (M \cdot d \cdot l + 1)} \cdot \epsilon.$$

□

Next we show how we can approximate a function which satisfies a hierarchical composition model by a Transformer encoder. In order to formulate this result, we introduce some additional notation. In order to compute a function $h_1^{(\kappa)} \in \mathcal{H}(\kappa, \mathcal{P})$ one has to compute different hierarchical composition models of some level $i$ ($i \in \{1, \dots, \kappa - 1\}$). Let $\tilde{N}_i$ denote the number of hierarchical composition models of level $i$, needed to compute $h_1^{(\kappa)}$. Let

$$h_j^{(i)} : \mathbb{R}^{d \cdot l} \to \mathbb{R} \tag{56}$$

be the $j$–th hierarchical composition model of some level $i$ ($j \in \{1, \dots, \tilde{N}_i\}, i \in \{1, \dots, \kappa\}$), that applies a $(p_j^{(i)}, C)$–smooth function $g_j^{(i)} : \mathbb{R}^{K_j^{(i)}} \to \mathbb{R}$ with $p_j^{(i)} = q_j^{(i)} + s_j^{(i)}$, $q_j^{(i)} \in \mathbb{N}_0$ and $s_j^{(i)} \in (0, 1]$, where $(p_j^{(i)}, K_j^{(i)}) \in \mathcal{P}$ (and $K_j^{(1)} = d \cdot l$ ($j = 1, \dots, \tilde{N}_1$)). With this notation we can describe the computation of $h_1^{(\kappa)}(\mathbf{x})$ recursively as follows:

$$h_j^{(i)}(\mathbf{x}) = g_j^{(i)} \left( h^{(i-1)}_{\sum_{t=1}^{j-1} K_t^{(i)} + 1}(\mathbf{x}), \dots, h^{(i-1)}_{\sum_{t=1}^{j} K_t^{(i)}}(\mathbf{x}) \right) \tag{57}$$

holds for $j \in \{1, \dots, \tilde{N}_i\}$ and $i \in \{2, \dots, \kappa\}$, and

$$h_j^{(1)}(\mathbf{x}) = g_j^{(1)} \left( x_{\pi(\sum_{t=1}^{j-1} K_t^{(1)} + 1)}, \dots, x_{\pi(\sum_{t=1}^{j} K_t^{(1)})} \right) \tag{58}$$

holds for $j \in \{1, \dots, \tilde{N}_1\}$ for some function $\pi : \{1, \dots, \tilde{N}_1\} \to \{1, \dots, d\}$. Here the recursion

$$\tilde{N}_l = 1 \text{ and } \tilde{N}_i = \sum_{j=1}^{\tilde{N}_{i+1}} K_j^{(i+1)} \quad (i \in \{1, \dots, \kappa - 1\}) \tag{59}$$

holds.

**Theorem 3** *Let $\tau \in \{l+1, l+2, \dots, l+d+1\}$. Let $A \geq 1$, let $m : \mathbb{R}^{d \cdot l} \to \mathbb{R}$ be contained in the class $\mathcal{H}(\kappa, \mathcal{P})$ for some $\kappa \in \mathbb{N}$ and $\mathcal{P} \subseteq [1, \infty) \times \mathbb{N}$. Let $\tilde{N}_i$ be defined as in (59). Each $m$ consists of different functions $h_j^{(i)}$ ($j \in \{1, \dots, \tilde{N}_i\}$, $i \in \{1, \dots, \kappa\}$) defined as in (56), (57) and (58). Assume that the corresponding functions $g_j^{(i)}$ are Lipschitz continuous with Lipschitz constant $C_{Lip} \geq 1$ and satisfy*

$$\|g_j^{(i)}\|_{C^{q_j^{(i)}}(\mathbb{R}^{K_j^{(i)}})} \leq c_{43}$$

*for some constant $c_{43} > 0$. Denote by $K_{max} = \max_{i,j} K_j^{(i)} < \infty$ the maximal input dimension and set $q_{max} = \max_{i,j} q_j^{(i)} < \infty$, where $q_j^{(i)}$ is the integer part of the smoothness $p_j^{(i)}$ of $g_j^{(i)}$. Let $A \geq 1$. Choose $h \in \mathbb{N}$ such that*

$$c_{44} \leq h \leq n \tag{60}$$

*holds for $n$ large and for some sufficiently large constant $c_{44}$, choose*

$$I \geq \sum_{i=1}^{\kappa} \tilde{N}_i + d + l + 4 \quad \text{and} \quad d_{key} \geq 3$$

*and set*

$$N = I \cdot (q_{max} \cdot K_{max} + 1), \quad d_{model} = h \cdot I, \quad d_v = I.$$

*Then there exists a transformer network $f_\vartheta$, where the matrices in the attention heads have in each row at most nonzero $\tau$ entries, where all matrices $W_{r,1}$ and $W_{r,2}$ in the pointwise feedforward neural networks have the property that in each row of $W_{r,1}$ and in each column of $W_{r,2}$ there are at most $\tau$ nonzero components, and where all parameters are bounded in absolute value by $c_{45} \cdot n^{c_{46}}$ provided $c_{45}, c_{46} > 0$ are sufficiently large, such that for each Transformer network $f_{\tilde{\vartheta}}$ which has the same structure and which weights are in supremum norm not further away than*

$$0 \leq \epsilon \leq \frac{1}{c_{47}}$$

*from the weights of this network for some suitable large constant $c_{47} \geq 1$ and where the matrices $\tilde{W}_{query,r,s}$, $\tilde{W}_{key,r,s}$ and $\tilde{W}_{value,r,s}$ satisfy the assumptions of Lemma 9, satisfies for $n$ large*

$$\begin{aligned} \|f_{\tilde{\vartheta}} - m\|_{\infty,[-A,A]^{d \cdot l}} \quad &\leq \quad c_{48} \cdot (K_{max}+1)^\kappa \cdot \max_{j,i} h^{-p_j^{(i)}/K_j^{(i)}} \\ &\qquad + c_{49} \cdot n^{(I+1)\cdot(q_{max} \cdot K_{max}+2)} \cdot d_{model}^{(I+1)\cdot(q_{max}\cdot K_{max}+1)} \cdot \epsilon \end{aligned}$$

*provided $\epsilon \geq 0$ satisfies*

$$\epsilon \leq \frac{1}{2 \cdot c_{50} \cdot n^{8^{(\sum_{i=1}^{\kappa} \tilde{N}_s)\cdot(2\cdot q_{max}\cdot K_{max}+4)}} \cdot d_{model}^{8^{(\sum_{i=1}^{\kappa} \tilde{N}_s)\cdot(2\cdot q_{max}\cdot K_{max}+4)}}}. \tag{61}$$

**Proof.** From the Lipschitz continuity of the $g_j^{(i)}$ and the recursive definition of the $h_j^{(i)}$ we conclude that there exists $1 \leq \bar{A} \leq c_{51} \cdot A$ such that

$$h_j^{(i)}(x) \in [-\bar{A}, \bar{A}] \tag{62}$$

holds for all $x \in [-A, A]^{d \cdot l}$, $j \in \{1, \dots, \tilde{N}_i\}$ and $i \in \{1, \dots, \kappa - 1\}$.

Our transformer encoder successively approximates $h_1^{(1)}(x)$, $\dots$, $h_{N_1}^{(1)}(x)$, $h_1^{(2)}(x)$, $\dots$, $h_{N_2}^{(2)}(x)$, $\dots$, $h_1^{(\kappa)}(x)$ and saves the computed values successively in $z_{r,1}^{(d+l+5)}$, $z_{r,1}^{(d+l+6)}$, $\dots$, $z_{r,1}^{(d+l+4+\sum_{i=1}^{\kappa} \tilde{N}_i)}$. Here $h_i^{(j)}$ is approximated by computing in a first step a truncated power basis of a tensor product spline space of degree $q_i^{(j)}$ on an equidistant grid in

$$[-\bar{A} - 1, \bar{A} + 1]^{K_i^{(j)}}$$

consisting of $h-1$ basis functions, which are evaluated at the arguments of $h_i^{(j)}$ in (58), and by using in a second step a linear combination of these basis functions to approximate

$$g_j^{(i)}\left(h^{(i-1)}_{\sum_{t=1}^{j-1}K_t^{(i)}+1}(\mathbf{x}),\dots,h^{(i-1)}_{\sum_{t=1}^{j}K_t^{(i)}}(\mathbf{x})\right).$$

The approximate computation of this truncated power basis can be done as in Lemma 11 using layers $(\tilde{N}_{i-1}+j-1)\cdot(q_{max}\cdot K_{max}+1)+1$ till $(\tilde{N}_{i-1}+j)\cdot(q_{max}\cdot K_{max}+1)$ of our transformer encoder. Here the computed values of this basis will have an error not exceeding

$$c_{51}\cdot n^{8^{(\sum_{s=1}^{i-1}\tilde{N}_s+j)\cdot(2\cdot q_{max}\cdot K_{max}+4)}}\cdot d_{model}^{8^{(\sum_{s=1}^{i-1}\tilde{N}_s+j)\cdot(2\cdot q_{max}\cdot K_{max}+4)}}\cdot\epsilon.$$

Using standard approximation results from spline theory (cf., e.g., Theorem 15.2 and proof of Theorem 15.1 in Györfi et al. (2002) and Lemma 1 in Kohler (2014)) and the Lipschitz continuity of $g_j^{(i)}$ this results in an approximation

$$\tilde{g}_j^{(i)}$$

of $g_j^{(i)}$ which satisfies

$$\begin{aligned}
&\|\tilde{g}_j^{(i)}-g_j^{(i)}\|_{\infty,[-\bar{A}-1,\bar{A}+1]^{K_j^{(i)}}} &(63)\\
&\leq c_{52}\cdot h^{-p_j^{(i)}/K_j^{(i)}}+c_{51}\cdot h\cdot n^{8^{(\sum_{s=1}^{i-1}\tilde{N}_s+j)\cdot(2\cdot q_{max}\cdot K_{max}+4)}}\cdot d_{model}^{8^{(\sum_{s=1}^{i-1}\tilde{N}_s+j)\cdot(2\cdot q_{max}\cdot K_{max}+4)}}\cdot\epsilon\\
&\leq c_{52}\cdot h^{-p_j^{(i)}/K_j^{(i)}}+c_{51}\cdot n^{8^{(\sum_{s=1}^{i-1}\tilde{N}_s+j)\cdot(2\cdot q_{max}\cdot K_{max}+4)}}\cdot d_{model}^{8^{(\sum_{s=1}^{i-1}\tilde{N}_s+j)\cdot(2\cdot q_{max}\cdot K_{max}+4)}}\cdot\epsilon.
\end{aligned}$$

The approximation $\tilde{h}_1^{(\kappa)}(\mathbf{x})$ of $h_1^{(\kappa)}(\mathbf{x})$ which our transformer encoder computes is defined as follows:

$$\tilde{h}_j^{(1)}(\mathbf{x})=\tilde{g}_j^{(1)}\left(x_{\pi(\sum_{t=1}^{j-1}K_t^{(1)}+1)},\dots,x_{\pi(\sum_{t=1}^{j}K_t^{(1)})}\right)$$

for $j\in\{1,\dots,\tilde{N}_1\}$ and

$$\tilde{h}_j^{(i)}(\mathbf{x})=\tilde{g}_j^{(i)}\left(\tilde{h}^{(i-1)}_{\sum_{t=1}^{j-1}K_t^{(i)}+1}(\mathbf{x}),\dots,\tilde{h}^{(i-1)}_{\sum_{t=1}^{j}K_t^{(i)}}(\mathbf{x})\right)$$

for $j\in\{1,\dots,\tilde{N}_i\}$ and $i\in\{2,\dots,\kappa\}$.

Assume that (62) holds. From (61), (62) and (63) we conclude

$$|\tilde{h}_j^{(i)}(\mathbf{x})|\leq|\tilde{h}_j^{(i)}(\mathbf{x})-h_j^{(i)}(\mathbf{x})|+|h_j^{(i)}(\mathbf{x})|\leq\bar{A}+1.$$

Consequently we get from (63) for $n$ sufficiently large

$$\begin{aligned}
&|\tilde{h}_j^{(i)}(\mathbf{x})-h_j^{(i)}(\mathbf{x})|\\
&\leq|\tilde{g}_j^{(i)}\left(\tilde{h}^{(i-1)}_{\sum_{t=1}^{j-1}K_t^{(i)}+1}(\mathbf{x}),\dots,\tilde{h}^{(i-1)}_{\sum_{t=1}^{j}K_t^{(i)}}(\mathbf{x})\right)-g_j^{(i)}\left(\tilde{h}^{(i-1)}_{\sum_{t=1}^{j-1}K_t^{(i)}+1}(\mathbf{x}),\dots,\tilde{h}^{(i-1)}_{\sum_{t=1}^{j}K_t^{(i)}}(\mathbf{x})\right)|
\end{aligned}$$

$$
\begin{aligned}
&+|g_j^{(i)}\left(\tilde{h}^{(i-1)}_{\sum_{t=1}^{j-1}K_t^{(i)}+1}(\mathbf{x}),\dots,\tilde{h}^{(i-1)}_{\sum_{t=1}^{j}K_t^{(i)}}(\mathbf{x})\right)-g_j^{(i)}\left(h^{(i-1)}_{\sum_{t=1}^{j-1}K_t^{(i)}+1}(\mathbf{x}),\dots,h^{(i-1)}_{\sum_{t=1}^{j}K_t^{(i)}}(\mathbf{x})\right)|\\
\leq\; & c_{52}\cdot h^{-p_j^{(i)}/K_j^{(i)}}+c_{51}\cdot n^{8^{(\sum_{s=1}^{i-1}\tilde{N}_s+j)\cdot(2\cdot q_{max}\cdot K_{max}+4)}}\cdot d_{model}^{8^{(\sum_{s=1}^{i-1}\tilde{N}_s+j)\cdot(2\cdot q_{max}\cdot K_{max}+4)}}\cdot\epsilon\\
&+|g_j^{(i)}\left(\tilde{h}^{(i-1)}_{\sum_{t=1}^{j-1}K_t^{(i)}+1}(\mathbf{x}),\dots,\tilde{h}^{(i-1)}_{\sum_{t=1}^{j}K_t^{(i)}}(\mathbf{x})\right)-g_j^{(i)}\left(h^{(i-1)}_{\sum_{t=1}^{j-1}K_t^{(i)}+1}(\mathbf{x}),\dots,h^{(i-1)}_{\sum_{t=1}^{j}K_t^{(i)}}(\mathbf{x})\right)|\\
\leq\; & c_{52}\cdot h^{-p_j^{(i)}/K_j^{(i)}}+c_{51}\cdot n^{8^{(\sum_{s=1}^{i-1}\tilde{N}_s+j)\cdot(2\cdot q_{max}\cdot K_{max}+4)}}\cdot d_{model}^{8^{(\sum_{s=1}^{i-1}\tilde{N}_s+j)\cdot(2\cdot q_{max}\cdot K_{max}+4)}}\cdot\epsilon\\
&+c_{53}\cdot\sum_{s=1}^{K_j^{(i)}}|\tilde{h}^{(i-1)}_{\sum_{t=1}^{j-1}K_t^{(i)}+s}(\mathbf{x})-h^{(i-1)}_{\sum_{t=1}^{j-1}K_t^{(i)}+s}(\mathbf{x})|,
\end{aligned}
$$

where the last inequality follows from the Lipschitz continuity of $g_j^{(i)}$. Together with

$$
|\tilde{h}_j^{(1)}(\mathbf{x})-h_j^{(1)}(\mathbf{x})|\leq c_{54}\cdot h^{-p_j^{(1)}/K_j^{(1)}}+c_{55}\cdot n^{8^{j\cdot(2\cdot q_{max}\cdot K_{max}+4)}}\cdot d_{model}^{8^{j\cdot(2\cdot q_{max}\cdot K_{max}+4)}}\cdot\epsilon,
$$

which follows again from (63), an easy induction shows

$$
\begin{aligned}
|\tilde{h}_1^{(\kappa)}(\mathbf{x})-h_1^{(\kappa)}(\mathbf{x})| \;\leq\; & c_{56}\cdot(K_{max}+1)^{\kappa}\cdot\max_{j,i}h^{-p_j^{(i)}/K_j^{(i)}}\\
&+c_{57}\cdot n^{8^{(\sum_{s=1}^{\kappa}\tilde{N}_s)\cdot(2\cdot q_{max}\cdot K_{max}+4)}}\cdot d_{model}^{8^{(\sum_{s=1}^{\kappa}\tilde{N}_s)\cdot(2\cdot q_{max}\cdot K_{max}+4)}}\cdot\epsilon.
\end{aligned}
$$

□

**Remark 6** *It follows from the proof of Theorem 3 (i.e., in particular from the proof of Lemma 9) that even if $\epsilon$ does not satisfy (63) then the all maximal attentions are attended at some data-independent indices provided $0\leq\epsilon\leq 1/c_{58}$ holds.*

**Lemma 12** *Set*

$$
f(z)=\begin{cases}\infty & ,z=1\\ \log\frac{z}{1-z} & ,0<z<1\\ -\infty & ,z=0,\end{cases}
$$

*let $K\in\mathbb{N}$ with $K\geq 6$ and let $A\geq 1$. Let $m:\mathbb{R}^{d\cdot l}\to[0,1]$ and let $\bar{g}:\mathbb{R}^{d\cdot l}\to\mathbb{R}$ such that $\|\bar{g}-m\|_{\infty,[-A,A]^{d\cdot l}}\leq\epsilon$ for some*

$$
0\leq\epsilon\leq\frac{1}{K}.
$$

*Then there exists a neural network $\bar{f}:\mathbb{R}\to\mathbb{R}$ with ReLU activation function, and one hidden layer with $3\cdot K+9$ neurons, where all the weights are bounded in absolute value by $K$, such that for each network $\tilde{f}:\mathbb{R}\to\mathbb{R}$ which has the same structure and which has weights which are in supremum norm not more than*

$$
0\leq\bar{\epsilon}\leq 1
$$

*away from the weights of the above network we have*

$$\sup_{x\in[-A,A]^{d\cdot l}} \Bigg( \Big| m(x) \cdot \Big( \varphi(\tilde{f}(\bar{g}(x)) - \varphi(f(m(x)) \Big) \Big| + \Big| (1-m(x)) \cdot \Big( \varphi(-\tilde{f}(\bar{g}(x))) - \varphi(-f(m(x)) \Big) \Big| \Bigg)$$
$$\leq c_{58} \cdot \left( \frac{\log K}{K} + \epsilon \right) + 132 \cdot K^2 \cdot \bar{\epsilon}.$$

**Proof.** In the *first part part of the proof* we show the assertion in case $\bar{\epsilon} = 0$.

For $k \in \{-1, 0, \dots, K+1\}$ define

$$B_k(z) = \begin{cases} 0 & , z < \frac{k-1}{K} \\ K \cdot (z - \frac{k-1}{K}) & , \frac{k-1}{K} \leq z < \frac{k}{K} \\ K \cdot (\frac{k+1}{K} - z) & , \frac{k}{K} \leq z < \frac{k+1}{K} \\ 0 & , z \geq \frac{k+1}{K}, \end{cases}$$

(which implies $B_k(k/K) = 1$ and $B_k(j/K) = 0$ for $j \in \mathbb{Z} \setminus \{k\}$) and set

$$\begin{aligned} \bar{f}(z) &= f(1/K) \cdot (B_{-1}(z) + B_0(z)) + \sum_{k=1}^{K-1} f(k/K) \cdot B_k(z) + f(1-1/K) \cdot (B_K(z) + B_{K+1}(z)) \\ &=: \sum_{k=-1}^{K+1} a_k \cdot B_k(z). \end{aligned}$$

Then $\bar{f}$ interpolates the points $(-1/K, f(1/K))$, $(0, f(1/K))$, $(1/K, f(1/K))$, $(2/K, f(2/K))$, $\dots$, $((K-1)/K, f((K-1)/K))$, $(1, f((K-1)/K))$ and $(1+1/K, f((K-1)/K))$, is zero outside of $(-2/K, 1+2/K)$ and is linear on each interval $[k/K, (k+1)/K]$ $(k \in \{-2, \dots, K+1\})$. Because of

$$B_k(z) = \sigma\left(K \cdot \left(z - \frac{k-1}{K}\right)\right) - 2 \cdot \sigma\left(K \cdot \left(z - \frac{k}{K}\right)\right) + \sigma\left(K \cdot \left(z - \frac{k+1}{K}\right)\right),$$

$\bar{f}$ can be computed by a neural network with ReLU activation function and one hidden layer with $3 \cdot (K+3) = 3 \cdot K + 9$ neurons. Set

$$h_1(z) = \varphi(f(z)) = \log\left(1 + \exp\left(-\log\frac{z}{1-z}\right)\right) = \log\left(1 + \frac{1-z}{z}\right) = -\log z$$

and

$$h_2(z) = \varphi(-f(z)) = \log\left(1 + \exp\left(\log\frac{z}{1-z}\right)\right) = -\log(1-z).$$

First we consider the case $m(\mathbf{x}) \in [0, 2/K]$, which implies

$$f(m(\mathbf{x})) \leq f(2/K) = -\log(K/2 - 1) < 0.$$

In this case we have $-1/K \le \bar{g}(\mathbf{x}) \le 3/K$ and

$$-\log(K-1) = f(1/K) \le \bar{f}(\bar{g}(\mathbf{x})) \le f(3/K) = -\log(K/3-1)$$

(where we have used that $\bar{f}$ is monotone increasing and satisfies $\bar{f}(-\frac{1}{K}) = f(\frac{1}{K})$ and $\bar{f}(\frac{3}{K}) = f(\frac{3}{K})$). Consequently we get

$$\begin{aligned}\left|m(\mathbf{x}) \cdot \left(\varphi(\bar{f}(\bar{g}(\mathbf{x})) - \varphi(f(m(\mathbf{x}))\right)\right| &\le m(\mathbf{x}) \cdot \varphi(\bar{f}(\bar{g}(\mathbf{x})) + m(\mathbf{x}) \cdot h_1(m(\mathbf{x})) \\ &\le \frac{2}{K} \cdot \log(1+\exp(\log(K-1))) + m(\mathbf{x}) \cdot \log(\frac{1}{m(\mathbf{x})}) \\ &\le 4 \cdot \frac{\log K}{K}\end{aligned}$$

(where we have used the inequality $z \cdot \log(1/z) \le (2/K) \cdot \log(K/2)$ for $0 < z < 2/K$) and

$$\begin{aligned}&\left|(1-m(\mathbf{x})) \cdot \left(\varphi(-\bar{f}(\bar{g}(\mathbf{x})) - \varphi(-f(m(\mathbf{x}))\right)\right| \\ &\le \varphi(-\bar{f}(\bar{g}(\mathbf{x})) + \varphi(-f(m(\mathbf{x})) \\ &= \log(1+\exp(\bar{f}(\bar{g}(\mathbf{x})))) + \log(1+\exp(f(m(\mathbf{x})))) \\ &\le \log(1+\exp(-\log(K/3-1))) + \log(1+\exp(-\log(K/2-1))) \\ &\le 2 \cdot \exp(-\log(K/3-1)) = \frac{6}{K-3}.\end{aligned}$$

Similarly we get in case $m(\mathbf{x}) \ge 1 - 2/K$

$$\begin{aligned}&\left|m(\mathbf{x}) \cdot \left(\varphi(\bar{f}(\bar{g}(\mathbf{x})) - \varphi(f(m(\mathbf{x}))\right)\right| + \left|(1-m(\mathbf{x})) \cdot \left(\varphi(-\bar{f}(\bar{g}(\mathbf{x})) - \varphi(-f(m(\mathbf{x}))\right)\right| \\ &\le 12 \cdot \frac{\log K}{K-3}.\end{aligned}$$

Hence it suffices to show

$$\begin{aligned}\sup_{\substack{\mathbf{x} \in \mathbb{R}^d, \\ m(\mathbf{x}) \in [2/K, 1-2/K]}} &\Bigg( \left|m(\mathbf{x}) \cdot \left(\varphi(\bar{f}(\bar{g}(\mathbf{x})) - \varphi(f(m(\mathbf{x}))\right)\right| \\ &\quad + \left|(1-m(\mathbf{x})) \cdot \left(\varphi(-\bar{f}(\bar{g}(\mathbf{x})) - \varphi(-f(m(\mathbf{x}))\right)\right| \Bigg) \\ &\le c_{59} \cdot (\frac{\log K}{K} + \epsilon).\end{aligned}$$

By the monotonicity of $f$, $|f'(z)| = \frac{1}{z \cdot (1-z)} \ge 1$ for $z \in (0,1)$, the mean value theorem and the definition of $\bar{f}$ we conclude that for any $\mathbf{x} \in \mathbb{R}^{d \cdot l}$ with $m(\mathbf{x}) \in [2/K, 1-2/K]$ we find $\xi_x, \delta_{\mathbf{x}} \in \mathbb{R}$ with $|\xi_x| \le \frac{1}{K}$, $|\delta_{\mathbf{x}}| \le \frac{1}{K} + \epsilon$ and $m(\mathbf{x}) + \delta_{\mathbf{x}} \in [1/K, 1-1/K]$ such that

$$\bar{f}(\bar{g}(\mathbf{x})) = f(\bar{g}(\mathbf{x}) + \xi_x) = f(m(\mathbf{x}) + \delta_{\mathbf{x}}). \tag{64}$$

This implies

$$\sup_{\substack{\mathbf{x} \in \mathbb{R}^d, \\ m(\mathbf{x}) \in [2/K, 1-2/K]}} \Bigg( \left|m(\mathbf{x}) \cdot \left(\varphi(\bar{f}(\bar{g}(\mathbf{x})) - \varphi(f(m(\mathbf{x}))\right)\right|$$

$$
\left. + \left| (1 - m(\mathbf{x})) \cdot \left( \varphi(-\bar{f}(\bar{g}(\mathbf{x})) - \varphi(-f(m(\mathbf{x}))) \right| \right)
$$

$$
= \sup_{\substack{\mathbf{x} \in \mathbb{R}^d, \\ m(\mathbf{x}) \in [2/K, 1-2/K]}} \Bigg( |m(\mathbf{x})| \cdot |h_1(m(\mathbf{x}) + \delta_{\mathbf{x}}) - h_1(m(\mathbf{x}))|
$$

$$
+ |1 - m(\mathbf{x})| \cdot |h_2(m(\mathbf{x}) + \delta_{\mathbf{x}}) - h_2(m(\mathbf{x}))| \Bigg).
$$

Consequently it suffices to show that there exist constants $c_{60}, c_{61} > 0$ such that we have for any $z \in [2/K, 1-2/K]$ and any $\delta \in \mathbb{R}$ with $|\delta| \leq \frac{1}{K} + \epsilon$ and $z + \delta \in [1/K, 1-1/K]$

$$
|z| \cdot |h_1(z + \delta) - h_1(z)| \leq c_{60} \cdot \left( \frac{1}{K} + \epsilon \right) \tag{65}
$$

and

$$
|1 - z| \cdot |h_2(z + \delta) - h_2(z)| \leq c_{61} \cdot \left( \frac{1}{K} + \epsilon \right). \tag{66}
$$

Obviously

$$
h_1'(z) = -\frac{1}{z}.
$$

By the mean value theorem we get for some $\xi \in [\min\{z + \delta, z\}, \max\{z + \delta, z\}]$

$$
|z| \cdot |h_1(z + \delta) - h_1(z)| = |z| \cdot \frac{1}{|\xi|} \cdot |\delta| \leq 4 \cdot |\delta| \leq 4 \cdot \left( \frac{1}{K} + \epsilon \right),
$$

where we have used that $z, z + \delta \in [1/K, 1-1/K]$ and $|\delta| \leq 2/K$ imply $4|\xi| \geq |z|$.

In the same way we get

$$
h_2'(z) = \frac{1}{1 - z}
$$

and

$$
|1 - z| \cdot |h_2(z + \delta) - h_2(z)| = |1 - z| \cdot \frac{1}{|1 - \xi|} \cdot |\delta| \leq 4 \cdot |\delta| \leq 4 \cdot \left( \frac{1}{K} + \epsilon \right).
$$

In the *second step of the proof* we show that if a network $\tilde{f}$ has the same structure as the network $\bar{f}$ in the first step of the proof and if the supremum norm distance between the weights of $\tilde{f}$ and $\bar{f}$ is at most $\epsilon$, then we have:

$$
\sup_{x \in \mathbb{R}^{d \cdot l}} \left| \tilde{f}(\bar{g}(x) - \bar{f}(\bar{g}(x)) \right| \leq 11 \cdot (3 \cdot K + 9) \cdot K \cdot \bar{\epsilon}.
$$

Let

$$
f(z) = \sum_{j=1}^{J_n} v_j^{(1)} \cdot \sigma \left( v_{j,1}^{(0)} \cdot z + v_{j,0}^{(0)} \right)
$$

be a neural network with one hidden layer with $J_n$ neurons, where all the weights are bounded in absolute value by $\beta = K$. It suffices to show that for any $z \in [-1, 2]$ and any network $\tilde{f}$ which has the same structure as $f$ and where the weights are in supremum norm not further away from the weights of $f$ than $\bar{\epsilon}$, it holds

$$|\tilde{f}(z) - f(z)| \leq 11 \cdot \beta \cdot J_n \cdot \bar{\epsilon}.$$

To prove this we observe

$$|\tilde{v}_{i,1}^{(0)} \cdot z + \tilde{v}_{i,0}^{(0)} - (v_{i,1}^{(0)} \cdot z + v_{i,0}^{(0)})| \leq |(\tilde{v}_{i,1}^{(0)} - v_{i,1}^{(0)}) \cdot z| + \bar{\epsilon} \leq 2 \cdot \bar{\epsilon} + \bar{\epsilon} = 3 \cdot \bar{\epsilon},$$

which implies

$$|\sigma(\tilde{v}_{i,1}^{(0)} \cdot \tilde{z} + \tilde{v}_{i,0}^{(0)}) - \sigma(v_{i,1}^{(0)} \cdot z + v_{i,0}^{(0)})| \leq 3 \cdot \bar{\epsilon},$$

and

$$|\sigma(\tilde{v}_{i,1}^{(0)} \cdot \tilde{z} + \tilde{v}_{i,0}^{(0)})| \leq 3 \cdot \bar{\epsilon} + |\sigma(v_{i,1}^{(0)} \cdot z + v_{i,0}^{(0)})| \leq 3 \cdot \bar{\epsilon} + 5\beta \leq 8\beta.$$

Hence we have

$$\begin{aligned}
|\tilde{f}(z) - f(z)| &= \left| \sum_{j=1}^{J_n} \tilde{v}_j^{(1)} \cdot \sigma\left(\tilde{v}_{j,1}^{(0)} \cdot z + \tilde{v}_{j,0}^{(0)}\right) - \sum_{j=1}^{J_n} v_j^{(1)} \cdot \sigma\left(v_{j,1}^{(0)} \cdot z + v_{j,0}^{(0)}\right) \right| \\
&\leq \sum_{j=1}^{J_n} |\tilde{v}_j^{(1)} - v_j^{(1)}| \cdot \sigma\left(\tilde{v}_{j,1}^{(0)} \cdot z + \tilde{v}_{j,0}^{(0)}\right) \\
&\quad + \sum_{j=1}^{J_n} |v_j^{(1)}| \cdot \left| \sigma\left(\tilde{v}_{j,1}^{(0)} \cdot z + \tilde{v}_{j,0}^{(0)}\right) - \sigma\left(v_{j,1}^{(0)} \cdot z + v_{j,0}^{(0)}\right) \right| \\
&\leq 8 \cdot \beta \cdot J_n \cdot \bar{\epsilon} + \beta \cdot J_n \cdot 3 \cdot \bar{\epsilon} \leq 11 \cdot \beta \cdot J_n \cdot \bar{\epsilon},
\end{aligned}$$

which yields the assertion.

Since $\varphi$ is Lipschitz continuous the assertion of Lemma 12 follows from steps 1 and 2. □

**Lemma 13** *Let $A \geq 1$ and let $0 \leq \epsilon \leq 1/(2 \cdot c_{62})$. Let $f_\vartheta$ be a transformer classifier defined as in Section 2 where the weights in all attention units and in all piecewise feedforward networks are in supremum norm not further away than $1/(2 \cdot c_{62})$ from the weights of the transformer network in Theorem 3, and where the weights in the feedfoward network are not further away than $\epsilon$ from the weights of the feedforward neural network in Lemma 12. Let $f_{\tilde{\vartheta}}$ be a transformer classifier of the same form which satisfies*

$$\|\tilde{\vartheta} - \vartheta\|_\infty \leq \epsilon.$$

*Assume $d_{ff} = 2 \cdot h + 2$, $h \leq c_{64} \cdot n$, $d_{model} = h \cdot I$ and $I = \lceil \log n \rceil$. Then we have for $c_{65}, c_{66} > 0$ sufficiently large*

$$\|f_{\tilde{\vartheta}} - f_\vartheta\|_{[-A,A]^{d \cdot l}, \infty} \leq c_{65} \cdot n^{c_{66}} \cdot \|\tilde{\vartheta} - \vartheta\|_\infty.$$

**Proof.** Since the weights in the transformer classifiers $f_\vartheta$ and $f_{\vartheta^*}$ are not further away than $1/c_{63}$ from the weights of the transformer network in Theorem 3, it follows from the proof of Theorem 3 that in both transformer network all maximal attention are attained at the same indices, namely at the indices where the transformer network in Theorem 3 attains its maximal attentions (cf., Remark 6). Consequently we can ignore the selection of the maximal attention in the rest of the proof.

Let

$$q_{k,r-1,s,i} = W_{query,k,r,s} \cdot z_{k,r-1,i}, \quad k_{k,r-1,s,i} = W_{key,k,r,s} \cdot z_{k,r-1,i},$$
$$v_{k,r-1,s,i} = W_{value,k,r,s} \cdot z_{k,r-1,i}, \quad \tilde{q}_{k,r-1,s,i} = \tilde{W}_{query,k,r,s} \cdot \tilde{z}_{k,r-1,i},$$
$$\tilde{k}_{k,r-1,s,i} = \tilde{W}_{key,k,r,s} \cdot \tilde{z}_{k,r-1,i}, \quad \tilde{v}_{k,r-1,s,i} = \tilde{W}_{value,k,r,s} \cdot \tilde{z}_{k,r-1,i}$$

where all the weights in the matrices above are bounded in absolute value by $B \geq 1$, and set

$$y_{k,r,i} = z_{k,r-1,i} + v_{k,r-1,s,r_1} \cdot < q_{k,r-1,s,i}, k_{k,r-1,s,r_1} >$$

and

$$\tilde{y}_{k,r,i} = \tilde{z}_{k,r-1,i} + \tilde{v}_{k,r-1,s,r_1} \cdot < \tilde{q}_{k,r-1,s,i}, \tilde{k}_{k,r-1,s,r_1} > .$$

In the *first step of the proof* we show

$$\begin{aligned}
&\|\tilde{y}_{k,r,i} - y_{k,r,i}\|_\infty \\
&\leq c_{67} \cdot d_{key} \cdot d_{model}^3 \cdot B^2 \cdot (\max\{\|z_{k,r-1,i}\|_\infty, \|\tilde{z}_{k,r-1,i}\|_\infty, 1\})^3 \\
&\qquad \cdot \max\Big\{\|\tilde{W}_{query,k,r-1,s} - W_{query,k,r-1,s}\|_\infty, \|\tilde{W}_{key,k,r-1,s} - W_{key,k,r-1,s}\|_\infty, \\
&\qquad\qquad \|\tilde{W}_{value,k,r-1,s} - W_{value,k,r-1,s}\|_\infty\Big\} \\
&\quad + c_{68} \cdot d_{key} \cdot d_{model}^3 \cdot B^3 \cdot (\max\{\|z_{k,r-1,i}\|_\infty, \|\tilde{z}_{k,r-1,i}\|_\infty, 1\})^2 \cdot \|\tilde{z}_{k,r-1} - z_{k,r-1}\|_\infty.
\end{aligned}$$

We have

$$\begin{aligned}
&\|\tilde{y}_{k,r,i} - y_{k,r,i}\|_\infty \\
&\leq \|\tilde{z}_{k,r,i} - z_{k,r,i}\|_\infty + \|\tilde{v}_{k,r,r_1} - v_{k,r,r_1}\|_\infty \cdot | < \tilde{q}_{k,r,i}, \tilde{k}_{k,r,r_1} > | \\
&\quad + \|v_{k,r,r_1}\|_\infty \cdot | < \tilde{q}_{k,r,i}, \tilde{k}_{k,r,r_1} > - < q_{k,r,i}, k_{k,r,r_1} > |.
\end{aligned}$$

With

$$\begin{aligned}
&\|\tilde{v}_{k,r,r_1} - v_{k,r,r_1}\|_\infty \\
&\leq \|(\tilde{W}_{value,k,r,s} - W_{value,k,r,s}) \cdot \tilde{z}_{k,r-1,r_1}\|_\infty + \|W_{value,k,r,s} \cdot (\tilde{z}_{k,r-1,r_1} - z_{k,r-1,r_1})\|_\infty \\
&\leq d_{model} \cdot \|\tilde{W}_{value,k,r,s} - W_{value,k,r,s}\|_\infty \cdot \|\tilde{z}_{k,r-1,r_1}\|_\infty \\
&\quad + d_{model} \cdot B \cdot \|\tilde{z}_{k,r-1,r_1} - z_{k,r-1,r_1}\|_\infty,
\end{aligned}$$

$$| < \tilde{q}_{k,r,i}, \tilde{k}_{k,r,r_1} > | \leq d_{key} \cdot B^2 \cdot d_{model} \cdot \|\tilde{z}_{k,r-1,r_1}\|_\infty^2,$$

$$\|v_{k,r,i}\|_\infty \leq d_{model} \cdot B \cdot \|z_{k,r-1,i}\|_\infty$$

and

$$
\begin{aligned}
& | < \tilde{q}_{k,r,i}, \tilde{k}_{k,r,r_1} > - < q_{k,r,i}, k_{k,r,r_1} > | \\
& \leq | < \tilde{q}_{k,r,i} - q_{k,r,i}, \tilde{k}_{k,r,r_1} > | + | < q_{k,r,i}, \tilde{k}_{k,r,r_1} - k_{k,r,r_1} > | \\
& \leq d_{key} \cdot \Big( \|(\tilde{W}_{query,k,r,s} - W_{query,k,r,s}) \cdot \tilde{z}_{k,r-1,i}\|_\infty \\
& \qquad\qquad + \|W_{query,k,r,s} \cdot (\tilde{z}_{k,r-1,i} - z_{k,r-1,i})\|_\infty \Big) \cdot \|\tilde{k}_{k,r,r_1}\|_\infty \\
& \quad + d_{key} \cdot \|q_{k,r,i}\|_\infty \cdot \Big( \|(\tilde{W}_{key,k,r,s} - W_{key,k,r,s}) \cdot \tilde{z}_{k,r-1,i}\|_\infty \\
& \qquad\qquad + \|W_{key,k,r,s} \cdot (\tilde{z}_{k,r-1,i} - z_{k,r-1,i})\|_\infty \Big) \\
& \leq d_{key} \cdot (d_{model} \cdot \|\tilde{W}_{query,k,r,s} - W_{query,k,r,s}\|_\infty \cdot \|\tilde{z}_{k,r-1}\|_\infty \\
& \qquad\qquad + d_{model} \cdot B \cdot \|\tilde{z}_{k,r-1} - z_{k,r-1}\|_\infty) \cdot d_{model} \cdot B \cdot \|\tilde{z}_{k,r-1}\|_\infty \\
& \quad + d_{key} \cdot d_{model} \cdot B \cdot \|z_{k,r-1}\|_\infty \cdot (d_{model} \cdot \|\tilde{W}_{key,k,r,s} - W_{key,k,r,s}\|_\infty \cdot \|\tilde{z}_{k,r-1}\|_\infty \\
& \qquad\qquad + d_{model} \cdot B \cdot \|\tilde{z}_{k,r-1} - z_{k,r-1}\|_\infty)
\end{aligned}
$$

we get the assertion.

Set

$$z_{r,s} = y_{r,s} + W_{r,2} \cdot \sigma\left(W_{r,1} \cdot y_{r,s} + b_{r,1}\right) + b_{r,2}$$

and

$$\tilde{z}_{r,s} = \tilde{y}_{r,s} + \tilde{W}_{r,2} \cdot \sigma\left(\tilde{W}_{r,1} \cdot \tilde{y}_{r,s} + \tilde{b}_{r,1}\right) + \tilde{b}_{r,2},$$

where all weights of the neural networks above are bounded in absolute value by $B \geq 1$. In the *second step of the proof* we show

$$
\begin{aligned}
& \|z_{r,s} - \tilde{z}_{r,s}\|_\infty \\
& \leq c_{69} \cdot d_{ff} \cdot d_{model} \cdot B \cdot \max\{\|y_{r,s}\|_\infty, \|\tilde{y}_{r,s}\|_\infty, 1\} \\
& \qquad\qquad \cdot \max\left\{\|\tilde{W}_{r,2} - W_{r,2}\|_\infty, \|\tilde{W}_{r,1} - W_{r,1}\|_\infty, \|\tilde{b}_{r,2} - b_{r,2}\|_\infty, \|\tilde{b}_{r,1} - b_{r,1}\|_\infty\right\} \\
& \quad + c_{70} \cdot d_{ff} \cdot d_{model} \cdot B^2 \cdot \|\tilde{y}_{r,s} - y_{r,s}\|_\infty.
\end{aligned}
$$

We have

$$
\begin{aligned}
& \|z_{r,s} - \tilde{z}_{r,s}\|_\infty \\
& \leq \|y_{r,s} - \tilde{y}_{r,s}\|_\infty + \|\tilde{W}_{r,2} \cdot \sigma\left(\tilde{W}_{r,1} \cdot \tilde{y}_{r,s} + \tilde{b}_{r,1}\right) - W_{r,2} \cdot \sigma\left(W_{r,1} \cdot y_{r,s} + b_{r,1}\right)\|_\infty \\
& \qquad + \|\tilde{b}_{r,2} - b_{r,2}\|_\infty \\
& \leq \|y_{r,s} - \tilde{y}_{r,s}\|_\infty + \|(\tilde{W}_{r,2} - W_{r,2}) \cdot \sigma\left(\tilde{W}_{r,1} \cdot \tilde{y}_{r,s} + \tilde{b}_{r,1}\right)\|_\infty \\
& \quad + \|W_{r,2} \cdot (\sigma\left(\tilde{W}_{r,1} \cdot \tilde{y}_{r,s} + \tilde{b}_{r,1}\right) - \sigma\left(W_{r,1} \cdot y_{r,s} + b_{r,1}\right))\|_\infty + \|\tilde{b}_{r,2} - b_{r,2}\|_\infty \\
& \leq \|y_{r,s} - \tilde{y}_{r,s}\|_\infty + d_{ff} \cdot \|\tilde{W}_{r,2} - W_{r,2}\|_\infty \cdot \|\tilde{W}_{r,1} \cdot \tilde{y}_{r,s} + \tilde{b}_{r,1}\|_\infty \\
& \quad + d_{ff} \cdot B \cdot \|\tilde{W}_{r,1} \cdot \tilde{y}_{r,s} - W_{r,1} \cdot y_{r,s} + \tilde{b}_{r,1} - b_{r,1}\|_\infty + \|\tilde{b}_{r,2} - b_{r,2}\|_\infty.
\end{aligned}
$$

Using

$$\|\tilde{W}_{r,1} \cdot \tilde{y}_{r,s} + \tilde{b}_{r,1}\|_\infty \le d_{model} \cdot B \cdot \|\tilde{y}_{r,s}\|_\infty + B$$

and

$$\begin{aligned}
&\|\tilde{W}_{r,1} \cdot \tilde{y}_{r,s} - W_{r,1} \cdot y_{r,s}\|_\infty \\
&\le \|(\tilde{W}_{r,1} - W_{r,1}) \cdot \tilde{y}_{r,s}\|_\infty + \|W_{r,1} \cdot (\tilde{y}_{r,s} - y_{r,s})\|_\infty \\
&\le d_{model} \cdot \|\tilde{W}_{r,1} - W_{r,1}\|_\infty \cdot \|\tilde{y}_{r,s}\|_\infty + d_{model} \cdot B \cdot \|\tilde{y}_{r,s} - y_{r,s}\|_\infty
\end{aligned}$$

we get the assertion.

Let

$$f(z) = \sum_{j=1}^{J_n} v_j^{(1)} \cdot \sigma\left(v_{j,1}^{(0)} \cdot z + v_{j,0}^{(0)}\right)$$

and

$$\tilde{f}(\tilde{z}) = \sum_{j=1}^{J_n} \tilde{v}_j^{(1)} \cdot \sigma\left(\tilde{v}_{j,1}^{(0)} \cdot \tilde{z} + \tilde{v}_{j,0}^{(0)}\right),$$

where all the weights of the networks above are bounded in absolute value by $B \ge 1$. In the *third part of the proof* we show

$$\begin{aligned}
|\tilde{f}(\tilde{z}) - f(z)| \quad \le \quad & c_{71} \cdot J_n \cdot B \cdot \max\{\tilde{z}, z, 1\} \cdot \max\{|\tilde{v}_j^{(1)} - v_j^{(1)}|, |\tilde{v}_{j,1}^{(0)} - v_{j,1}^{(0)}|, |\tilde{v}_{j,0}^{(0)} - v_{j,0}^{(0)}|\} \\
& + c_{72} \cdot J_n \cdot B^2 \cdot |\tilde{z} - z|.
\end{aligned}$$

We have

$$\begin{aligned}
&|\tilde{f}(\tilde{z}) - f(z)| \\
&\le \sum_{j=1}^{J_n} |\tilde{v}_j^{(1)} \cdot \sigma\left(\tilde{v}_{j,1}^{(0)} \cdot \tilde{z} + \tilde{v}_{j,0}^{(0)}\right) - v_j^{(1)} \cdot \sigma\left(v_{j,1}^{(0)} \cdot z + v_{j,0}^{(0)}\right)| \\
&\le J_n \cdot \max_{j=1,\dots,J_n} \Big(|\tilde{v}_j^{(1)} - v_j^{(1)}| \cdot \sigma\left(\tilde{v}_{j,1}^{(0)} \cdot \tilde{z} + \tilde{v}_{j,0}^{(0)}\right) \\
&\qquad\qquad + |v_j^{(1)}| \cdot |\sigma\left(\tilde{v}_{j,1}^{(0)} \cdot \tilde{z} + \tilde{v}_{j,0}^{(0)}\right) - \sigma\left(v_{j,1}^{(0)} \cdot z + v_{j,0}^{(0)}\right)|\Big).
\end{aligned}$$

With

$$\sigma\left(\tilde{v}_{j,1}^{(0)} \cdot \tilde{z} + \tilde{v}_{j,0}^{(0)}\right) \le |\tilde{v}_{j,1}^{(0)} \cdot \tilde{z} + \tilde{v}_{j,0}^{(0)}| \le B \cdot |\tilde{z}| + B$$

and

$$\begin{aligned}
&|\sigma\left(\tilde{v}_{j,1}^{(0)} \cdot \tilde{z} + \tilde{v}_{j,0}^{(0)}\right) - \sigma\left(v_{j,1}^{(0)} \cdot z + v_{j,0}^{(0)}\right)| \\
&\le |\tilde{v}_{j,1}^{(0)} \cdot \tilde{z} - v_{j,1}^{(0)} \cdot z + \tilde{v}_{j,0}^{(0)} - v_{j,0}^{(0)}| \le |\tilde{v}_{j,1}^{(0)} \cdot (\tilde{z} - z)| + |(\tilde{v}_{j,1}^{(0)} - v_{j,1}^{(0)}) \cdot z| + |\tilde{v}_{j,0}^{(0)} - v_{j,0}^{(0)}| \\
&\le B \cdot |\tilde{z} - z| + |\tilde{v}_{j,1}^{(0)} - v_{j,1}^{(0)}| \cdot |z| + |\tilde{v}_{j,0}^{(0)} - v_{j,0}^{(0)}|
\end{aligned}$$

we get the assertion.

In the *fourth part of the proof* we use the above results to show the assertion of the lemma.

All weights in the above transformer classifiers are bounded in absolute value by

$$c_{73} \cdot n^{c_{74}} + \epsilon \leq 2 \cdot c_{73} \cdot n^{c_{74}} =: B.$$

Because of

$$\|y_{k,r}\|_\infty \leq 2 \cdot d_{ff} \cdot d_{key}^2 \cdot d_{model} \cdot B^3 \cdot \max\{\|z_{k,r-1}\|_\infty^3, 1\},$$
$$\|z_{k,r}\|_\infty \leq 4 \cdot d_{ff} \cdot d_{model} \cdot B^2 \cdot \max\{\|y_{k,r}\|_\infty, 1\}$$

and

$$\|z_{0,r}\|_\infty \leq A$$

an easy induction shows

$$\|y_{k,r}\|_\infty \leq 128^{3^{2\cdot r-2}} \cdot d_{ff}^{4^{2\cdot r-2}} \cdot d_{model}^{4^{2\cdot r-2}} \cdot B^{9^{2\cdot r-2}} \cdot A^{3^r}$$

and

$$\|z_{k,r}\|_\infty \leq 128^{3^{2\cdot r-2}} \cdot d_{ff}^{4^{2\cdot r-2}+1} \cdot d_{model}^{4^{2\cdot r-2}+1} \cdot B^{9^{2\cdot r-2}+2} \cdot A^{3^r}$$

for $r \geq 1$. The same inequalities also hold for $\tilde{y}_{k,r}$ and $\tilde{z}_{k,r}$. This implies

$$\max\{\|y_{k,r}\|_\infty, \|\tilde{y}_{k,r}\|_\infty, \|z_{k,r}\|_\infty, \|\tilde{z}_{k,r}\|_\infty\} \leq c_{75} \cdot n^{c_{76}}$$

for $r \leq N$, where $c_{75} = c_{75}(N), c_{76} = c_{76}(N) > 0$ are finite constants.

Consequently we can conclude from Step 1

$$\begin{aligned}
&\|\tilde{y}_{k,r,i} - y_{k,r,i}\|_\infty \\
&\leq c_{77} \cdot n^{c_{78}} \cdot \max\Big\{\|\tilde{W}_{query,k,r-1,s} - W_{query,k,r-1,s}\|, \|\tilde{W}_{key,k,r-1,s} - W_{key,k,r-1,s}\|, \\
&\qquad \|\tilde{W}_{value,k,r-1,s} - W_{value,k,r-1,s}\|, \|\tilde{z}_{k,r-1} - z_{k,r-1}\|_\infty\Big\},
\end{aligned}$$

from Step 2

$$\begin{aligned}
&\|z_{r,s} - \tilde{z}_{r,s}\|_\infty \\
&\leq c_{79} \cdot n^{c_{80}} \cdot \max\Big\{\|\tilde{W}_{r,2} - W_{r,2}\|_\infty, \|\tilde{W}_{r,1} - W_{r,1}\|_\infty, \|\tilde{b}_{r,2} - b_{r,2}\|_\infty, \\
&\qquad \|\tilde{b}_{r,1} - b_{r,1}\|_\infty, \|\tilde{y}_{r,s} - y_{r,s}\|_\infty\Big\},
\end{aligned}$$

and from Step 3

$$\begin{aligned}
&\|f_{\tilde{\vartheta}} - f_\vartheta\|_{[-A,A]^{d\cdot l},\infty} \\
&\leq c_{81} \cdot n^{c_{82}} \cdot \max\{|\tilde{v}_j^{(1)} - v_j^{(1)}|, |\tilde{v}_{j,1}^{(0)} - v_{j,1}^{(0)}|, |\tilde{v}_{j,0}^{(0)} - v_{j,0}^{(0)}|, \|\tilde{z}_{N,1} - z_{N,1}\|_\infty\}.
\end{aligned}$$

Using these relations recursively we conclude

$$\|f_{\tilde{\vartheta}} - f_\vartheta\|_{[-A,A]^{d\cdot l},\infty} \leq c_{83} \cdot n^{c_{84}}$$

for $c_{83}, c_{84} > 0$ sufficiently large (and depending on $N$). □