Estimation of a regression function from dependent data by over-parametrized deep neural networks learned by gradient descent *

Michael Kohler[1], Adam Krzyżak[2,†] and Vincent Molinero Römer[1]
[1] *Fachbereich Mathematik, Technische Universität Darmstadt, Schlossgartenstr. 7, 64289 Darmstadt, Germany, email: kohler@mathematik.tu-darmstadt.de, roemer@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*

March 31, 2026

Abstract
Estimation of a regression function from exponentially β-mixing data is considered. The L_2 error with integration with respect to the design is used as the error criterion. Deep neural network estimates with logistic activation function are defined, where all parameters are learned by gradient descent. The rate of convergence of the expected L_2 error is analyzed for (p, C)-smooth regression functions. In the special case that the design is concentrated on a d^*-dimensional manifold, it is shown that the expected L_2 error of the estimate achieves a rate of convergence which depends on d^* and not on the dimension d of the design.

AMS classification: Primary 62G08; secondary 62G20.

Key words and phrases: Deep neural networks, dependent data, dimension reduction, gradient descent, over-parametrization, rate of convergence, regression estimation.

1. Introduction

Deep neural networks are nowadays among the most successful statistical methods applied in practice, e.g. in image classification (cf., e.g., Krizhevsky, Sutskever and Hinton (2012)), in language recognition (cf., e.g., Kim (2014)) in machine translation (cf., e.g., Wu et al. (2016)) in game playing (cf., e.g., Silver et al. (2017)) or in simulation of human conversation (cf., e.g., Minaee et al. (2025)). Motivated by these successes in applications, deep learning was also intensively studied theoretically in the past ten years. Often this is done in the context of nonparametric regression, where a sample

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

*Running title: *Estimation of a regression function from dependent data*
†Corresponding author. Tel: +1-514-848-2424 ext. 3007, Fax:+1-514-848-2830

of a $\mathbb{R}^d \times \mathbb{R}$–valued random vector (X, Y) satisfying $\mathbf{E}\{Y^2\} < \infty$ is given, and the task is to construct an estimate

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

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

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

is small (see Chapter 1 in Györfi et al. (2002) for a motivation for using the L_2 error as the error criterion in nonparametric regression).

It is well-known that the rate of convergence of the L_2 error might be arbitrarily slow in case that one does not impose smoothness assumptions on the regression function (cf., e.g., Chapter 3 in Györfi et al. (2002)). In the sequel we will assume that the regression function is (p, C)–smooth according to the following definition.

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

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

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

Stone (1982) showed that the optimal minimax rate of convergence for estimation of (p, C)–smooth regression functions in the case of i.i.d. data is

$$n^{-\frac{2p}{2p+d}}.$$

This rate suffers from the so–called "curse of dimensionality": it becomes slow when d is large relative to p.

If the data $\mathcal{D}_n$ is an i.i.d. sample independent of (X, Y), then least squares estimates based on deep neural networks can achieve very fast rate of convergence results even for large d. The reason behind this is the structure of the deep neural networks which enables us to use approximation results for smooth functions by neural networks (cf., e.g., Yarotsky (2018)) to approximate the composition of smooth functions by deep neural networks. Here dimension reduction occurs in the sense that the rate of convergence of the corresponding least squares regression estimates does not depend on the dimension d of X, but on the maximum number of variables in the individual functions to be composed, and therefore the estimates achieve good rate of convergence results under such a hierarchical composition model for the regression function even for very large d. The first paper showing that deep neural networks can circumvent the 'curse of dimensionality' in the above setting was Kohler and Krzyżak (2017), and later on this was extended by many additional results (see, e.g., Bauer and Kohler (2019), Schmidt-Hieber (2020) and Kohler and Langer (2021)). Another setting where it was shown that

deep neural networks are able to circumvent the curse of dimensionality is manifold case, where it is assumed that X takes on values on some d^* dimensional manifold defined as follows:

Definition 2 *Let $\mathcal{M} \subseteq \mathbb{R}^d$ be compact and let $d^* \in \{1, \dots, d\}$.*
a) *We say that $U_1, \dots, U_r$ is an* open covering of $\mathcal{M}$*, if $U_1, \dots, U_r \subset \mathbb{R}^d$ are open (with respect to the Euclidean topology on $\mathbb{R}^d$) and satisfy*

$$\mathcal{M} \subseteq \bigcup_{l=1}^{r} U_l.$$

b) *We say that*

$$\psi_1, \dots, \psi_r : [0,1]^{d^*} \to \mathbb{R}^d$$

are bi-Lipschitz functions*, if there exists $0 < C_{\psi,1} \leq C_{\psi,2} < \infty$ such that*

$$C_{\psi,1} \cdot \|x_1 - x_2\| \leq \|\psi_l(x_1) - \psi_l(x_2)\| \leq C_{\psi,2} \cdot \|x_1 - x_2\| \tag{2}$$

holds for any $x_1, x_2 \in [0,1]^{d^*}$ and any $l \in \{1, \dots, r\}$.
c) *We say that $\mathcal{M}$ is a d^*-*dimensional Lipschitz-manifold *if there exist* bi-Lipschitz functions $\psi_i : [0,1]^{d^*} \to \mathbb{R}^d$ $(i \in \{1, \dots, r\})$*, and an open covering $U_1, \dots, U_r$ of $\mathcal{M}$ such that*

$$\psi_l\left((0,1)^{d^*}\right) = \mathcal{M} \cap U_l$$

holds for all $l \in \{1, \dots, r\}$. Here we call $\psi_1, \dots, \psi_r$ the parametrizations *of the manifold.*

In this case Kohler, Langer and Reif (2023) and Jiao et al. (2023) showed that deep neural networks are able to achieve a rate of convergence which depends on d^* rather than on d.

In many applications, it is not possible to observe an i.i.d. sample of (X, Y). In the past few years several results were shown which demonstrate that the above results also holds for dependent data sets $\mathcal{D}_n$ which are exponentially β-mixing according to the following definition.

Definition 3 **a)** *Let $N, K \in \mathbb{N}$, let $X : \Omega \to \mathbb{R}^N$ and $Y : \Omega \to \mathbb{R}^K$ be Borel measurable random variables defined on the same probability space. The β-mixing coefficient between X and Y is defined by*

$$\beta(X,Y) = \mathbf{E}\{\operatorname*{ess\,sup}_{C \in \sigma(Y)} |\mathbf{P}(C|X) - \mathbf{P}(C)|\} = \mathbf{E}\{\operatorname*{ess\,sup}_{A \in \mathcal{B}_K} |\mathbf{P}(Y \in A|X) - \mathbf{P}(Y \in A)|\},$$

where $\sigma(Y)$ is the σ-field generated by Y.
b) *A sequence of random variables $X_1, X_2, \dots$ is called β-mixing if*

$$\beta_s(X_1, X_2, \dots) = \sup_{k \in \mathbb{N}} \beta((X_1, \dots, X_k), X_{k+s}) \to 0 \ \ (s \to \infty).$$

c) *We say that* (X_1, Y_1), (X_2, Y_2), *... is exponentially* β*-mixing if there exists constants* $c_1, c_2 > 0$ *such that*

$$\beta_s((X_1, Y_1), (X_2, Y_2), \dots) \leq c_1 \cdot e^{-c_2 \cdot s} \quad (s \in \mathbb{N}),$$

i.e., if

$$\sup_{k \in \mathbb{N}} \mathbf{E}\{\operatorname*{ess\,sup}_{A \in \mathcal{B}_{d+1}} |\mathbf{P}\{(X_{k+s}, Y_{k+s}) \in A | (X_1, Y_1), \dots, (X_k, Y_k)\} - \mathbf{P}\{(X_{k+s}, Y_{k+s}) \in A\}|\}$$
$$\leq c_1 \cdot e^{-c_2 \cdot s} \quad (s \in \mathbb{N}).$$

It was shown in Ma and Safikhani (2022) that least squares estimates based on deep neural networks achieve dimension reduction also in the case of exponentially β–mixing data provided the regression function satisfies a hierarchical composition model. In a time series setting a related result, which covers in addition the case of manifold data, has been obtained by Padilla et al. (2024).

The above least squares estimates based on deep neural networks cannot be computed in applications since the computation of the least squares estimates requires minimization of the empirical risk of the networks, which is a nonlinear and non-convex function of the weights. Instead one computes these estimates approximately by using gradient descent. Here, even in i.i.d case, it is not known whether the corresponding estimates achieve dimension reduction in case that the regression function satisfies a hierarchical composition model or in case that the predictor takes on values on some submanifold of $\mathbb{R}^d$ as long as the size of the network or the number of gradient descent steps are not exponentially growing in the sample size.

In this article we focus on the case where the regression function is (p, C)–smooth. In this situation it was shown in Kohler (2026) that there exists a deep neural network estimate (with logistic activation function) learned by gradient descent, where both the size of the network and the number of gradient descent steps are bounded by a polynomial in the data size n, which satisfies for any $\epsilon > 0$

$$\mathbf{E} \int |m_n(x) - m(x)|^2 \mathbf{P}_X(dx) \leq c_3 \cdot n^{-\frac{2p}{2p+d} + \epsilon} \tag{3}$$

for some $c_3 = c_3(\epsilon) > 0$ in case where the data are i.i.d, that X takes on with probability one values in some bounded set, that the regression function is (p, C)–smooth and that the distribution of Y is subgaussian (cf., (A1) below).

In this article we use the following model for our data: We assume that

$$(X, Y), (X_1, Y_1), (X_2, Y_2), \dots$$

are $\mathbb{R}^d \times \mathbb{R}$-valued random variables defined on the same probability space satisfying

(A1) Y is subgaussian, i.e., $\mathbf{E}\left\{e^{c_4 \cdot Y^2}\right\} < \infty$ for some $c_4 > 0$,

(A2) $\operatorname{supp}(X)$ is bounded,

(A3) $m(x) = \mathbf{E}\{Y|X = x\}$ is (p, C)-smooth for some $p, C > 0$,

(A4) (X, Y), (X_1, Y_1), ... are identically distributed,

(A5) (X, Y) is independent of (X_1, Y_1), (X_2, Y_2), ...,

(A6) (X_1, Y_1), (X_2, Y_2), ... is exponentially β-mixing.

Our main results in this article extend the above-described result from Kohler (2026) in two ways: First, we show that (3) also holds in case that the data is exponentially β-mixing. Second, we consider the case that X takes on values on some d^*-dimensional manifold. In this case we show that our deep neural network regression estimate learned by gradient descent satisfies for any $\epsilon > 0$

$$\mathbf{E} \int |m_n(x) - m(x)|^2 \mathbf{P}_X(dx) \leq c_5 \cdot n^{-\frac{2p}{2p+d^*}+\epsilon}$$

for some $c_5 = c_5(\epsilon) > 0$ in case that the data is exponentially β-mixing, that X takes on with probability one values in some d^*–dimensional manifold, that the regression function is (p, C)–smooth and that the distribution of Y is subexponential.

1.1. Discussion of related results

In recent years, a large number of research studies were devoted to analysis of regression estimates based on deep neural networks and trained by independent, identically distributed (i.i.d.) data. In order to establish theoretical guarantees for these estimates one needs to simultaneously study approximation, estimation and optimization issues.

Approximation abilities of neural network regression estimates have been investigated by Yarotsky (2017, 2018), Yarotsky and Zhevnerchuk (2019), Lu et al. (2020), Langer (2020) and in the literature cited therein. The authors studied expressive powers of shallow and deep neural networks regression estimates with specific activation functions such as ReLU.

Generalization aspects of deep neural network regression estimates have been established by using the classical VC theory to bound the VC dimension of classes of neural networks, see Bartlett et al. (2017), Bartlett et al. (2019) and by means of bounding the Rademacher complexity, see, e.g., Bartlett and Mendelson (2002), Liang, Rakhlin and Sridharan (2015), Golowich, Rakhlin and Shamir (2019), Lin and Zhang (2019) and Wang and Ma (2022).

The asymptotic properties of the neural network regression estimates minimizing least squares have been studied by Kohler and Krzyżak (2017). They obtained the rate of convergence independent of the input dimension within the class of regression functions following the hierarchical interaction model with additional smoothness constraints. These results were later extended to arbitrary smooth functions by Bauer and Kohler (2019). Furthermore, Schmidt-Hieber (2020) proved that the least squares neural network regression estimates with ReLU activations that satisfy certain sparsity constraints achieve the minimax rates of convergence up to logarithmic factors for a more general class of functions called hierarchical composition model. Kohler and Langer (2021) obtained the same

rate for a linear combination of fully connected networks without sparsity constraints. Further extensions of these results were obtained by Imaizumi and Fukamizu (2018), Suzuki (2018) and Suzuki and Nitanda (2019).

Solving least squares problem for deep neural network regression estimation in practice is prohibitively expensive. Alternatively, gradient descent has been established as a method of choice for optimizing the weights in deep neural networks. Zou et al. (2018), Du et al. (2019), Allen-Zhu, Li and Song (2019) and Kawaguchi and Huang (2019) proved that gradient descent applied to over-parameterized deep neural networks yields neural networks globally minimizing the empirical risk, however Kohler and Krzyżak (2021) showed the resulting estimates do not generalize well. Cao and Gu (2019) and Chen et al. (2021) proved generalization ability of the neural network estimates trained by gradient descent, but they did not analyze their approximation abilities.

In order to establish theoretical guarantees for the estimates trained by gradient descent the approximation, generalization and optimization errors have to be studied simultaneously. Such approach has been used by Braun et al. (2024) for shallow (or one hidden layer) neural network and it led to dimension-free rate of convergence for regression functions satisfying Barron (1993, 1994) condition, i.e., for functions whose Fourier transform has a finite first moment. In the case of deep neural networks with multiple hidden layers Kohler and Krzyżak (2025a) used over-parametrized deep neural networks defined as a linear combination of a huge number of deep neural networks computed in parallel with randomly initialized weights and they applied gradient descent to effectively select a subset of the neural networks within the linear combination. They apply metric entropy bounds, see Birman and Solomjak (1967) and Li, Gu and Ding (2021), to establish generalization abilities of over-parametrized neural networks and show the rate of convergence of order close to $n^{-1/(1+d)}$, whereas for interaction models the rate becomes $n^{1/(1+d^*)}$, i.e., it is independent of dimension, where d^* is effective dimension of the interaction model. The above results are valid for (p, C)–smooth regression functions with $p = 1/2$. In Kohler (2026) over-parametrized deep neural neural network estimates learned by gradient descent have been analyzed in the case of a (p, C)–smooth regression function with general p. An interesting by-product of these studies is that learning of inner weights of the deep network regression estimate is not important. E.g., Gonon (2021) does not train inner weights at all, whereas Braun et al. (2024), Kohler and Krzyżak (2025a) and Kohler (2026) use the fact that the relevant inner weights remain close to their starting values during gradient descent and they use gradient descent to only train output weights of L_2 regularized network. Similar approach has also been taken by Andoni et al. (2014), Daniely (2017), Huang, Chen and Siew (2006), Rahimi and Recht (2008a), Rahimi and Recht (2008b) and Rahimi and Recht (2009). Universal consistency of over-parametrized deep learning regression estimates trained by gradient descent has been demonstrated by Drews and Kohler (2024). Statistically guided deep learning has been investigated by Kohler and Krzyżak (2025b), who have established a bound on the expected L_2 error of a deep neural network regression estimate.

There exist alternative approaches for analyzing gradient descent training. One is the neural tangent kernel proposed by Jacot, Gabriel and Hongler (2020) and studied by Hanin and Nica (2019), Wilson et al. (2025) and by Mahowald et al (2026). Another

approach is the mean-field approach, cf., Mei, Montanari and Nguyen (2018), Chizat and Bach (2018) (further extended by Wojtowytsch (2020)) and Nguyen and Pham (2020). A comprehensive survey of over-parametrized deep neural network estimates trained by gradient descent is presented in Bartlett, Montanari and Rakhlin (2021).

Recently introduced deep transformer networks have been very successful in large language models such as Chat GPT and in image and video processing and they became very popular in research and applications. They were introduced by Vaswani et al. (2017) and their approximation and generalization abilities were established by Gurevych, Kohler and Sahin (2022). The rates of convergence of over-parametrized transformer classifiers learned by gradient descent have been obtained by Kohler and Krzyżak (2023).

All results presented above concern the case of independent, identically distributed (i.i.d.) data. In recent years a number of results for deep learning with dependent data have become available. The most common notions of dependence are so-called mixing coefficients. The α-mixing has been introduced by Rosenblatt (1956). The β-mixing has been defined by Kolmogorov, but first appeared in the paper by Wolkonski and Rozanov (1959). Ibragimov (1962) introduced the ϕ-mixing coefficient. Blum, Hanson and Koopmans (1963) introduced the $*$-mixing coefficient. Properties of various dependencies were proven in Doob (1963). A general survey of mixing coefficients and examples is presented by Doukhan (1994). Barrera and Gobet (2021) generalize uniform deviation inequalities for the empirical process from the independent data to dependent β-mixing case. Their results make it possible to analyze errors of the least-squares regression schemes for dependent data. Kengne and Wade (2025a) studied deep learning with weakly dependent data and Kengne and Wade (2025b) considered strongly mixing observations. A general framework for deep learning with dependent data was proposed by Kengne and Wade (2025c). They obtained rates of convergence for dependencies described by C-mixing processes, strong mixing processes and ϕ-mixing processes. Alquier and Kengne (2025) proved minimax optimality of deep neural network estimates with ReLU activation for the least squares regression problem and Markov chain dependence. Ma and Safikhani (2022) established the non-asymptotic bounds for prediction errors of sparse neural networks with ReLU activation functions under various mixing conditions including α-mixing process and autoregressive time-series process. None of the above papers considers deep neural network estimates learned by gradient descent. In the present paper we analyze deep neural network estimates learned by gradient descent in the case of exponential β-mixing data.

1.2. Notation

The sets of natural numbers, real numbers and nonnegative real numbers are denoted by $\mathbb{N}$, $\mathbb{R}$ and $\mathbb{R}_+$, respectively. For $z \in \mathbb{R}$, we denote the smallest integer greater than or equal to z by $\lceil z \rceil$. And the largest integer less than or equal to z is denoted by $\lfloor z \rfloor$. The Euclidean norm of $x \in \mathbb{R}^d$ is denoted by $\|x\|$, and the scalar product of $x, y \in \mathbb{R}^d$ is denoted by $< x, y >$. For $f : \mathbb{R}^d \to \mathbb{R}$

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

is its supremum norm. If $(X_i)_{i\in I}$ is a nonempty family of real valued random variables, we define $\operatorname{ess\,sup}_{i\in I} X_i$ as the (almost surely) unique variable Y which satisfies

(i) $Y \geq X_i$ *a.s.* for every $i \in I$,

(ii) for any variable $\tilde{Y}$ which satisfies (i) we have $Y \leq \tilde{Y}$ *a.s.*

A finite collection $f_1, \ldots, f_N : \mathbb{R}^d \to \mathbb{R}$ is called an L_p ε–covering of $\mathcal{F}$ on x_1^n if for all $f \in \mathcal{F}$

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

hold. The L_p ε–covering number of $\mathcal{F}$ on x_1^n is the size N of the smallest L_p ε–covering of $\mathcal{F}$ on x_1^n and is denoted by $\mathcal{N}_p(\varepsilon, \mathcal{F}, x_1^n)$.

For $z \in \mathbb{R}$ and $\kappa \geq 0$ we define $T_\kappa z = \max\{-\kappa, \min\{\kappa, z\}\}$. If $f : \mathbb{R}^d \to \mathbb{R}$ is a function then we set $(T_\kappa f)(x) = T_\kappa(f(x))$, and if $\mathcal{F}$ is a class of functions $f : \mathbb{R}^d \to \mathbb{R}$ we set $T_\kappa \mathcal{F} = \{T_\kappa f : f \in \mathcal{F}\}$.

1.3. Outline

Section 2 contains the definition of the estimate. The main results are presented in Section 3 and proven in Section 4. Auxiliary results concerning mixing are presented in the appendix.

2. Definition of the estimate

We use the so–called logistic squasher $\sigma(x) = 1/(1 + e^{-x})$ as our activation function throughout the paper. The topology we use is the same as in Kohler (2026) and is defined as follows: For parameters $K_n, L, r \in \mathbb{N}$ we set

$$f_{\mathbf{w}}(x) = \sum_{j=1}^{K_n} w_{1,1,j}^{(L)} \cdot f_{j,1}^{(L)}(x) \tag{4}$$

for some $w_{1,1,1}^{(L)}, \ldots, w_{1,1,K_n}^{(L)} \in \mathbb{R}$, where $f_{j,1}^{(L)} = f_{\mathbf{w},j,1}^{(L)}$ are recursively defined by

$$f_{k,i}^{(l)}(x) = f_{\mathbf{w},k,i}^{(l)}(x) = \sigma\left(\sum_{j=1}^{r} w_{k,i,j}^{(l-1)} \cdot f_{k,j}^{(l-1)}(x) + w_{k,i,0}^{(l-1)} \right) \tag{5}$$

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

$$f_{k,i}^{(1)}(x) = f_{\mathbf{w},k,i}^{(1)}(x) = \sigma\left(\sum_{j=1}^{d} w_{k,i,j}^{(0)} \cdot x^{(j)} + w_{k,i,0}^{(0)} \right) \tag{6}$$

for some $w_{k,i,0}^{(0)}, \dots, w_{k,i,d}^{(0)} \in \mathbb{R}$.
That is, we consider neural networks consisting of K_n parallel fully connected subnetworks of depth L and width r, with the final output given by a linear combination of their outputs. We write $(w_{k,i,j}^{(l)})_{i,j,l}$ for the weights in the k-th subnetwork, where $w_{k,i,j}^{(l)}$ denotes the weight between neuron j in layer l and neuron i in layer $l+1$.

The weights $\mathbf{w}^{(0)} = ((\mathbf{w}^{(0)})_{k,i,j}^{(l)})_{k,i,j,l}$ are first initialized as follows: Set

$$(\mathbf{w}^{(0)})_{1,1,k}^{(L)} = 0 \quad (k = 1, \dots, K_n), \tag{7}$$

choose the weights of layer $l \in \{1, \dots, L-1\}$ $(\mathbf{w}^{(0)})_{k,i,j}^{(l)}$ uniformly distributed on $[-c_6, c_6]$, and choose the weights of layer 0 $(\mathbf{w}^{(0)})_{k,i,j}^{(0)}$ uniformly distributed on

$$[-c_7 \cdot (\log n) \cdot n^{\tau}, c_7 \cdot (\log n) \cdot n^{\tau}],$$

for parameters $c_6, c_7, \tau > 0$. We assume that all components of $\mathbf{w}^{(0)}$ chosen in this way are independent.

Subsequently, we perform $t_n \in \mathbb{N}$ gradient descent steps of step size $\lambda_n > 0$ in an attempt to minimize the empirical L_2 risk

$$F_n(\mathbf{w}) = \frac{1}{n} \sum_{i=1}^{n} |Y_i - f_{\mathbf{w}}(X_i)|^2. \tag{8}$$

Therefore, we set

$$\mathbf{w}^{(t+1)} = \mathbf{w}^{(t)} - \lambda_n \cdot \nabla_{\mathbf{w}} F_n(\mathbf{w}^{(t)}) \quad (t = 0, \dots, t_n - 1). \tag{9}$$

Note that the choice of the weights implies

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

The final estimate m_n is a truncated version of the neural network with weights according to the weight vector $\mathbf{w}^{(t_n)}$, i.e.

$$m_n(x) = T_{\kappa_n}(f_{\mathbf{w}^{(t_n)}}(x)), \tag{10}$$

where $\kappa_n = c_8 \cdot \log n$ $(c_8 > 0)$ and $T_\kappa z = \max\{\min\{z, \kappa\}, -\kappa\}$ for $z \in \mathbb{R}$ and $\kappa > 0$.

3. Main results

Our main results are presented in the following theorem.

Theorem 1 *Assume (A1) - (A6) holds, set $p = q + s$ where $q \in \mathbb{N}_0$ and $s \in (0, 1]$, set $\kappa_n = c_8 \cdot \log n$ for some $c_8 > 0$ and assume $c_8 \cdot c_4 \geq 3$. Define the estimate m_n as in Section 2 and assume $c_6, c_7 > 0$ are sufficiently large.*

a) *Set*

$$L = \lceil \log_2(q+d) \rceil + 1, \quad r = 4 \cdot \lceil (p+d)^2 \rceil \quad \textit{and} \quad \tau = \frac{1}{2p+d}$$

and choose $K_n \in \mathbb{N}$ *such that for some* $c_9 > 0$

$$\frac{K_n}{n^{c_9}} \to 0 \quad (n \to \infty)$$

and

$$\frac{K_n}{n^{((2p+2d)\tau+1.5)\, r((r+1)(L-1)+(d+1))+\tau(r(d+1)+4(p+d))+2}} \to \infty \quad (n \to \infty). \tag{11}$$

Set

$$\lambda_n = \frac{c_{10}}{K_n^{3/2} \cdot \kappa_n} \quad \textit{and} \quad t_n = \left\lceil c_{11} \cdot \frac{K_n^{3/2}}{\kappa_n} \right\rceil$$

for some $c_{10}, c_{11} > 0$. *Then we have for any* $\epsilon > 0$

$$\mathbf{E} \int |m_n(x) - m(x)|^2 \mathbf{P}_X(dx) \le c_{12} \cdot n^{-\frac{2p}{2p+d}+\epsilon}$$

for some $c_{12} = c_{12}(\epsilon) > 0$.

b) *Let* $\mathcal{M} \subseteq \mathbb{R}^d$ *be a* d^**-dimensional Lipschitz-manifold and assume* $\mathrm{supp}(X) \subseteq \mathcal{M}$. *Set* L, r *as in part a) and set*

$$\tau = \frac{1}{2p+d^*}.$$

Choose K_n, λ_n *and* t_n *as in part a) with this new value for* τ. *Then we have for any* $\epsilon > 0$

$$\mathbf{E} \int |m_n(x) - m(x)|^2 \mathbf{P}_X(dx) \le c_{13} \cdot n^{-\frac{2p}{2p+d^*}+\epsilon}$$

for some $c_{13} = c_{13}(\epsilon) > 0$.

Remark 1. We say that (X_1, Y_1), (X_2, Y_2), ... is subexponentially β-mixing if there exists constants $c_{14}, c_{15} > 0$ and $\delta \in (0,1)$ such that

$$\beta_s((X_1,Y_1),(X_2,Y_2),\dots) \le c_{14} \cdot e^{-c_{15} \cdot s^\delta} \quad (s \in \mathbb{N}),$$

i.e., if

$$\sup_{k \in \mathbb{N}} \mathbf{E}\{\operatorname*{ess\,sup}_{A \in \mathcal{B}_{d+1}} |\mathbf{P}\{(X_{k+s}, Y_{k+s}) \in A | (X_1,Y_1),\dots,(X_k,Y_k)\} - \mathbf{P}\{(X_{k+s},Y_{k+s}) \in A\}|\}$$
$$\le c_{14} \cdot e^{-c_{15} \cdot s^\delta} \quad (s \in \mathbb{N}).$$

It follows from the proof of Theorem 1 that the assertion also holds if the sample is subexponential β-mixing. To prove this it suffices to choose

$$N_n = \lceil (\log n)^{2/\delta} \rceil$$

in the proof of Lemma 8.

Remark 2. Let $(X_t)_{t\in\mathbb{N}}$ be a sequence of exponentially β-mixing $\mathbb{R}^d$–valued random variables, let $(\epsilon_t)_{t\in\mathbb{N}}$ be a sequence of independent identically distributed square integrable real-valued random variables with expectation zero which are independent of $(X_t)_{t\in\mathbb{N}}$, let $m : \mathbb{R}^d \to \mathbb{R}$ be a Borel-measurable function such that $\mathbf{E}\{|m(X_t)|^2\} < \infty$, and set

$$Y_t = m(X_t) + \epsilon_t \quad (t \in \mathbb{N}).$$

Then $((X_t, Y_t))_{t\in\mathbb{N}}$ is exponentially β-mixing since

$$\begin{aligned}
&\sup_{k\in\mathbb{N}} \mathbf{E}\{\operatorname*{ess\,sup}_{A\in\mathcal{B}_{d+1}} |\mathbf{P}\{(X_{k+s}, Y_{k+s}) \in A|(X_1, Y_1), \ldots, (X_k, Y_k)\} - \mathbf{P}\{(X_{k+s}, Y_{k+s}) \in A\}|\} \\
&\leq \sup_{k\in\mathbb{N}} \mathbf{E}\{\operatorname*{ess\,sup}_{A\in\mathcal{B}_{d+1}} |\mathbf{P}\{(X_{k+s}, Y_{k+s}) \in A|X_1, \epsilon_1, \ldots, X_k, \epsilon_k\} - \mathbf{P}\{(X_{k+s}, Y_{k+s}) \in A\}|\} \\
&= \sup_{k\in\mathbb{N}} \mathbf{E}\{\operatorname*{ess\,sup}_{A\in\mathcal{B}_{d+1}} |\mathbf{P}\{(X_{k+s}, Y_{k+s}) \in A|X_1 \ldots, X_k\} - \mathbf{P}\{(X_{k+s}, Y_{k+s}) \in A\}|\}
\end{aligned}$$

where the last equality holds because of the independence of $(X_t)_t$ and $(\epsilon_t)_t$, and by using the symmetry of the β-mixing coefficient (cf., e.g., Remark 2.4 in Barrera and Gobet (2021)) we can conclude in the same way that this in turn is bounded by

$$\sup_{k\in\mathbb{N}} \mathbf{E}\{\operatorname*{ess\,sup}_{A\in\mathcal{B}_d} |\mathbf{P}\{X_{k+s} \in A|X_1, \ldots, X_k\} - \mathbf{P}\{X_{k+s} \in A\}|\}.$$

Examples of sequences of exponentially β-mixing $\mathbb{R}^d$–valued random variables can be found, e.g., in Kurisu, Fukami and Koike (2024). A simple example of a exponentially β-mixing sequence of identically distributed real valued random variables is given by the $AR(1)$-model

$$X_{t+1} = \frac{1}{\sqrt{2}} \cdot X_t + \frac{1}{\sqrt{2}} \cdot \bar{\epsilon}_{t+1} \quad (t \in \mathbb{N})$$

where $X_1, \bar{\epsilon}_2, \bar{\epsilon}_3, \ldots$ are independent and identically standard normally distributed random variables.

4. Proofs

In the proof of Theorem 1 we will need results which help us to analyze the optimization error of the estimate, the approximation error of the estimate, and the generalization error of the estimate, which we present before the proof in separate subsections.

4.1. Neural network optimization

Let $d, J_n \in \mathbb{N}$, and for

$$\mathbf{w} = (w_1, \ldots, w_{J_n}) \in \mathbb{R}^{J_n}$$

let $f_{\mathbf{w}} : \mathbb{R}^d \to \mathbb{R}$ be a (deep) neural network with weight vector $\mathbf{w}$ as defined by (4) - (6). Further let

$$f_{lin,\mathbf{w},\tilde{\mathbf{w}}}(x) = f_{\mathbf{w}}(x) + \sum_{j=1}^{J_n} \frac{\partial f_{\mathbf{w}}(x)}{\partial \mathbf{w}^{(j)}} \cdot (\tilde{\mathbf{w}}^{(j)} - \mathbf{w}^{(j)}) \tag{12}$$

be the linear Taylor polynomial of $f_{\tilde{\mathbf{w}}}(x)$ around $\mathbf{w}$.

Lemma 1 *Let F_n be defined by (8), $\mathbf{w}^{(t+1)}$ by (9) and $f_{lin,\mathbf{w},\tilde{\mathbf{w}}}$ by (12). Let $\mathbf{w}^* \in \mathbb{R}^{J_n}$. Assume*

$$F_n(\mathbf{w}^{(s+1)}) \leq F_n(\mathbf{w}^{(s)}) - \frac{\lambda}{2} \cdot \left\| \nabla_{\mathbf{w}} F_n(\mathbf{w}^{(s)}) \right\|^2 \tag{13}$$

for $s = 0, 1, \ldots, t_n - 1$,

$$|f_{lin,\mathbf{w},\mathbf{w}^*}(x) - f_{\mathbf{w}^*}(x)| \leq C_n \cdot \|\mathbf{w}^* - \mathbf{w}\|^2 \tag{14}$$

for all $x \in \{X_1, \ldots, X_n\}$ and all $\mathbf{w} \in \mathbb{R}^{J_n}$ with $\|\mathbf{w}^* - \mathbf{w}\| \leq \|\mathbf{w}^* - \mathbf{w}^{(0)}\|$,

$$|Y_i| \leq \kappa_n \quad (i = 1, \ldots, n), \tag{15}$$

$$|f_{\mathbf{w}^*}(X_i)| \leq \kappa_n \quad (i = 1, \ldots, n) \tag{16}$$

and

$$C_n \cdot \|\mathbf{w}^* - \mathbf{w}^{(0)}\|^2 \leq \kappa_n. \tag{17}$$

Then

$$F_n(\mathbf{w}^{(t_n)}) \leq F_n(\mathbf{w}^*) + \left(5 \cdot \kappa_n \cdot C_n + \frac{1}{2 \cdot \lambda \cdot t_n} \right) \cdot \|\mathbf{w}^* - \mathbf{w}^{(0)}\|^2 + \frac{F_n(\mathbf{w}^{(0)})}{t_n}.$$

In the proof of Lemma 1 we will need the following auxiliary result.

Lemma 2 *Let $t \in \{1, \ldots, t_n\}$. Set*

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

and assume (13) and

$$F_n(\mathbf{w}^{(s+1)}) \geq F_{n,lin,\mathbf{w}^{(s)}}(\mathbf{w}^*) \tag{18}$$

for $s = 0, 1, \ldots, t - 1$. Then

$$\|\mathbf{w}^{(t)} - \mathbf{w}^*\| \leq \|\mathbf{w}^{(0)} - \mathbf{w}^*\|.$$

Proof. Let $s \in \{0, 1, \ldots, t-1\}$ be arbitrary. We have

$$F_{n,lin,\mathbf{w}^{(s)}}(\mathbf{w}^{(s)}) = F_n(\mathbf{w}^{(s)}) \quad \text{and} \quad \nabla_{\mathbf{w}} F_{n,lin,\mathbf{w}^{(s)}}(\mathbf{w}^{(s)}) = \nabla_{\mathbf{w}} F_n(\mathbf{w}^{(s)})$$

and (since $\mathbf{w} \mapsto F_{n,lin,\mathbf{w}^{(s)}}(\mathbf{w})$ is convex and differentiable)

$$F_{n,lin,\mathbf{w}^{(s)}}(\mathbf{w}^{(s)}) - F_{n,lin,\mathbf{w}^{(s)}}(\mathbf{w}^*) \leq < \nabla_{\mathbf{w}} F_{n,lin,\mathbf{w}^{(s)}}(\mathbf{w}^{(s)}), \mathbf{w}^{(s)} - \mathbf{w}^* >,$$

which implies

$$< \nabla_{\mathbf{w}} F_{n,}(\mathbf{w}^{(s)}), \mathbf{w}^{(s)} - \mathbf{w}^* > \;\; = \;\; < \nabla_{\mathbf{w}} F_{n,lin,\mathbf{w}^{(s)}}(\mathbf{w}^{(s)}), \mathbf{w}^{(s)} - \mathbf{w}^* >$$

$$
\begin{aligned}
&\geq F_{n,lin,\mathbf{w}^{(s)}}(\mathbf{w}^{(s)}) - F_{n,lin,\mathbf{w}^{(s)}}(\mathbf{w}^*) \\
&= F_n(\mathbf{w}^{(s)}) - F_{n,lin,\mathbf{w}^{(s)}}(\mathbf{w}^*).
\end{aligned}
$$

Hence

$$
\begin{aligned}
&\|\mathbf{w}^{(s+1)} - \mathbf{w}^*\|^2 \\
&= \left\|\mathbf{w}^{(s)} - \lambda \cdot \nabla_{\mathbf{w}} F_n(\mathbf{w}^{(s)}) - \mathbf{w}^*\right\|^2 \\
&= \left\|\mathbf{w}^{(s)} - \mathbf{w}^*\right\|^2 - 2 \cdot \lambda \cdot < \nabla_{\mathbf{w}} F_n(\mathbf{w}^{(s)}), \mathbf{w}^{(s)} - \mathbf{w}^* > + \lambda^2 \cdot \left\|\nabla_{\mathbf{w}} F_n(\mathbf{w}^{(s)})\right\|^2 \\
&\leq \left\|\mathbf{w}^{(s)} - \mathbf{w}^*\right\|^2 - 2 \cdot \lambda \left(F_n(\mathbf{w}^{(s)}) - F_{n,lin,\mathbf{w}^{(s)}}(\mathbf{w}^*)\right) \\
&\quad + 2 \cdot \lambda \cdot (F_n(\mathbf{w}^{(s)}) - F_n(\mathbf{w}^{(s+1)})) \\
&= \left\|\mathbf{w}^{(s)} - \mathbf{w}^*\right\|^2 - 2 \cdot \lambda \left(F_n(\mathbf{w}^{(s+1)}) - F_{n,lin,\mathbf{w}^{(s)}}(\mathbf{w}^*)\right) \\
&\leq \left\|\mathbf{w}^{(s)} - \mathbf{w}^*\right\|^2.
\end{aligned}
$$

Here we have used (13) in the first inequality and (18) in the second inequality. □

Proof of Lemma 1. Set

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

In the *first step of the proof* we show that for any $\mathbf{w} \in \mathbb{R}^{J_n}$ with $\|\mathbf{w}^* - \mathbf{w}\| \leq \|\mathbf{w}^* - \mathbf{w}^{(0)}\|$, we have

$$
|F_{n,lin,\mathbf{w}}(\mathbf{w}^*) - F_n(\mathbf{w}^*)| \leq 5 \cdot \kappa_n \cdot C_n \cdot \|\mathbf{w}^* - \mathbf{w}\|^2. \tag{19}
$$

Using (14)–(17) we get

$$
\begin{aligned}
&|F_{n,lin,\mathbf{w}}(\mathbf{w}^*) - F_n(\mathbf{w}^*)| \\
&= \left| \frac{1}{n} \sum_{i=1}^{n} |Y_i - f_{lin,\mathbf{w},\mathbf{w}^*}(X_i)|^2 - \frac{1}{n} \sum_{i=1}^{n} |Y_i - f_{\mathbf{w}^*}(X_i)|^2 \right| \\
&\leq \frac{1}{n} \sum_{i=1}^{n} |2 \cdot Y_i - 2 \cdot f_{\mathbf{w}^*}(X_i) + f_{\mathbf{w}^*}(X_i) - f_{lin,\mathbf{w},\mathbf{w}^*}(X_i)| \cdot |f_{lin,\mathbf{w},\mathbf{w}^*}(X_i) - f_{\mathbf{w}^*}(X_i)| \\
&\leq \frac{1}{n} \sum_{i=1}^{n} (2 \cdot \kappa_n + 2 \cdot \kappa_n + C_n \cdot \|\mathbf{w}^* - \mathbf{w}\|^2) \cdot C_n \cdot \|\mathbf{w}^* - \mathbf{w}\|^2 \\
&\leq 5 \cdot \kappa_n \cdot C_n \cdot \|\mathbf{w}^* - \mathbf{w}\|^2.
\end{aligned}
$$

In the *second step of the proof* we show that the assertion holds in case that we have for some $s \in \{0, 1, \ldots, t_n - 1\}$

$$
F_n(\mathbf{w}^{(s+1)}) < F_{n,lin,\mathbf{w}^{(s)}}(\mathbf{w}^*). \tag{20}
$$

So assume that (20) holds for some $s \in \{0, 1, \ldots, t_n - 1\}$. By choosing s minimal with this property we can assume

$$F_n(\mathbf{w}^{(t+1)}) \geq F_{n,lin,\mathbf{w}^{(t)}}(\mathbf{w}^*)$$

for all $t \in \{0, 1, \ldots, s-1\}$. By Lemma 2 we can conclude

$$\|\mathbf{w}^{(s)} - \mathbf{w}^*\| \leq \|\mathbf{w}^{(0)} - \mathbf{w}^*\|.$$

Furthermore, we know by assumption (13)

$$F_n(\mathbf{w}^{(t_n)}) \leq F_n(\mathbf{w}^{(s+1)}).$$

By using the result of the first step of the proof we get

$$\begin{aligned} F_n(\mathbf{w}^{(t_n)}) &\leq F_n(\mathbf{w}^{(s+1)}) \\ &< F_{n,lin,\mathbf{w}^{(s)}}(\mathbf{w}^*) \\ &= F_n(\mathbf{w}^*) + F_{n,lin,\mathbf{w}^{(s)}}(\mathbf{w}^*) - F_n(\mathbf{w}^*) \\ &\leq F_n(\mathbf{w}^*) + 5 \cdot \kappa_n \cdot C_n \cdot \|\mathbf{w}^* - \mathbf{w}^{(s)}\|^2 \\ &\leq F_n(\mathbf{w}^*) + 5 \cdot \kappa_n \cdot C_n \cdot \|\mathbf{w}^* - \mathbf{w}^{(0)}\|^2. \end{aligned}$$

In the *third step of the proof* we show the assertion in case that (20) does not hold for all $s \in \{0, 1, \ldots, t_n - 1\}$.

In this case we have

$$F_n(\mathbf{w}^{(s+1)}) \geq F_{n,lin,\mathbf{w}^{(s)}}(\mathbf{w}^*)$$

for all $s \in \{0, 1, \ldots, t_n - 1\}$, so by Lemma 2 we know

$$\|\mathbf{w}^{(t)} - \mathbf{w}^*\| \leq \|\mathbf{w}^{(0)} - \mathbf{w}^*\|$$

for all $t \in \{0, 1, \ldots, t_n\}$.

Using (13) and the result of the first step of the proof we conclude

$$\begin{aligned} &F_n(\mathbf{w}^{(t_n)}) - F_n(\mathbf{w}^*) \\ &\leq \frac{1}{t_n} \sum_{t=0}^{t_n-1} F_n(\mathbf{w}^{(t)}) - F_n(\mathbf{w}^*) \\ &= \frac{1}{t_n} \sum_{t=0}^{t_n-1} \left(F_n(\mathbf{w}^{(t)}) - F_{n,lin,\mathbf{w}^{(t)}}(\mathbf{w}^*) \right) + \frac{1}{t_n} \sum_{t=0}^{t_n-1} \left(F_{n,lin,\mathbf{w}^{(t)}}(\mathbf{w}^*) - F_n(\mathbf{w}^*) \right) \\ &\leq \frac{1}{t_n} \sum_{t=0}^{t_n-1} \left(F_n(\mathbf{w}^{(t)}) - F_{n,lin,\mathbf{w}^{(t)}}(\mathbf{w}^*) \right) + \frac{1}{t_n} \sum_{t=0}^{t_n-1} 5 \cdot \kappa_n \cdot C_n \cdot \|\mathbf{w}^* - \mathbf{w}^{(t)}\|^2 \\ &\leq \frac{1}{t_n} \sum_{t=0}^{t_n-1} \left(F_n(\mathbf{w}^{(t)}) - F_{n,lin,\mathbf{w}^{(t)}}(\mathbf{w}^*) \right) + 5 \cdot \kappa_n \cdot C_n \cdot \|\mathbf{w}^* - \mathbf{w}^{(0)}\|^2. \end{aligned}$$

Since

$$F_{n,lin,\mathbf{w}^{(t)}}(\mathbf{w}^{(t)}) = F_n(\mathbf{w}^{(t)}) \quad \text{and} \quad \nabla_{\mathbf{w}} F_{n,lin,\mathbf{w}^{(t)}}(\mathbf{w}^{(t)}) = \nabla_{\mathbf{w}} F_n(\mathbf{w}^{(t)})$$

and $\mathbf{w} \mapsto F_{n,lin,\mathbf{w}^{(t)}}(\mathbf{w})$ is convex and differentiable we get furthermore

$$\begin{aligned}
&\frac{1}{t_n}\sum_{t=0}^{t_n-1}\Big(F_n(\mathbf{w}^{(t)}) - F_{n,lin,\mathbf{w}^{(t)}}(\mathbf{w}^*)\Big) \\
&= \frac{1}{t_n}\sum_{t=0}^{t_n-1}\Big(F_{n,lin,\mathbf{w}^{(t)}}(\mathbf{w}^{(t)}) - F_{n,lin,\mathbf{w}^{(t)}}(\mathbf{w}^*)\Big) \\
&\leq \frac{1}{t_n}\sum_{t=0}^{t_n-1} < \nabla_{\mathbf{w}} F_{n,lin,\mathbf{w}^{(t)}}(\mathbf{w}^{(t)}), \mathbf{w}^{(t)} - \mathbf{w}^* > \\
&= \frac{1}{t_n}\sum_{t=0}^{t_n-1} < \nabla_{\mathbf{w}} F_n(\mathbf{w}^{(t)}), \mathbf{w}^{(t)} - \mathbf{w}^* > \\
&= \frac{1}{2\cdot\lambda\cdot t_n}\sum_{t=0}^{t_n-1} 2\cdot < \lambda\cdot\nabla_{\mathbf{w}} F_n(\mathbf{w}^{(t)}), \mathbf{w}^{(t)} - \mathbf{w}^* > \\
&= \frac{1}{2\cdot\lambda\cdot t_n}\sum_{t=0}^{t_n-1}\Bigg(\|\mathbf{w}^{(t)} - \mathbf{w}^*\|^2 - \|\mathbf{w}^{(t)} - \mathbf{w}^* - \lambda\cdot\nabla_{\mathbf{w}} F_n(\mathbf{w}^{(t)})\|^2 \\
&\qquad\qquad +\lambda^2\cdot\|\nabla_{\mathbf{w}} F_n(\mathbf{w}^{(t)})\|^2\Bigg) \\
&= \frac{1}{2\cdot\lambda\cdot t_n}\sum_{t=0}^{t_n-1}\Big(\|\mathbf{w}^{(t)} - \mathbf{w}^*\|^2 - \|\mathbf{w}^{(t+1)} - \mathbf{w}^*\|^2\Big) + \frac{\lambda}{2\cdot t_n}\sum_{t=0}^{t_n-1}\|\nabla_{\mathbf{w}} F_n(\mathbf{w}^{(t)})\|^2 \\
&\leq \frac{1}{2\cdot\lambda\cdot t_n}\cdot\Big(\|\mathbf{w}^{(0)} - \mathbf{w}^*\|^2 - \|\mathbf{w}^{(t_n)} - \mathbf{w}^*\|^2\Big) + \frac{1}{t_n}\sum_{t=0}^{t_n-1}(F_n(\mathbf{w}^{(t)}) - F_n(\mathbf{w}^{(t+1)})) \\
&\leq \frac{1}{2\cdot\lambda\cdot t_n}\cdot\|\mathbf{w}^{(0)} - \mathbf{w}^*\|^2 + \frac{F_n(\mathbf{w}^{(0)})}{t_n},
\end{aligned}$$

where the second to last inequality follows from (13). □

Next we investigate when our topology of the deep neural network satisfies the assumptions of Lemma 1. The following localization lemma for gradient descent proven in Braun et al. (2024) helps with inequality (13).

Lemma 3 *Let $F : \mathbb{R}^K \to \mathbb{R}_+$ be a nonnegative differentiable function. Let $t \in \mathbb{N}$, $\bar{L} > 0$, $\mathbf{a}_0 \in \mathbb{R}^K$, choose*

$$0 < \lambda \leq \frac{1}{\bar{L}}$$

and set

$$\mathbf{a}_{k+1} = \mathbf{a}_k - \lambda\cdot(\nabla_{\mathbf{a}} F)(\mathbf{a}_k) \quad (k \in \{0, 1, \ldots, t-1\}).$$

Assume

$$\|(\nabla_{\mathbf{a}} F)(\mathbf{a})\| \leq \sqrt{2 \cdot t \cdot \bar{L} \cdot \max\{F(\mathbf{a}_0), 1\}} \tag{21}$$

for all $\mathbf{a} \in \mathbb{R}^K$ *with* $\|\mathbf{a} - \mathbf{a}_0\| \leq \sqrt{2 \cdot t \cdot \max\{F(\mathbf{a}_0), 1\}/\bar{L}}$*, and*

$$\|(\nabla_{\mathbf{a}} F)(\mathbf{a}) - (\nabla_{\mathbf{a}} F)(\mathbf{b})\| \leq \bar{L} \cdot \|\mathbf{a} - \mathbf{b}\| \tag{22}$$

for all $\mathbf{a}, \mathbf{b} \in \mathbb{R}^K$ *satisfying*

$$\|\mathbf{a} - \mathbf{a}_0\| \leq \sqrt{8 \cdot \frac{t}{\bar{\bar{L}}} \cdot \max\{F(\mathbf{a}_0), 1\}} \quad \textit{and} \quad \|\mathbf{b} - \mathbf{a}_0\| \leq \sqrt{8 \cdot \frac{t}{\bar{\bar{L}}} \cdot \max\{F(\mathbf{a}_0), 1\}}. \tag{23}$$

Then we have

$$\|\mathbf{a}_k - \mathbf{a}_0\| \leq \sqrt{2 \cdot \frac{k}{\bar{\bar{L}}} \cdot (F(\mathbf{a}_0) - F(\mathbf{a}_k))} \quad \textit{for all } k \in \{1, \ldots, t\},$$

$$\sum_{k=0}^{s-1} \|\mathbf{a}_{k+1} - \mathbf{a}_k\|^2 \leq \frac{2}{\bar{\bar{L}}} \cdot (F(\mathbf{a}_0) - F(\mathbf{a}_s)) \quad \textit{for all } s \in \{1, \ldots, t\}$$

and

$$F(\mathbf{a}_k) \leq F(\mathbf{a}_{k-1}) - \frac{\lambda}{2} \cdot \|(\nabla_{\mathbf{a}} F)(\mathbf{a}_{k-1})\|^2 \quad \textit{for all } k \in \{1, \ldots, t\}.$$

Proof. See Lemma A.1 in Braun et al. (2024). □

We use the following two lemmata from Kohler (2026) to verify assumptions (21) and (22) of Lemma 3.

Lemma 4 *Let* $\sigma : \mathbb{R} \to \mathbb{R}$ *be the logistic squasher, let* $L, r, K_n \in \mathbb{N}$*, let* $f_{\mathbf{w}}$ *be defined by (4) - (6) and let* F_n *be defined by (8). Let* $a \geq 1$*,* $\gamma_n^* \geq 1$*,* $B_n \geq 1$*,* $\kappa_n \geq 1$ *and assume* $X_i \in [-a, a]^d$*,* $|Y_i| \leq \kappa_n$ $(i = 1, \ldots, n)$*,*

$$|w_{1,1,k}^{(L)}| \leq \gamma_n^* \textit{ for } k = 1, \ldots, K_n,$$

$$|w_{k,i,j}^{(l)}| \leq B_n \textit{ for } l = 1, \ldots, L-1 \textit{ and all } k, i, j$$

and

$$K_n \cdot \gamma_n^* \geq \kappa_n.$$

Then

$$\|\nabla_{\mathbf{w}} F_n(\mathbf{w})\| \leq c_{16} \cdot K_n^{3/2} \cdot (\gamma_n^*)^2 \cdot B_n^L$$

for some $c_{16} = c_{16}(d, L, r, a) > 0$*.*

Proof. See Lemma 3 in Kohler (2026). □

Lemma 5 *Let $\sigma : \mathbb{R} \to \mathbb{R}$ be the logistic squasher, let $L, r, K_n \in \mathbb{N}$, let $f_{\mathbf{w}}$ be defined by (4) - (6) and let F_n be defined by (8). Let $a \geq 1$, $\gamma_n^* \geq 1$, $B_n \geq 1$, $\kappa_n \geq 1$ and assume $X_i \in [-a, a]^d$, $|Y_i| \leq \kappa_n$ $(i = 1, \ldots, n)$,*

$$\max\{|(\mathbf{w}_1)_{1,1,k}^{(L)}|, |(\mathbf{w}_2)_{1,1,k}^{(L)}|\} \leq \gamma_n^* \text{ for } k = 1, \ldots, K_n, \tag{24}$$

$$\max\{|(\mathbf{w}_1)_{k,i,j}^{(l)}|, |(\mathbf{w}_2)_{k,i,j}^{(l)}|\} \leq B_n \quad \text{for } l = 1, \ldots, L-1 \text{ and all } k, i, j \tag{25}$$

and

$$K_n \cdot \gamma_n^* \geq \kappa_n.$$

Then we have

$$\|(\nabla_{\mathbf{w}} F_n)(\mathbf{w}_1) - (\nabla_{\mathbf{w}} F_n)(\mathbf{w}_2)\| \leq c_{17} \cdot K_n^{3/2} \cdot B_n^{2L} \cdot (\gamma_n^*)^2 \cdot \|\mathbf{w}_1 - \mathbf{w}_2\|$$

for some $c_{17} = c_{17}(d, L, r, a) > 0$.

Proof. See Lemma 5 in Kohler (2026). □

The following two lemmata consider inequality (14) for the special topology of our networks.

Lemma 6 *Let $f_{\mathbf{w}} : \mathbb{R}^d \to \mathbb{R}$ be a deep neural network defined by (4) - (6) with weight vector*

$$\mathbf{w} = (w_j)_{j=1,\ldots,J_n} \in \mathbb{R}^{J_n}$$

and a twice differentiable activation function, and denote the linear Taylor polynomial of $f_{\mathbf{w}}$ around $\mathbf{w}_0$ by

$$f_{lin,\mathbf{w}_0,\mathbf{w}}(x) = f_{\mathbf{w}_0}(x) + \sum_{j=1}^{J_n} \frac{\partial}{\partial w_j} f_{\mathbf{w}_0}(x) \cdot (w_j - (\mathbf{w}_0)_j).$$

Then for every $x \in \mathbb{R}^d$ there exists $\xi \in [0, 1]$ such that

$$|f_{lin,\mathbf{w}_0,\mathbf{w}}(x) - f_{\mathbf{w}}(x)| \leq \frac{1}{2} \cdot \|\mathbf{H}(\mathbf{w}_0 + \xi \cdot (\mathbf{w} - \mathbf{w}_0))\|_2 \cdot \|\mathbf{w} - \mathbf{w}_0\|^2,$$

where

$$\mathbf{H}(\mathbf{w}) = \left(\frac{\partial^2 f_{\mathbf{w}}(x)}{\partial w_i \partial w_j} \right)_{1 \leq i,j \leq J_n}$$

is the Hessian matrix of $f_{\mathbf{w}}(x)$ and

$$\|\mathbf{H}(\mathbf{w})\|_2 = \sup_{\tilde{\mathbf{w}} \in \mathbb{R}^{J_n} : \tilde{\mathbf{w}} \neq 0} \frac{\|\mathbf{H}(\mathbf{w})\tilde{\mathbf{w}}\|}{\|\tilde{\mathbf{w}}\|}$$

denotes its spectral norm.

Proof. For $x \in \mathbb{R}^d$ and $s \in [0,1]$ define

$$F(s) = f_{\mathbf{w}_0 + s\cdot(\mathbf{w}-\mathbf{w}_0)}(x).$$

Then the chain rule and the formula for Taylor polynomials of order 2 imply that for some $\xi \in [0,1]$ we have

$$\begin{aligned}
&|f_{\mathbf{w}}(x) - f_{lin,\mathbf{w}_0,\mathbf{w}}(x)| \\
&= |F(1) - F(0) - F'(0)\cdot(1-0)| \\
&= \left|\frac{1}{2}\cdot F''(\xi)\cdot(1-0)^2\right| \\
&= \left|\frac{1}{2}\cdot(\mathbf{w}-\mathbf{w}_0)^T\cdot\mathbf{H}(\mathbf{w}_0+\xi\cdot(\mathbf{w}-\mathbf{w}_0))\cdot(\mathbf{w}-\mathbf{w}_0)\right| \\
&\leq \frac{1}{2}\cdot\|\mathbf{w}-\mathbf{w}_0\|\cdot\|\mathbf{H}(\mathbf{w}_0+\xi\cdot(\mathbf{w}-\mathbf{w}_0))\cdot(\mathbf{w}-\mathbf{w}_0)\| \\
&\leq \frac{1}{2}\cdot\|\mathbf{H}(\mathbf{w}_0+\xi\cdot(\mathbf{w}-\mathbf{w}_0))\|_2\cdot\|\mathbf{w}-\mathbf{w}_0\|^2.
\end{aligned}$$

□

Lemma 7 *Let σ be the logistic squasher and define $f_{\mathbf{w}}$ by (4) - (6). Let $c_{18} > 0$ and assume*

$$|w_{k,i,j}^{(l)}| \leq c_{18}$$

for all $l \in \{1,\ldots,L\}$ and all k,i,j. Let $R > 0$. Then we have for all $x \in \mathbb{R}^d$ with $\|x\| \leq R$

$$\|\mathbf{H}(\mathbf{w})\|_2 \leq c_{19}$$

for some constant $c_{19} = c_{19}(L,r,d,R,c_{18}) > 0$ which does not depend on K_n.

Proof. Since

$$\frac{\partial^2 f_{\mathbf{w}}(x)}{\partial w_{k_1,i_1,j_1}^{(l_1)}\partial w_{k_2,i_2,j_2}^{(l_2)}} = 0$$

holds whenever $k_1 \neq k_2$, the Hessian matrix is a block diagonal matrix given by

$$\begin{pmatrix} \mathbf{A}_1 & 0 & \ldots & 0 \\ 0 & \mathbf{A}_2 & \ldots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \ldots & \mathbf{A}_{K_n} \end{pmatrix}.$$

If we split $\tilde{\mathbf{w}}$ accordingly into $\tilde{\mathbf{w}} = (\tilde{\mathbf{w}}_1,\ldots,\tilde{\mathbf{w}}_{K_n})^T$, then we have

$$\|\mathbf{H}(\mathbf{w})\cdot\tilde{\mathbf{w}}\|^2 = \left\|\begin{pmatrix} \mathbf{A}_1\tilde{\mathbf{w}}_1 \\ \vdots \\ \mathbf{A}_{K_n}\tilde{\mathbf{w}}_{K_n} \end{pmatrix}\right\|^2 = \sum_{k=1}^{K_n}\|\mathbf{A}_k\tilde{\mathbf{w}}_k\|^2 \leq \sum_{k=1}^{K_n}\|\mathbf{A}_k\|_2^2\cdot\|\tilde{\mathbf{w}}_k\|^2$$

$$\leq \max\{\|\mathbf{A}_1\|_2^2, \ldots, , \|\mathbf{A}_{K_n}\|_2^2\} \cdot \sum_{k=1}^{K_n} \|\tilde{\mathbf{w}}_k\|^2$$
$$= \max\{\|\mathbf{A}_1\|_2^2, \ldots, , \|\mathbf{A}_{K_n}\|_2^2\} \cdot \|\tilde{\mathbf{w}}\|^2,$$

which implies

$$\|\mathbf{H}(\mathbf{w})\|_2 \leq \max\{\|\mathbf{A}_1\|_2, \ldots, , \|\mathbf{A}_{K_n}\|_2\}.$$

By construction, each matrix $\mathbf{A}_k$ is a square matrix of size

$$K = r + 2 + (L-2) \cdot r \cdot (r+1) + r \cdot (d+1),$$

where all entries are bounded by a constant depending only on L, r, d, c_{18} and $\|x\|_\infty$. It is easy to see that the spectral norm of a matrix is bounded by its Frobenius norm, i.e., that

$$\|A_k\|_2 = \|(a_{i,j})_{1 \leq i,j \leq K}\|_2 \leq \sqrt{\sum_{i=1}^{K} \sum_{j=1}^{K} a_{i,j}^2}$$

holds, which implies the assertion. □

We summarize all our results concerning neural network optimization in the following theorem.

Theorem 2 *Let $\sigma : \mathbb{R} \to \mathbb{R}$ be the logistic squasher, let $f_{\mathbf{w}}$ be defined by (4) - (6), let $L, r, K_n \in \mathbb{N}$ and let F_n be defined by (8). Let $c_{20} > 0$, $c_{21} \geq 1$, $c_{22} \geq 1$, $\kappa_n \geq 1$, $A \geq 1$ and assume $X_1, \ldots X_n \in [-A, A]^d$, $|Y_i| \leq \kappa_n$ $(i \in \{1, \ldots, n\})$ and*

$$c_{21} \cdot K_n \geq \kappa_n \text{ and } \frac{\kappa_n}{n} \leq c_{20}.$$

Choose a starting vector $\mathbf{w}^{(0)}$ which satisfies

$$(\mathbf{w}^{(0)})_{1,1,k}^{(L)} = 0 \text{ and } |(\mathbf{w}^{(0)})_{k,i,j}^{(l)}| \leq c_{22}$$

for all $l \in \{1, \ldots, L-1\}$ and all i, j, k, and set

$$\mathbf{w}^{(t+1)} = \mathbf{w}^{(t)} - \lambda_n \cdot \nabla_{\mathbf{w}} F_n(\mathbf{w}^{(t)})$$

for $t = 0, 1, \ldots, t_n - 1$. Set

$$\lambda_n = \frac{c_{23}}{K_n^{3/2} \cdot \kappa_n} \quad \text{and} \quad t_n = \left\lceil c_{24} \cdot \frac{K_n^{3/2}}{\kappa_n} \right\rceil$$

for $c_{23}, c_{24} > 0$. Let

$$\mathbf{w}^* \in \left\{ \mathbf{w} : \|\mathbf{w} - \mathbf{w}^{(0)}\| \leq c_{25} \cdot \frac{\sqrt{\kappa_n}}{\sqrt{n}} \right\} \tag{26}$$

for $c_{25} > 0$ and assume (16) holds. Then we have for $c_{26}, c_{27} > 0$ and n sufficiently large

$$F_n(\mathbf{w}^{(t_n)}) \leq F_n(\mathbf{w}^*) + c_{26} \cdot \kappa_n^2 \cdot \|\mathbf{w}^* - \mathbf{w}^{(0)}\|^2 + c_{27} \cdot \frac{\kappa_n^3}{K_n^{3/2}}.$$

Proof. The assertion follows from Lemma 1 with $\lambda = \lambda_n$ and we will now check that the assumptions (13) - (17) are fulfilled. Assumptions (15) and (16) are trivially fulfilled. We will now consider assumption (13). This assumption follows directly from Lemma 3. Noticing that

$$\kappa_n \cdot \sqrt{t_n \cdot \lambda_n} \leq c_{28} \text{ and } F_n(\mathbf{w}^{(0)}) \leq \kappa_n^2,$$

we can see that the assumptions of Lemma 3 are fulfilled if we use Lemma 4 and Lemma 5 with

$$\tilde{\gamma}^* = c_{21} + c_{29} \cdot \kappa_n \cdot \sqrt{t_n \cdot \lambda_n} \text{ and } B_n = c_{22} + c_{30} \cdot \kappa_n \cdot \sqrt{t_n \cdot \lambda_n}$$

and

$$\bar{L} = \frac{1}{\lambda_n} = c_{31} \cdot K_n^{3/2} \cdot \kappa_n.$$

Assumption (13) is hence fulfilled. Remember that $\left|\left(\mathbf{w}^{(0)}\right)_{k,i,j}^{(l)}\right|$ is bounded by a constant for $l > 0$ and all k, i, j. Using this and (26) we can show that $\left|(\mathbf{w} + \xi \cdot (\mathbf{w}^* - \mathbf{w}))_{k,i,j}^{(l)}\right|$ is bounded by a constant for all $\mathbf{w} \in \mathbb{R}^{J_n}$ with $\|\mathbf{w}^* - \mathbf{w}\| \leq \|\mathbf{w}^* - \mathbf{w}^{(0)}\|$ and all $\xi \in [0, 1]$, $l \in \{1, \ldots, L\}$ and k, i, j. Assumption (14) follows consequently from Lemma 6 and Lemma 7 with $C_n = c_{19}(L, r, d, A, c_{32})$ for some $c_{32} > 0$, where we use $R = A$. Assumption (17) follows from (26) for large enough n. With Lemma 1 we conclude

$$\begin{aligned} F_n(\mathbf{w}^{(t_n)}) &\leq F_n(\mathbf{w}^*) + \left(c_{33} \cdot \kappa_n + c_{34} \cdot \kappa_n^2\right) \cdot \|\mathbf{w}^* - \mathbf{w}^{(0)}\|^2 + c_{35} \cdot \frac{\kappa_n^3}{K_n^{3/2}} \\ &\leq F_n(\mathbf{w}^*) + c_{36} \cdot \kappa_n^2 \cdot \|\mathbf{w}^* - \mathbf{w}^{(0)}\|^2 + c_{35} \cdot \frac{\kappa_n^3}{K_n^{3/2}}. \end{aligned}$$

□

4.2. Neural network approximation

We use the following result from Kohler (2026) for bounding the approximation error.

Theorem 3 *Let $d \in \mathbb{N}$, $p = q + s$ where $s \in (0, 1]$ and $q \in \mathbb{N}_0$, $C > 0$, $A \geq 1$ and $A_n, B_n, \gamma_n^* \geq 1$. Let σ be the logistic squasher. For $L, r, K \in \mathbb{N}$ let $\mathcal{F}$ be the set of all networks $f_{\mathbf{w}}$ defined by*

$$f_{\mathbf{w}}(x) = \sum_{j=1}^{r} w_{1,1,j}^{(L)} \cdot f_{j,1}^{(L)}(x) \tag{27}$$

for some $w_{1,1,1}^{(L)}, \ldots, w_{1,1,r}^{(L)} \in \mathbb{R}$, where $f_{j,1}^{(L)} = f_{\mathbf{w},j,1}^{(L)}$ are recursively defined by

$$f_{k,i}^{(l)}(x) = f_{\mathbf{w},k,i}^{(l)}(x) = \sigma\left(\sum_{j=1}^{r} w_{k,i,j}^{(l-1)} \cdot f_{k,j}^{(l-1)}(x) + w_{k,i,0}^{(l-1)}\right) \tag{28}$$

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

$$f_{k,i}^{(1)}(x) = f_{\mathbf{w},k,i}^{(1)}(x) = \sigma\left(\sum_{j=1}^{d} w_{k,i,j}^{(0)} \cdot x^{(j)} + w_{k,i,0}^{(0)}\right) \tag{29}$$

for some $w_{k,i,0}^{(0)}, \ldots, w_{k,i,d}^{(0)} \in \mathbb{R}$*, where the weight vector satisfies*

$$|w_{k,i,j}^{(0)}| \leq A_n, \quad |w_{k,i,j}^{(l)}| \leq B_n \quad \textit{and} \quad |w_{k,i,j}^{(L)}| \leq \gamma_n^*$$

for all $l \in \{1, \ldots, L-1\}$ *and all* k, i, j*, and set*

$$\mathcal{H} = \left\{ \sum_{k=1}^{K^d} f_k \quad : \quad f_k \in \mathcal{F} \quad (k = 1, \ldots, K^d) \right\}.$$

Let $L, r \in \mathbb{N}$ *with*

$$L \geq \lceil \log_2(q+d) \rceil \quad \textit{and} \quad r \geq 4 \cdot (p+d)^2,$$

and set

$$A_n = A \cdot K \cdot \log K, \quad B_n = c_{37} \quad \textit{and} \quad \gamma_n^* = c_{38} \cdot K^{q+d}.$$

Assume $K \geq c_{39}$ *for* $c_{39} > 0$ *sufficiently large. Then there exists for any* (p, C)*–smooth* $f : \mathbb{R}^d \to \mathbb{R}$ *a neural network* $h \in \mathcal{H}$ *such that*

$$\sup_{x \in [-A,A)^d} |f(x) - h(x)| \leq \frac{c_{40}}{K^p}.$$

Proof. See Theorem 3 in Kohler (2026). □

4.3. Neural network generalization

Our next lemma is our main tool to analyze the generalization error for exponentially β-mixing data.

Lemma 8 *Assume (A4) - (A5), let* $n \in \mathbb{N}$ *with* $n > 1$*, and let* $\mathcal{F}_n$ *be a set of functions* $f : \mathbb{R}^d \to \mathbb{R}$*. Then we have*

$$\begin{aligned}
&\mathbf{E}\left\{ \sup_{f \in \mathcal{F}_n} \mathbf{E}\left\{ |(T_{\kappa_n} f)(X) - T_{\kappa_n} Y|^2 \right\} - \mathbf{E}\{ |m_{\kappa_n}(X) - T_{\kappa_n} Y|^2 \} \right. \\
&\qquad \left. - 2 \cdot \frac{1}{n} \sum_{i=1}^{n} \left(|(T_{\kappa_n} f)(X_i) - T_{\kappa_n} Y_i|^2 - |m_{\kappa_n}(X_i) - T_{\kappa_n} Y_i|^2 \right) \right\} \\
&\leq 4 \cdot \kappa_n^2 \cdot \frac{n}{(\log n)^2} \cdot \beta_{\lceil (\log n)^2 \rceil}((X_1, Y_1), (X_2, Y_2), \ldots,) + \epsilon \\
&\quad + \int_{\epsilon}^{\infty} 14 \cdot \sup_{x_1^n \in \mathrm{supp}(X)^n} \mathcal{N}_1\left(\frac{u}{80 \cdot \kappa_n}, \mathcal{F}_n, x_1^n \right) \cdot \exp\left(-c_{41} \cdot \frac{u \cdot n}{\kappa_n^2 \cdot (\log n)^2} \right) du
\end{aligned}$$

for every $\epsilon > 0$ *and some* $c_{41} > 0$*.*

We will use the following auxiliary result to prove Lemma 8, which allows us to reduce our problem to the case of independent data and controls the probability of the resulting error with β-mixing coefficients.

Lemma 9 *Let Z_0, Z_1, …, Z_n be identically distributed $\mathbb{R}^d$–valued random variables defined on a probability space $(\Omega, \mathcal{A}, \mathbf{P})$ such that Z_0 is independent from Z_1, …, Z_n. Then there exists a probability space $(\bar{\Omega}, \bar{\mathcal{A}}, \bar{\mathbf{P}})$ and random variables $\bar{Z}_0$, $\bar{Z}_1$, …, $\bar{Z}_n$, $\bar{Z}_1^*$, …, $\bar{Z}_n^*$ defined on this probability space such that*

$$\mathbf{P}_{(Z_0,Z_1,\dots,Z_n)} = \bar{\mathbf{P}}_{(\bar{Z}_0,\bar{Z}_1,\dots,\bar{Z}_n)}, \tag{30}$$

$$\bar{Z}_0, \bar{Z}_1^*, \dots, \bar{Z}_n^* \quad \textit{are i.i.d.} \tag{31}$$

and

$$\begin{aligned}&\bar{\mathbf{P}}\{\exists k \in \{1,\dots,n\} : \bar{Z}_k \neq \bar{Z}_k^*\}\\ &\leq (n-1) \cdot \max_{s\in\{2,\dots,n\}} \mathbf{E}\left\{\sup_{C\in\mathcal{C}} |\mathbf{P}\{Z_s \in C | Z_1,\dots,Z_{s-1}\} - \mathbf{P}\{Z_s \in C\}|\right\},\end{aligned} \tag{32}$$

where $\mathcal{C}$ is some countable subset of $\mathcal{B}_d$.

Lemma 9 is closely related to classical coupling results (see Berbee (1979), Corollary 4.2.4 and Doukhan (1994), Theorem 1, Section 1.1) and its extensions (c.f. Barrera and Gobet (2021), Lemma 2.10), and follows from these well–known results. For the sake of completeness, we provide nevertheless a complete and self-contained proof of this result in the appendix.

Proof of Lemma 8. Set $N_n = \lceil (\log n)^2 \rceil$ and $I_{n,k} = \{k, k+N_n, k+2\cdot N_n, \dots\} \cap \{1,\dots,n\}$ for $k \in \{1,\dots,N_n\}$. For each $n \in \mathbb{N}$ we have

$$\bigcup_{k=1}^{N_n} I_{n,k} = \{1,\dots,n\} \text{ and } I_{n,k_1} \cap I_{n,k_2} = \emptyset \text{ for all } k_1, k_2 \in \{1,\dots,N_n\}, k_1 \neq k_2.$$

Notice that

$$|I_{n,k}| \in \left\{ \left\lfloor \frac{n}{N_n} \right\rfloor, \left\lceil \frac{n}{N_n} \right\rceil \right\} = \left\{ \left\lfloor \frac{n}{\lceil (\log n)^2 \rceil} \right\rfloor, \left\lceil \frac{n}{\lceil (\log n)^2 \rceil} \right\rceil \right\}$$

and hence

$$|I_{n,k}| \leq \frac{n}{(\log n)^2} + 1. \tag{33}$$

We get with the convexity of the supremum

$$\begin{aligned}\mathbf{E}\Bigg\{ \sup_{f\in\mathcal{F}_n} \Bigg(&\mathbf{E}\left\{|(T_{\kappa_n} f)(X) - T_{\kappa_n} Y|^2\right\} - \mathbf{E}\{|m_{\kappa_n}(X) - T_{\kappa_n} Y|^2\}\\ &- 2 \cdot \frac{1}{n} \sum_{i=1}^{n} \left(|(T_{\kappa_n} f)(X_i) - T_{\kappa_n} Y_i|^2 - |m_{\kappa_n}(X_i) - T_{\kappa_n} Y_i|^2 \right) \Bigg) \Bigg\}\end{aligned}$$

$$
\begin{aligned}
&\le \frac{1}{n}\sum_{k=1}^{N_n} |I_{n,k}| \mathbf{E}\left\{ \sup_{f\in\mathcal{F}_n} \left(\mathbf{E}\left\{ |(T_{\kappa_n} f)(X) - T_{\kappa_n} Y|^2 \right\} - \mathbf{E}\{ |m_{\kappa_n}(X) - T_{\kappa_n} Y|^2 \right\} \right.\right. \\
&\qquad \left.\left. - 2\cdot \frac{1}{|I_{n,k}|} \sum_{i\in I_{n,k}} \left(|(T_{\kappa_n} f)(X_i) - T_{\kappa_n} Y_i|^2 - |m_{\kappa_n}(X_i) - T_{\kappa_n} Y_i|^2 \right) \right) \right\} \\
&\le \max_{k=1,\dots,N_n} \mathbf{E}\left\{ \sup_{f\in\mathcal{F}_n} \left(\mathbf{E}\left\{ |(T_{\kappa_n} f)(X) - T_{\kappa_n} Y|^2 \right\} - \mathbf{E}\{ |m_{\kappa_n}(X) - T_{\kappa_n} Y|^2 \right\} \right.\right. \\
&\qquad \left.\left. - 2\cdot \frac{1}{|I_{n,k}|} \sum_{i\in I_{n,k}} \left(|(T_{\kappa_n} f)(X_i) - T_{\kappa_n} Y_i|^2 - |m_{\kappa_n}(X_i) - T_{\kappa_n} Y_i|^2 \right) \right) \right\} \\
&=: \max_{k=1,\dots,N_n} \mathbf{E}\{ g_{n,k}((X,Y), ((X_i,Y_i))_{i\in I_{n,k}}) \}.
\end{aligned}
$$

Here we used that $\sum_{k=1}^{N_n} |I_{n,k}| = n$.

Now we use Lemma 9 with

$$
n = |I_{n,k}|,\ Z_0 = (X,Y)
$$

and

$$
(Z_i)_{i=1}^{|I_{n,k}|} = ((X_{k+(i-1)\cdot N_n}, Y_{k+(i-1)\cdot N_n}))_{i=1}^{|I_{n,k}|} (= ((X_j,Y_j))_{j\in I_{n,k}})
$$

for each $k \in \{1,\dots,N_n\}$. It follows that for each $k \in \{1,\dots,N_n\}$ there exists a probability space $(\Omega^{(k)}, \mathcal{A}^{(k)}, \mathbf{P}^{(k)})$ and random variables $(X^{(k)}, Y^{(k)})$, $((X_j^{(k)}, Y_j^{(k)}))_{j\in I_{n,k}}$ and $((X_j^{*(k)}, Y_j^{*(k)}))_{j\in I_{n,k}}$ defined on this probability space such that

$$
\mathbf{P}_{((X,Y),((X_j,Y_j))_{j\in I_{n,k}})} = \mathbf{P}^{(k)}_{((X^{(k)},Y^{(k)}),((X_j^{(k)},Y_j^{(k)}))_{j\in I_{n,k}})}, \tag{34}
$$

$$
(X^{(k)}, Y^{(k)}), ((X_j^{*(k)}, Y_j^{*(k)}))_{j\in I_{n,k}} \text{ are i.i.d.} \tag{35}
$$

and

$$
\begin{aligned}
&\mathbf{P}^{(k)}\{ \exists j \in I_{n,k} : (X_j^{(k)}, Y_j^{(k)}) \neq (X_j^{*(k)}, Y_j^{*(k)}) \} \\
&\le (|I_{n,k}| - 1) \max_{1\in\{1,\dots,|I_{n,k}|-1\}} \mathbf{E}\left\{ \sup_{C\in\mathcal{C}} \left| \mathbf{P}\{ (X_{k+s\cdot N_n}, Y_{k+s\cdot N_n}) \in C | \right.\right. \\
&\qquad (X_k,Y_k), (X_{k+N_n}, Y_{k+N_n}), \dots, (X_{k+(s-1)\cdot N_n}, Y_{k+(s-1)\cdot N_n}) \} \\
&\qquad\qquad \left.\left. - \mathbf{P}\{ (X_{k+s\cdot N_n}, Y_{k+s\cdot N_n}) \in C \} \right| \right\} \\
&= (|I_{n,k}| - 1) \max_{i\in I_{n,k}\setminus\{k\}} \\
&\qquad \mathbf{E}\left\{ \sup_{C\in\mathcal{C}} |\mathbf{P}\{ (X_i,Y_i) \in C | (X_j,Y_j) \,:\, j \in I_{n,k}, j < i \} - \mathbf{P}\{ (X_i,Y_i) \in C \}| \right\}
\end{aligned}
$$

for some countable set $\mathcal{C} \subset B_{d+1}$.

We have for $i \in I_{n,k} \setminus \{k\}$ and each $C \in \mathcal{C}$ by the tower property and Jensen's inequality

$$\begin{aligned}
&|\mathbf{P}\{(X_i, Y_i) \in C | (X_j, Y_j) \,:\, j \in I_{n,k}, j < i\} - \mathbf{P}\{(X_i, Y_i) \in C\}| \\
&\quad = |\mathbf{E}\{1_{\{(X_i,Y_i)\in C\}} - \mathbf{P}\{(X_i, Y_i) \in C\} | (X_j, Y_j) \,:\, j \in I_{n,k}, j < i\}| \\
&\quad = |\mathbf{E}\{\mathbf{E}\{1_{\{(X_i,Y_i)\in C\}} - \mathbf{P}\{(X_i, Y_i) \in C\} | (X_m, Y_m) \,:\, m = 1, \dots, i - N_n\} \\
&\qquad\quad | (X_j, Y_j) \,:\, j \in I_{n,k}, j < i\}| \\
&\quad \le \mathbf{E}\{|\mathbf{E}\{1_{\{(X_i,Y_i)\in C\}} - \mathbf{P}\{(X_i, Y_i) \in C\} | (X_m, Y_m) \,:\, m = 1, \dots, i - N_n\}| \\
&\qquad\quad | (X_j, Y_j) \,:\, j \in I_{n,k}, j < i\} \\
&\quad = \mathbf{E}\{|\mathbf{P}\{(X_i, Y_i) \in C | (X_m, Y_m) \,:\, m = 1, \dots, i - N_n\} - \mathbf{P}\{(X_i, Y_i) \in C\}| \\
&\qquad\quad | (X_j, Y_j) \,:\, j \in I_{n,k}, j < i\} \\
&\quad \le \mathbf{E}\{\operatorname*{ess\,sup}_{A \in \mathcal{B}_{d+1}} |\mathbf{P}\{(X_i, Y_i) \in A | (X_m, Y_m) \,:\, m = 1, \dots, i - N_n\} - \mathbf{P}\{(X_i, Y_i) \in A\}| \\
&\qquad\quad | (X_j, Y_j) \,:\, j \in I_{n,k}, j < i\}
\end{aligned}$$

almost surely. This implies

$$\begin{aligned}
&\mathbf{E}\left\{\sup_{C \in \mathcal{C}} |\mathbf{P}\{(X_i, Y_i) \in C | (X_j, Y_j) \,:\, j \in I_{n,k}, j < i\} - \mathbf{P}\{(X_i, Y_i) \in C\}|\right\} \\
&\le \mathbf{E}\{\operatorname*{ess\,sup}_{A \in \mathcal{B}_{d+1}} |\mathbf{P}\{(X_i, Y_i) \in A | (X_m, Y_m) \,:\, m = 1, \dots, i - N_n\} - \mathbf{P}\{(X_i, Y_i) \in A\}|\}
\end{aligned}$$

and hence

$$\begin{aligned}
&\mathbf{P}^{(k)}(\exists j \in I_{n,k} : (X_j^{(k)}, Y_j^{(k)}) \neq (X_j^{*(k)}, Y_j^{*(k)})) \\
&\quad \le (|I_{n,k}| - 1) \cdot \beta_{N_n}((X_1, Y_1), (X_2, Y_2), \dots).
\end{aligned} \tag{36}$$

Denote with $\mathbf{E}^{(k)}$ the expectation with respect to $\mathbf{P}^{(k)}$. We have by (34)

$$\begin{aligned}
&\mathbf{E}\{g_{n,k}((X, Y), ((X_i, Y_i))_{i \in I_{n,k}})\} = \mathbf{E}^{(k)}\{g_{n,k}((X^{(k)}, Y^{(k)}), ((X_i^{(k)}, Y_i^{(k)}))_{i \in I_{n,k}})\} \\
&\quad = \mathbf{E}^{(k)}\{g_{n,k}((X^{(k)}, Y^{(k)}), ((X_i^{(k)}, Y_i^{(k)}))_{i \in I_{n,k}}) \cdot 1_{\{\exists j \in I_{n,k} : (X_j^{(k)}, Y_j^{(k)}) \neq (X_j^{*(k)}, Y_j^{*(k)})\}}\} \\
&\quad + \mathbf{E}^{(k)}\{g_{n,k}((X^{(k)}, Y^{(k)}), ((X_i^{(k)}, Y_i^{(k)}))_{i \in I_{n,k}}) \cdot 1_{\{\forall j \in I_{n,k} : (X_j^{(k)}, Y_j^{(k)}) = (X_j^{*(k)}, Y_j^{*(k)})\}}\} \\
&\quad =: \mathbf{E}^{(k)}\{S_{1,n,k}\} + \mathbf{E}^{(k)}\{S_{2,n,k}\}.
\end{aligned}$$

Since $g_{n,k}((x, y), ((x_i, y_i))_{i \in I_{n,k}}) \le 4 \cdot \kappa_n^2$, we get by (33) and (36)

$$\begin{aligned}
\mathbf{E}^{(k)}\{S_{1,n,k}\} &\le 4 \cdot \kappa_n^2 \cdot \mathbf{P}^{(k)}(\exists j \in I_{n,k} : (X_j^{(k)}, Y_j^{(k)}) \neq (X_j^{*(k)}, Y_j^{*(k)})) \\
&\le 4 \cdot \kappa_n^2 \cdot (|I_{n,k}| - 1) \cdot \beta_{N_n}((X_1, Y_1), (X_2, Y_2), \dots) \\
&\le 4 \cdot \kappa_n^2 \cdot \frac{n}{(\log n)^2} \cdot \beta_{N_n}((X_1, Y_1), (X_2, Y_2), \dots).
\end{aligned}$$

We also have

$$\mathbf{E}^{(k)}\{S_{2,n,k}\} \leq \mathbf{E}^{(k)}\{(g_{n,k}((X^{(k)},Y^{(k)}),((X_i^{*(k)},Y_i^{*(k)}))_{i\in I_{n,k}}))_+\}$$

(where $z_+ = \max\{z,0\}$ for $z \in \mathbb{R}$), and the $(X_i^{*(k)},Y_i^{*(k)})$ $(i \in I_{n,k})$ are identically distributed for each $k \in \{1,\dots,N_n\}$. This implies, that there exists $K_n \in \left\{ \left\lfloor \frac{n}{\lfloor (\log n)^2 \rfloor} \right\rfloor, \left\lceil \frac{n}{\lfloor (\log n)^2 \rfloor} \right\rceil \right\}$ and random variables

$$(\bar{X},\bar{Y}),(\bar{X}_1,\bar{Y}_1),(\bar{X}_2,\bar{Y}_2),\dots,(\bar{X}_{K_n},\bar{Y}_{K_n})$$

on a probability space $(\bar{\Omega},\bar{\mathcal{A}},\bar{\mathbf{P}})$ that are i.i.d and have the same distribution as (X,Y) such that

$$\begin{aligned}
&\max_{k\in\{1,\dots,N_n\}} \mathbf{E}^{(k)}\{S_{2,n,k}\} = \mathbf{E}^{(k^*)}\{S_{2,n,k}\} \\
&\quad \leq \bar{\mathbf{E}}\left\{ \left(\sup_{f\in\mathcal{F}_n} \left(\bar{\mathbf{E}}\left\{ |(T_{\kappa_n} f)(\bar{X}) - T_{\kappa_n}\bar{Y}|^2 \right\} - \bar{\mathbf{E}}\{ |m_{\kappa_n}(\bar{X}) - T_{\kappa_n}\bar{Y}|^2 \right. \right. \right. \right. \\
&\qquad\qquad \left. \left. \left. \left. - 2\cdot \frac{1}{K_n}\sum_{i=1}^{K_n} \left(|(T_{\kappa_n} f)(\bar{X}_i) - T_{\kappa_n}\bar{Y}_i|^2 - |m_{\kappa_n}(\bar{X}_i) - T_{\kappa_n}\bar{Y}_i|^2 \right) \right) \right)_+ \right\} \\
&\quad =: \bar{\mathbf{E}}\{S_{3,n}\},
\end{aligned}$$

e.g.,

$$(\bar{X},\bar{Y}) = (X,Y), \{(\bar{X}_i,\bar{Y}_i)\}_{i=1}^{K_n} = \{(X_j^*,Y_j^*)\}_{j\in I_{n,k^*}},$$

and

$$(\bar{\Omega},\bar{\mathcal{A}},\bar{\mathbf{P}}) = (\Omega^{(k^*)},\mathcal{A}^{(k^*)},\mathbf{P}^{(k^*)}).$$

From here we proceed as in the proof of Theorem 1 in Kohler (2026) to get the assertion. We have for any $u > 0$

$$\begin{aligned}
&\bar{\mathbf{P}}(S_{3,n} > u) \\
&\quad \leq \bar{\mathbf{P}}\left(\exists f \in \mathcal{F}_n : \bar{\mathbf{E}}\left\{ |(T_{\kappa_n} f)(\bar{X}) - T_{\kappa_n}\bar{Y}|^2 \right\} - \bar{\mathbf{E}}\{ |m_{\kappa_n}(\bar{X}) - T_{\kappa_n}\bar{Y}|^2 \right. \\
&\qquad\qquad \left. - 2\cdot \frac{1}{K_n}\sum_{i=1}^{K_n} \left(|(T_{\kappa_n} f)(\bar{X}_i) - T_{\kappa_n}\bar{Y}_i|^2 - |m_{\kappa_n}(\bar{X}_i) - T_{\kappa_n}\bar{Y}_i|^2 \right) > u \right) \\
&\quad \leq \bar{\mathbf{P}}\left(\exists f \in \mathcal{F}_n : \bar{\mathbf{E}}\left\{ \left| \frac{(T_{\kappa_n} f)(\bar{X})}{\kappa_n} - \frac{T_{\kappa_n}\bar{Y}}{\kappa_n} \right|^2 \right\} - \bar{\mathbf{E}}\left\{ \left| \frac{m_{\kappa_n}(\bar{X})}{\kappa_n} - \frac{T_{\kappa_n}\bar{Y}}{\kappa_n} \right|^2 \right\} \right. \\
&\qquad\qquad \left. - \frac{1}{K_n}\sum_{i=1}^{K_n} \left(\left| \frac{(T_{\kappa_n} f)(\bar{X}_i)}{\kappa_n} - \frac{T_{\kappa_n}\bar{Y}_i}{\kappa_n} \right|^2 - \left| \frac{m_{\kappa_n}(\bar{X}_i)}{\kappa_n} - \frac{T_{\kappa_n}\bar{Y}_i}{\kappa_n} \right|^2 \right) \right. \\
&\qquad > \left. \frac{1}{2}\cdot \left(\frac{u}{\kappa_n^2} + \bar{\mathbf{E}}\left\{ \left| \frac{(T_{\kappa_n} f)(\bar{X})}{\kappa_n} - \frac{T_{\kappa_n}\bar{Y}}{\kappa_n} \right|^2 \right\} - \bar{\mathbf{E}}\left\{ \left| \frac{m_{\kappa_n}(\bar{X})}{\kappa_n} - \frac{T_{\kappa_n}\bar{Y}}{\kappa_n} \right|^2 \right\} \right) \right).
\end{aligned}$$

Theorem 11.4 in Györfi et al. (2002) allows us to conclude

$$\begin{aligned}
&\bar{\mathbf{P}}(S_{3,n} > u) \\
&\quad \leq 14 \cdot \sup_{x_1^n \in \mathrm{supp}(X)^n} \mathcal{N}_1\left(\frac{u}{80 \cdot \kappa_n^2}, \left\{\frac{1}{\kappa_n} \cdot f : f \in \mathcal{F}_n\right\}, x_1^n\right) \cdot \exp\left(-\frac{u \cdot K_n}{5136 \cdot \kappa_n^2}\right) \\
&\quad \leq 14 \cdot \sup_{x_1^n \in \mathrm{supp}(X)^n} \mathcal{N}_1\left(\frac{u}{80 \cdot \kappa_n}, \mathcal{F}_n, x_1^n\right) \cdot \exp\left(-c_{41} \cdot \frac{u \cdot n}{\kappa_n^2 \cdot (\log n)^2}\right).
\end{aligned}$$

Note that we take $\sup_{x_1^n \in \mathrm{supp}(X)^n}$ instead of $\sup_{x_1^n \in (\mathbb{R}^d)^n}$, which can be concluded from the proof of Theorem 11.4. Finally we can conclude for any $\epsilon > 0$

$$\begin{aligned}
\mathbf{E}\{S_{3,n}\} &\leq \int_0^\epsilon \mathbf{P}(S_{3,n} > u)\, du + \int_\epsilon^\infty \mathbf{P}(S_{3,n} > u)\, du \\
&\leq \epsilon + \int_\epsilon^\infty 14 \cdot \sup_{x_1^n \in \mathrm{supp}(X)} \mathcal{N}_1\left(\frac{u}{80 \cdot \kappa_n}, \mathcal{F}_n, x_1^n\right) \cdot \exp\left(-c_{41} \cdot \frac{u \cdot n}{\kappa_n^2 \cdot (\log n)^2}\right) du.
\end{aligned}$$

□

To bound the covering number in Lemma 8 we proceed similarly as in Kohler (2026) and extend Lemma 12 from Kohler (2026) to our setting.

Lemma 10 *Let $\mathcal{M}$ be a d^*-dimensional Lipschitz-manifold. Let $k \geq 3$, $\kappa \geq 1$ and let $A, B, C \geq 1$. Let $\sigma : \mathbb{R} \to \mathbb{R}$ be k-times differentiable such that all derivatives up to order k are bounded on $\mathbb{R}$. Let $L, r, K_n \in \mathbb{N}$ and let $\mathcal{F}$ be the set of all functions $f_{\mathbf{w}}$ defined by*

$$f_{\mathbf{w}}(x) = \sum_{j=1}^{K_n} w_{1,1,j}^{(L)} \cdot f_{j,1}^{(L)}(x) \tag{37}$$

for some $w_{1,1,1}^{(L)}, \ldots, w_{1,1,K_n}^{(L)} \in \mathbb{R}$, where $f_{j,1}^{(L)} = f_{\mathbf{w},j,1}^{(L)}$ are recursively defined by

$$f_{k,i}^{(l)}(x) = f_{\mathbf{w},k,i}^{(l)}(x) = \sigma\left(\sum_{j=1}^{r} w_{k,i,j}^{(l-1)} \cdot f_{k,j}^{(l-1)}(x) + w_{k,i,0}^{(l-1)}\right) \tag{38}$$

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

$$f_{k,i}^{(1)}(x) = f_{\mathbf{w},k,i}^{(1)}(x) = \sigma\left(\sum_{j=1}^{d} w_{k,i,j}^{(0)} \cdot x^{(j)} + w_{k,i,0}^{(0)}\right) \tag{39}$$

for some $w_{k,i,0}^{(0)}, \ldots, w_{k,i,d}^{(0)} \in \mathbb{R}$, where the weight vector $\mathbf{w}$ satisfies

$$\sum_{j=1}^{K_n} |w_{1,1,j}^{(L)}| \leq C, \tag{40}$$

$$|w_{k,i,j}^{(l)}| \leq B \quad (k \in \{1,\dots,K_n\}, i,j \in \{1,\dots,r\}, l \in \{1,\dots,L-1\}) \tag{41}$$

and

$$|w_{k,i,j}^{(0)}| \leq A \quad (k \in \{1,\dots,K_n\}, i \in \{1,\dots,r\}, j \in \{1,\dots,d\}). \tag{42}$$

Then we have for any $1 \leq p < \infty$, $0 < \epsilon < 1$ *and* $x_1^n \in \mathcal{M}^n$

$$\mathcal{N}_p\left(\epsilon, \{T_\kappa f \,:\, f \in \mathcal{F}\}, x_1^n\right) \leq \left(c_{42} \cdot \frac{\kappa^p}{\epsilon^p}\right)^{c_{43} \cdot B^{(L-1)\cdot d^*} \cdot A^{d^*} \cdot \left(\frac{C}{\epsilon}\right)^{d^*/k} + c_{44}}.$$

Remark 3. It follows from Lemma 12 in Kohler (2026) that the assertion also holds if $d^* = d$ and $\mathcal{M}$ is contained in some compact subset of $\mathbb{R}^d$ (but not necessarily a Lipschitz-manifold).

To prove Lemma 10 we need the following result which is a slight modification of Lemma 1 a) in Kohler, Langer and Reif (2023).

Lemma 11 *Let* $\mathcal{M}$ *be a* d^**-dimensional Lipschitz-manifold. Let* $h \in (0,1]$ *and let* $\mathcal{P}$ *be a partition of* $\mathbb{R}^d$ *into cubes with side length* h*. Then*

$$|\{C \in \mathcal{P} \,:\, C \cap \mathcal{M} \neq \emptyset\}| \leq c_{45} \cdot \left(\frac{1}{h}\right)^{d^*},$$

where $c_{45} = r \cdot (2 \cdot C_{\psi,2} \cdot \sqrt{d^*} + 4)^{d^*}$.

Proof. For $k_1,\dots,k_{d^*} \in \{0,1,\dots,\lceil 1/h \rceil - 1\}$ set

$$A_{k_1,\dots,k_{d^*}} = [k_1 \cdot h, \min\{(k_1+1)\cdot h, 1\}) \times \cdots \times [k_{d^*} \cdot h, \min\{(k_{d^*}+1)\cdot h, 1\}).$$

We can conclude from the definition of a Lipschitz-manifold

$$\begin{aligned}\mathcal{M} &= \bigcup_{j=1}^{r} \mathcal{M} \cap U_j = \bigcup_{j=1}^{r} \psi_j\left((0,1)^{d^*}\right) \\ &\subseteq \bigcup_{j=1}^{r} \bigcup_{k_1,\dots,k_{d^*} \in \{0,1,\dots,\lceil 1/h \rceil - 1\}} \psi_j\left(A_{k_1,\dots,k_{d^*}}\right).\end{aligned}$$

This implies

$$\begin{aligned}&|\{C \in \mathcal{P} \,:\, C \cap \mathcal{M} \neq \emptyset\}| \\ &\leq \sum_{j=1}^{r} \sum_{k_1=0}^{\lceil \frac{1}{h} \rceil - 1} \cdots \sum_{k_{d^*}=0}^{\lceil \frac{1}{h} \rceil - 1} |\{C \in \mathcal{P} \,:\, C \cap \psi_j\left(A_{k_1,\dots,k_{d^*}}\right) \neq \emptyset\}| \\ &\leq r \cdot 2^{d^*} \cdot \left(\frac{1}{h}\right)^{d^*} \cdot \max_{j=1,\dots,r} |\{C \in \mathcal{P} \,:\, C \cap \psi_j\left(A_{k_1,\dots,k_{d^*}}\right) \neq \emptyset\}|\end{aligned}$$

and it suffices to show

$$\max_{j=1,\dots,r} |\{C \in \mathcal{P} \,:\, C \cap \psi_j\left(A_{k_1,\dots,k_{d^*}}\right) \neq \emptyset\}| \leq (C_{\psi,2} \cdot \sqrt{d^*} + 2)^{d^*}. \tag{43}$$

Fix $j \in \{1,\dots,r\}$. Each point of a cube

$$[k_1 \cdot h, (k_1+1) \cdot h) \times \cdots \times [k_{d^*} \cdot h, (k_{d^*}+1) \cdot h)$$

has a distance of at most $\frac{1}{2} \cdot \sqrt{d^*} \cdot h$ from its center. Hence, $A_{k_1,\dots,k_{d^*}}$ is contained in a ball of radius $\frac{1}{2} \cdot \sqrt{d^*} \cdot h$, and by Lipschitz continuity $\psi_j(A_{k_1,\dots,k_{d^*}})$ is contained in a cube of side length $C_{\psi,2} \cdot \sqrt{d^*} \cdot h$. The number of cubes from $\mathcal{P}$ that intersect that cube is bounded from above by

$$\left(\frac{C_{\psi,2} \cdot \sqrt{d^*} \cdot h}{h} + 2\right)^{d^*} = (C_{\psi,2} \cdot \sqrt{d^*} + 2)^{d^*},$$

since each cube from the partition $\mathcal{P}$ has side length h. □

Proof of Lemma 10. Let $R > 0$ be such that $\mathcal{M} \subseteq \{x \in \mathbb{R}^d \,:\, \|x\| \leq R\}$. The first step of the proof is to show, that for any $f_{\mathbf{w}} \in \mathcal{F}$, any $x \in \mathbb{R}^d$ with $\|x\| \leq R$, any $k \in \mathbb{N}$ and any $s_1,\dots,s_k \in \{1,\dots,d\}$

$$\left|\frac{\partial^k f_{\mathbf{w}}}{\partial x^{(s_1)} \dots \partial x^{(s_k)}}(x)\right| \leq c_{46} \cdot C \cdot B^{(L-1)\cdot k} \cdot A^k =: c. \tag{44}$$

This follows from the first step of the proof of Lemma 12 in Kohler (2024).

In the second step of the proof we will show

$$\mathcal{N}_p\left(\epsilon, \{T_\kappa f \,:\, f \in \mathcal{F}\}, x_1^n\right) \leq \mathcal{N}_p\left(\frac{\epsilon}{2}, T_\kappa \mathcal{G} \circ \Pi, x_1^n\right), \tag{45}$$

where $\mathcal{G}$ denotes the set of all polynomials of degree less than or equal to $k-1$ and Π is a partition of $\mathbb{R}^d$ into cubes of side length $\left(c_{47} \cdot \frac{\epsilon}{c}\right)^{1/k}$, where $c_{47} = c_{47}(d,k) > 0$ is a suitable small constant. Here $T_\kappa \mathcal{G} \circ \Pi$ is the class of all functions whose restriction on each cube in Π lies in $T_\kappa \mathcal{G}$.

Let $(T f_{\mathbf{w}})_{k-1,u}$ be the multivariate Taylor polynomial of $f_{\mathbf{w}}$ of degree $k-1$ around $u \in \mathbb{R}^d$, i.e.

$$\begin{aligned}&(Tf)_{k-1,u}(x)\\ &= \sum_{\substack{j_1,\dots,j_d \in \mathbb{N}_0,\\ j_1+\cdots+j_d \leq k-1}} \frac{1}{j_1! \cdots j_d!} \cdot \frac{\partial^{j_1+\cdots+j_d} f}{\partial^{j_1} x^{(1)} \dots \partial^{j_d} x^{(d)}}(u) \cdot (x^{(1)} - u^{(1)})^{j_1} \cdots (x^{(d)} - u^{(d)})^{j_d}.\end{aligned}$$

For each $I \in \Pi$ fix some $u \in I$. By a multivariate Taylor theorem we get as in the proof of Lemma 1 in Kohler (2014) for each $x \in I$

$$|f_{\mathbf{w}}(x) - (T f_{\mathbf{w}})_{k-1,u}(x)|$$

$$
\begin{aligned}
&= \Bigg| f_{\mathbf{w}}(x) - (T f_{\mathbf{w}})_{k-2,u}(x) \\
&\quad - \sum_{\substack{j_1,\dots,j_d \in \mathbb{N}_0, \\ j_1+\cdots+j_d=k-1}} \frac{1}{j_1! \cdots j_d!} \cdot \frac{\partial^{j_1+\cdots+j_d} f}{\partial^{j_1} x^{(1)} \dots \partial^{j_d} x^{(d)}}(u) \cdot (x^{(1)} - u^{(1)})^{j_1} \cdots (x^{(d)} - u^{(d)})^{j_d} \Bigg| \\
&= \Bigg| \sum_{\substack{j_1,\dots,j_d \in \mathbb{N}_0, \\ j_1+\cdots+j_d=k-1}} \frac{k-1}{j_1! \cdots j_d!} \cdot \int_0^1 (1-t)^{k-2} \cdot \frac{\partial^{j_1+\cdots+j_d} f}{\partial^{j_1} x^{(1)} \dots \partial^{j_d} x^{(d)}}(u + t \cdot (x-u))\, dt \\
&\qquad \cdot (x^{(1)} - u^{(1)})^{j_1} \cdots (x^{(d)} - u^{(d)})^{j_d} \\
&\quad - \sum_{\substack{j_1,\dots,j_d \in \mathbb{N}_0, \\ j_1+\cdots+j_d=k-1}} \frac{1}{j_1! \cdots j_d!} \cdot \frac{\partial^{j_1+\cdots+j_d} f}{\partial^{j_1} x^{(1)} \dots \partial^{j_d} x^{(d)}}(u) \cdot (x^{(1)} - u^{(1)})^{j_1} \cdots (x^{(d)} - u^{(d)})^{j_d} \Bigg| \\
&= \Bigg| \sum_{\substack{j_1,\dots,j_d \in \mathbb{N}_0, \\ j_1+\cdots+j_d=k-1}} \frac{k-1}{j_1! \cdots j_d!} \cdot (x^{(1)} - u^{(1)})^{j_1} \cdots (x^{(d)} - u^{(d)})^{j_d} \\
&\quad \cdot \int_0^1 (1-t)^{k-2} \cdot \left(\frac{\partial^{j_1+\cdots+j_d} f}{\partial^{j_1} x^{(1)} \dots \partial^{j_d} x^{(d)}}(u + t \cdot (x-u)) - \frac{\partial^{j_1+\cdots+j_d} f}{\partial^{j_1} x^{(1)} \dots \partial^{j_d} x^{(d)}}(u) \right) dt \Bigg| \\
&\leq \sum_{\substack{j_1,\dots,j_d \in \mathbb{N}_0, \\ j_1+\cdots+j_d=k-1}} \frac{k-1}{j_1! \cdots j_d!} \cdot |x^{(1)} - u^{(1)}|^{j_1} \cdots |x^{(d)} - u^{(d)}|^{j_d} \\
&\quad \cdot \int_0^1 (1-t)^{k-2} \cdot \left| \frac{\partial^{j_1+\cdots+j_d} f}{\partial^{j_1} x^{(1)} \dots \partial^{j_d} x^{(d)}}(u + t \cdot (x-u)) - \frac{\partial^{j_1+\cdots+j_d} f}{\partial^{j_1} x^{(1)} \dots \partial^{j_d} x^{(d)}}(u) \right| dt.
\end{aligned}
$$

By the mean value theorem and (44) we get for any $j_1, \dots, j_d \in \mathbb{N}_0$ with $j_1 + \cdots + j_d = k-1$

$$
\begin{aligned}
&\left| \frac{\partial^{j_1+\cdots+j_d} f}{\partial^{j_1} x^{(1)} \dots \partial^{j_d} x^{(d)}}(u + t \cdot (x-u)) - \frac{\partial^{j_1+\cdots+j_d} f}{\partial^{j_1} x^{(1)} \dots \partial^{j_d} x^{(d)}}(u) \right| \\
&\leq \sum_{i=1}^{d} \Bigg| \frac{\partial^{j_1+\cdots+j_d} f}{\partial^{j_1} x^{(1)} \dots \partial^{j_d} x^{(d)}}(u^{(1)}, \dots, u^{(i-1)}, u^{(i)} + t \cdot (x^{(i)} - u^{(i)}), \dots, u^{(d)} + t \cdot (x^{(d)} - u^{(d)})) \\
&\quad - \frac{\partial^{j_1+\cdots+j_d} f}{\partial^{j_1} x^{(1)} \dots \partial^{j_d} x^{(d)}}(u^{(1)}, \dots, u^{(i)}, u^{(i+1)} + t \cdot (x^{(i+1)} - u^{(i+1)}), \dots, u^{(d)} + t \cdot (x^{(d)} - u^{(d)})) \Bigg| \\
&\leq \sum_{i=1}^{d} c \cdot |u^{(i)} + t \cdot (x^{(i)} - u^{(i)}) - u^{(i)}| \\
&= c \cdot t \cdot \sum_{i=1}^{d} |x^{(i)} - u^{(i)}|.
\end{aligned}
$$

Hence,

$$
|f_{\mathbf{w}}(x) - (T f_{\mathbf{w}})_{k-1,u}(x)|
$$

$$
\begin{aligned}
&\leq \sum_{\substack{j_1,\cdots,j_d\in\mathbb{N}_0,\\ j_1+\cdots+j_d=k-1}} \frac{k-1}{j_1!\cdots j_d!}\cdot c\cdot\sum_{i=1}^{d}\left|x^{(i)}-u^{(i)}\right|\cdot\left|x^{(1)}-u^{(1)}\right|^{j_1}\cdots\left|x^{(d)}-u^{(d)}\right|^{j_d} \\
&\qquad\qquad \cdot\int_0^1 t\cdot(1-t)^{k-2}dt \\
&= \sum_{\substack{j_1,\cdots,j_d\in\mathbb{N}_0,\\ j_1+\cdots+j_d=k-1}} \frac{1}{j_1!\cdots j_d!\cdot k}\cdot c\cdot\sum_{i=1}^{d}\left|x^{(i)}-u^{(i)}\right|\cdot\left|x^{(1)}-u^{(1)}\right|^{j_1}\cdots\left|x^{(d)}-u^{(d)}\right|^{j_d} \\
&\leq \sum_{\substack{j_1,\cdots,j_d\in\mathbb{N}_0,\\ j_1+\cdots+j_d=k}} \frac{1}{j_1!\cdots j_d!}\cdot c\cdot\left|x^{(1)}-u^{(1)}\right|^{j_1}\cdots\left|x^{(d)}-u^{(d)}\right|^{j_d} \\
&= \frac{c}{k!}\cdot\left(\sum_{i=1}^{d}\left|x^{(i)}-u^{(i)}\right|\right)^k \\
&\leq c\cdot\frac{d^k}{k!}\cdot\|x-u\|_\infty^k \leq c\cdot\frac{d^k}{k!}\cdot c_{47}\cdot\frac{\epsilon}{c} = c_{47}\cdot\frac{d^k}{k!}\cdot\epsilon,
\end{aligned}
$$

where we have used the multinomial theorem. By repeating this argument for every cube $I\in\Pi$ and by choosing c_{47} small enough we can see that for each $f_{\mathbf{w}}$ we can find $g\in\mathcal{G}\circ\Pi$ such that

$$|f_{\mathbf{w}}(x)-g(x)|\leq\frac{\epsilon}{2}$$

holds for all $x\in\mathbb{R}^d$, which implies (45). The last step is to complete the proof of Lemma 10. For

$$\Pi^*:=\{I\in\Pi: I\cap\mathcal{M}\neq\emptyset\}\cup\left\{\mathbb{R}^d\setminus\bigcup_{I\in\Pi: I\cap\mathcal{M}\neq\emptyset} I\right\}$$

we have

$$\mathcal{N}_p\left(\frac{\epsilon}{2},T_\kappa\mathcal{G}\circ\Pi,x_1^n\right)\leq\mathcal{N}_p\left(\frac{\epsilon}{2},T_\kappa\mathcal{G}\circ\Pi^*,x_1^n\right), \tag{46}$$

since $x_1^n\in\mathcal{M}^n$. There are

$$\binom{d+k-1}{d}$$

many monomials in d variables of degree at most $k-1$. This together with Lemma 11 allows us to conclude that $\mathcal{G}\circ\Pi^*$ is a linear vector space of dimension

$$\binom{d+k-1}{d}\cdot|\Pi^*|\leq c_{48}\cdot\left(\frac{c}{\epsilon}\right)^{d^*/k}.$$

With this we derive from Theorem 9.4 and Theorem 9.5 in Györfi et al. (2002)

$$\mathcal{N}_p(\frac{\epsilon}{2},T_\kappa\mathcal{G}\circ\Pi^*,x_1^n)\leq 3\left(\frac{2e(2\kappa)^p}{(\epsilon/2)^p}\log\left(\frac{3e(2\kappa)^p}{(\epsilon/2)^p}\right)\right)^{c_{48}\cdot\left(\frac{c}{\epsilon}\right)^{d^*/k}+1}. \tag{47}$$

The claim follows from (45), (46) and (47).

□

4.4. Proof of Theorem 1

We prove a) and b) at the same time. To do this, we let $d^* \in \{1, \ldots, d\}$, assume

$$supp(X) \subseteq \mathcal{M},$$

and assume that for $d^* < d$ $\mathcal{M}$ is a d^*-dimensional Lipschitz-manifold, and for $d^* = d$ we assume that $\mathcal{M}$ is contained in a compact subset of $\mathbb{R}^d$.

We give a proof similar to the proof of Theorem 1 in Kohler (2026). Throughout the proof we assume w.l.o.g. that n is sufficiently large and that $\|m\|_\infty \leq \kappa_n$ holds. Since $\mathcal{M}$ is bounded, we can choose $A \geq 1$ with $\mathcal{M} \subseteq [-A, A)^d$. Set

$$K = \left\lceil c_{49} \cdot n^{\frac{1}{2p+d^*}} \right\rceil,$$

$$\tilde{K}_n = r \cdot K^d \text{ and}$$

$$N_n = \left\lceil c_{50} \cdot n^{1+\frac{4p+3d}{2p+d^*}} \right\rceil.$$

Set

$$J_n = K_n \cdot (r + 2 + (L-2) \cdot r \cdot (r+1) + r \cdot (d+1))$$

and

$$J_n^* = N_n \cdot \tilde{K}_n \cdot (r + 2 + (L-2) \cdot r \cdot (r+1) + r \cdot (d+1)).$$

By using Theorem 3 with K, A, L and r we know that there exists a weight vector

$$\mathbf{w}^* \in \mathbb{R}^{J_n^*}$$

of a neural network

$$f_{\mathbf{w}^*}(x) = \sum_{k=1}^{N_n \cdot \tilde{K}_n} (\mathbf{w}^*)_{1,1,k}^{(L)} \cdot f_{\mathbf{w}^*,k,1}^{(L)}(x),$$

where each $f_{\mathbf{w}^*,k,1}^{(L)}$ is a subnetwork consisting of L layers and r neurons per layer, such that

$$\sup_{x \in [-A,A)^d} |f_{\mathbf{w}^*}(x) - m(x)| \leq \frac{c_{51}}{\tilde{K}_n^{p/d}} \tag{48}$$

and

$$|\, (\mathbf{w}^*)_{1,1,k}^{(L)} \,| \leq \frac{c_{52} \cdot \tilde{K}_n^{(q+d)/d}}{N_n} \quad (k = 1, \ldots, N_n \cdot \tilde{K}_n).$$

Note that here we replace each of the f_k's in the outer sum of the space $\mathcal{H}$ of Theorem 3 with

$$f_k = \frac{1}{N_n} \sum_{i=1}^{N_n} f_k.$$

Furthermore, the weights of this network satisfy

$$|\, (\mathbf{w}^*)_{k,i,j}^{(l)} \,| \leq c_{53} \quad \text{for } l = 1, \ldots, L-1$$

and

$$|\,(\mathbf{w}^*)_{k,i,j}^{(0)}\,| \le c_{54} \cdot (\log n) \cdot n^{\tau}.$$

Set

$$\epsilon_n = \frac{c_{55}}{n \cdot \sqrt{N_n \cdot \tilde{K}_n}} \ge \frac{c_{56}}{n^{(2p+2d)\cdot\tau+1.5}},$$

where the inequality follows from

$$n \cdot \sqrt{N_n \cdot \tilde{K}_n} \le c_{57} \cdot n^{\frac{3}{2}} \cdot n^{\frac{4p+3d}{4p+2d^*}} \cdot n^{\frac{d}{4p+2d^*}} = c_{57} \cdot n^{(2p+2d)\cdot\tau+1.5}.$$

Let A_n be the event such that firstly there exist pairwise distinct $j_1, \dots, j_{N_n \cdot \tilde{K}_n} \in \{1, \dots, K_n\}$ such that the weight vector $\mathbf{w}^{(0)} \in \mathbb{R}^{J_n}$ satisfies

$$|(\mathbf{w}^{(0)})_{j_s,k,i}^{(l)} - (\mathbf{w}^*)_{s,k,i}^{(l)}| \le \epsilon_n \quad \text{for all } k, i \text{ and } l \in \{0, \dots, L-1\}, s \in \{1, \dots, N_n \cdot \tilde{K}_n\}$$

and such that secondly

$$\max_{i=1,\dots,n} |Y_i|^2 \le \kappa_n$$

holds. In the proof we split the L_2 error of m_n into a sum of several terms and bound each term separately. In the following we set

$$m_{\kappa_n}(x) = \mathbf{E}\{T_{\kappa_n} Y | X = x\}.$$

Then we have

$$\begin{aligned}
&\int |m_n(x) - m(x)|^2 \mathbf{P}_X(dx) \\
&= \left(\mathbf{E}\left\{|m_n(X) - Y|^2 | \mathcal{D}_n\right\} - \mathbf{E}\{|m(X) - Y|^2\}\right) \cdot 1_{A_n} \\
&\quad + \int |m_n(x) - m(x)|^2 \mathbf{P}_X(dx) \cdot 1_{A_n^c} \\
&= \Big[\mathbf{E}\left\{|m_n(X) - Y|^2 | \mathcal{D}_n\right\} - \mathbf{E}\{|m(X) - Y|^2\} \\
&\qquad\qquad - \left(\mathbf{E}\left\{|m_n(X) - T_{\kappa_n} Y|^2 | \mathcal{D}_n\right\} - \mathbf{E}\{|m_{\kappa_n}(X) - T_{\kappa_n} Y|^2\}\right)\Big] \cdot 1_{A_n} \\
&\quad + \Big[\mathbf{E}\left\{|m_n(X) - T_{\kappa_n} Y|^2 | \mathcal{D}_n\right\} - \mathbf{E}\{|m_{\kappa_n}(X) - T_{\kappa_n} Y|^2\} \\
&\qquad\qquad - 2 \cdot \frac{1}{n} \sum_{i=1}^{n} \left(|m_n(X_i) - T_{\kappa_n} Y_i|^2 - |m_{\kappa_n}(X_i) - T_{\kappa_n} Y_i|^2\right)\Big] \cdot 1_{A_n} \\
&\quad + \Big[2 \cdot \frac{1}{n} \sum_{i=1}^{n} |m_n(X_i) - T_{\kappa_n} Y_i|^2 - 2 \cdot \frac{1}{n} \sum_{i=1}^{n} |m_{\kappa_n}(X_i) - T_{\kappa_n} Y_i|^2 \\
&\qquad\qquad - \left(2 \cdot \frac{1}{n} \sum_{i=1}^{n} |m_n(X_i) - Y_i|^2 - 2 \cdot \frac{1}{n} \sum_{i=1}^{n} |m(X_i) - Y_i|^2\right)\Big] \cdot 1_{A_n} \\
&\quad + \Big[2 \cdot \frac{1}{n} \sum_{i=1}^{n} |m_n(X_i) - Y_i|^2 - 2 \cdot \frac{1}{n} \sum_{i=1}^{n} |m(X_i) - Y_i|^2\Big] \cdot 1_{A_n}
\end{aligned}$$

$$
\begin{aligned}
&+ \int |m_n(x) - m(x)|^2 \mathbf{P}_X(dx) \cdot 1_{A_n^c} \\
&=: \sum_{j=1}^{5} T_{j,n}.
\end{aligned}
$$

In the first step of the proof we show

$$\mathbf{E}\{T_{j,n}\} \leq c_{58} \cdot \frac{\log n}{n} \quad \text{for } j \in \{1, 3\}.$$

This follows as in the proof of Lemma 1 in Bauer and Kohler (2019).
In the second step of the proof we show

$$\mathbf{E}\{T_{5,n}\} \leq c_{59} \cdot \frac{(\log n)^2}{n^2}.$$

Since we assume $\|m\|_\infty \leq \kappa_n$, we have

$$\int |m_n(x) - m(x)|^2 \mathbf{P}_X(dx) \leq \int |2 \cdot \kappa_n|^2 \mathbf{P}_X(dx) \leq 4 \cdot c_8^2 \cdot (\log n)^2$$

and it suffices to show

$$\mathbf{P}(A_n^c) \leq \frac{c_{60}}{n^2}. \tag{49}$$

We consider the initial choice of the weights for the K_n fully connected neural networks sequentially. The weight in the first of these networks differs in all components in layers $l = 1, \ldots, L-1$ by at most ϵ_n from $(\mathbf{w}^*)_{1,i,j}^{(l)}$ with probability bounded from below by

$$
\begin{aligned}
&\left(\frac{\epsilon_n}{2 \cdot c_6}\right)^{r \cdot (r+1) \cdot (L-1)} \cdot \left(\frac{\epsilon_n}{2 \cdot c_7 \cdot (\log n) \cdot n^\tau}\right)^{r \cdot (d+1)} \\
&\geq \left(\frac{c_{56}}{2 \cdot c_6 \cdot n^{(2p+2d) \cdot \tau + 1.5}}\right)^{r \cdot (r+1) \cdot (L-1)} \cdot \left(\frac{c_{56}}{2 \cdot c_7 \cdot (\log n) \cdot n^\tau \cdot n^{(2p+2d) \cdot \tau + 1.5}}\right)^{r \cdot (d+1)} \\
&\geq n^{-((2p+2d) \cdot \tau + 1.5) \cdot r \cdot (r+1) \cdot (L-1) - ((2p+2d) \cdot \tau + 1.5) \cdot r \cdot (d+1) - \tau \cdot r \cdot (d+1) - 0.5} \\
&= n^{-\eta - 0.5},
\end{aligned}
$$

where

$$\eta = ((2p+2d) \cdot \tau + 1.5) \cdot r \cdot (r+1) \cdot (L-1) + ((2p+2d) \cdot \tau + 1.5) \cdot r \cdot (d+1) + \tau \cdot r \cdot (d+1).$$

Therefore, the probability that none of the first $n^{\eta+1}$ neural networks satisfy this condition is bounded from above by

$$\left(1 - n^{-\eta - 0.5}\right)^{n^{\eta+1}} \leq \left(\exp\left(-n^{-\eta-0.5}\right)\right)^{n^{\eta+1}} = \exp\left(-n^{0.5}\right).$$

Assumption (11) implies that $K_n \geq n^{\eta+1} \cdot N_n \cdot \tilde{K}_n$ for all sufficiently large n. Consequently, we can conclude, that the probability that there exists $s \in \{1, \ldots, N_n \cdot \tilde{K}_n\}$ such that all

K_n weight vectors differ from $((\mathbf{w}^*)^{(l)}_{s,k,i})_{k,i,l}$ in at least one component by more than ϵ_n is bounded from above by

$$N_n \cdot \tilde{K}_n \cdot \exp(-n^{0.5}) \qquad \leq c_{61} \cdot n^{(4p+3d)\cdot\tau+1} \cdot n^{d\cdot\tau} \cdot \exp(-n^{0.5}) \leq \frac{c_{62}}{n^2}.$$

We conclude with Markov's inequality, (A1) and $c_8 \cdot c_4 \geq 3$

$$\begin{aligned}
\mathbf{P}(A_n^c) &\leq \frac{c_{62}}{n^2} + \mathbf{P}\{\max_{i=1,\ldots,n} Y_i^2 > \kappa_n\} \leq \frac{c_{62}}{n^2} + n \cdot \mathbf{P}\{Y^2 > \kappa_n\} \\
&= \frac{c_{62}}{n^2} + n \cdot \mathbf{P}\{\exp(c_4 \cdot Y^2) > \exp(c_4 \cdot \kappa_n)\} \leq \frac{c_{62}}{n^2} + n \cdot \frac{\mathbf{E}\{\exp(c_4 \cdot Y^2)\}}{\exp(c_4 \cdot \kappa_n)} \\
&= \frac{c_{62}}{n^2} + n \cdot \frac{\mathbf{E}\{\exp(c_4 \cdot Y^2)\}}{\exp(c_8 \cdot c_4 \cdot \log(n))} = \frac{c_{62}}{n^2} + n \cdot \frac{\mathbf{E}\{\exp(c_4 \cdot Y^2)\}}{n^{c_8 \cdot c_4}} \leq \frac{c_{63}}{n^2},
\end{aligned}$$

for n sufficiently large.

Let $\epsilon > 0$ be arbitrary. The third step is to show

$$\mathbf{E}\{T_{2,n}\} \leq c_{64} \cdot \frac{n^{\tau \cdot d^* + \epsilon}}{n} = c_{64} \cdot n^{-\frac{2p}{2p+d^*}+\epsilon}.$$

Let $\mathcal{W}_n$ be the set of all weight vectors $\mathbf{w} = (w^{(l)}_{i,j,k})_{i,j,k,l} \in \mathbb{R}^{J_n}$ which satisfy

$$|w^{(L)}_{1,1,k}| \leq c_{65} \quad (k = 1, \ldots, K_n),$$

$$|w^{(l)}_{i,j,k}| \leq c_{66} \quad (l = 1, \ldots, L-1)$$

and

$$|w^{(0)}_{i,j,k}| \leq c_{67} \cdot (\log n) \cdot n^{\tau}.$$

By Lemma 3, Lemma 4 and Lemma 5 we can conclude

$$\|\mathbf{w}^{(t)} - \mathbf{w}^{(0)}\| \leq c_{68} \quad (t = 1, \ldots, t_n)$$

on A_n. This is done similarly as in the proof of Theorem 2 and uses

$$t_n \cdot \lambda_n \cdot \kappa_n \leq c_{69}.$$

The choice of $\mathbf{w}^{(0)}$ implies for c_{65}, c_{66}, c_{67} sufficiently large that we have on A_n

$$\mathbf{w}^{(t)} \in \mathcal{W}_n \quad (t = 0, \ldots, t_n).$$

This means that

$$m_n \in \mathcal{F}_n = \{T_{\kappa_n} f_{\mathbf{w}} \quad : \quad \mathbf{w} \in \mathcal{W}_n\}$$

and we conclude with Lemma 8 for $u_n > 0$

$$\mathbf{E}\{T_{2,n}\} \leq 4 \cdot \kappa_n^2 \cdot \frac{n}{(\log n)^2} \cdot \beta_{\lceil (\log n)^2 \rceil}((X_1, Y_1), (X_2, Y_2), \ldots) + u_n$$

$$+\int_{u_n}^{\infty} 14 \cdot \sup_{x_1^n \in \mathrm{supp}(X)^n} \mathcal{N}_1\left(\frac{u}{80\cdot\kappa_n}, \mathcal{F}_n, x_1^n\right) \cdot \exp\left(-c_{70}\cdot\frac{u\cdot n}{\kappa_n^2\cdot(\log n)^2}\right) du.$$

Lemma 10 (in case $d^* < d$) and Remark 3 (in case $d^* = d$) imply for $x_1^n \in \mathcal{M}^n$

$$\mathcal{N}_1\left(\frac{u}{80\cdot\kappa_n}, \mathcal{F}_n, x_1^n\right) \le \left(\frac{c_{71}}{u/(80\cdot\kappa_n)}\right)^{c_{72}\cdot(c_{73})^{(L-1)\cdot d^*}\cdot(\log n)^{d^*} n^{\tau\cdot d^*}\cdot\left(\frac{K_n\cdot c_{74}}{u/(80\cdot\kappa_n)}\right)^{d^*/k}+c_{75}}$$

and by choosing k large enough we get for $u > \frac{c_{76}}{n}$

$$\mathcal{N}_1\left(\frac{u}{80\cdot\kappa_n}, \mathcal{F}_n, x_1^n\right) \le c_{77}\cdot n^{c_{78}\cdot n^{\tau\cdot d^*+\epsilon/2}}.$$

We have by (A6)

$$\beta_{\lceil(\log n)^2\rceil}((X_1,Y_1),(X_2,Y_2),\dots) \le c_{79}\cdot e^{-c_{80}\cdot\lceil(\log n)^2\rceil} \le c_{79}\cdot e^{-c_{80}\cdot(\log n)^2}$$

For $u_n \ge c_{81}/n$ we conclude

$$\begin{aligned}
\mathbf{E}\{T_{2,n}\} &\le 4\cdot c_{79}\cdot\kappa_n^2\cdot\frac{n}{(\log n)^2}\cdot e^{-c_{80}\cdot(\log n)^2} + u_n \\
&\quad + \int_{u_n}^{\infty} 14\cdot c_{77}\cdot n^{c_{78}\cdot n^{\tau\cdot d^*+\epsilon/2}}\cdot\exp\left(-c_{82}\cdot\frac{u\cdot n}{\kappa_n^2\cdot(\log n)^2}\right) du \\
&\le 4\cdot c_{79}\cdot\kappa_n^2\cdot\frac{n}{(\log n)^2}\cdot e^{-c_{80}\cdot(\log n)^2} + u_n \\
&\quad + c_{83}\cdot n^{c_{78}\cdot n^{\tau\cdot d^*+\epsilon/2}}\cdot\exp\left(-c_{82}\cdot\frac{u_n\cdot n}{\kappa_n^2\cdot(\log n)^2}\right)\cdot\frac{\kappa_n^2\cdot(\log n)^2}{n}
\end{aligned}$$

The result of the third step of the proof follows by setting

$$u_n = \frac{\kappa_n^2\cdot(\log n)^2}{c_{82}\cdot n}\cdot c_{78}\cdot n^{\tau\cdot d^*+\epsilon/2}\cdot\log n.$$

In the last step of the proof we will show

$$\mathbf{E}\{T_{4,n}\} \le c_{84}\cdot n^{-\frac{2p}{2p+d^*}}.$$

Using

$$|T_{\kappa_n} z - y| \le |z-y| \quad \text{for } |y|\le\kappa_n$$

we get

$$\begin{aligned}
\frac{T_{4,n}}{2} &= \Big[\frac{1}{n}\sum_{i=1}^n |m_n(X_i)-Y_i|^2 - \frac{1}{n}\sum_{i=1}^n |m(X_i)-Y_i|^2\Big]\cdot 1_{A_n} \\
&= \Big[\frac{1}{n}\sum_{i=1}^n |T_{\kappa_n} f_{\mathbf{w}^{(t_n)}}(X_i)-Y_i|^2 - \frac{1}{n}\sum_{i=1}^n |m(X_i)-Y_i|^2\Big]\cdot 1_{A_n}
\end{aligned}$$

$$\leq \Big[\frac{1}{n}\sum_{i=1}^{n}|f_{\mathbf{w}^{(t_n)}}(X_i)-Y_i|^2-\frac{1}{n}\sum_{i=1}^{n}|m(X_i)-Y_i|^2\Big]\cdot 1_{A_n}$$
$$= \Big[F_n(\mathbf{w}^{(t_n)})-\frac{1}{n}\sum_{i=1}^{n}|m(X_i)-Y_i|^2\Big]\cdot 1_{A_n}.$$

On A_n let $\tilde{\mathbf{w}} \in \mathbb{R}^{J_n}$ be an extension of $\mathbf{w}^* \in \mathbb{R}^{J_n^*}$ onto $\mathbb{R}^{J_n}$, where we fill each new component with the value of $\mathbf{w}^{(0)}$, i.e.

$$(\tilde{\mathbf{w}})_{j_s,k,i}^{(l)} = (\mathbf{w}^*)_{s,k,i}^{(l)} \text{ for } s \in \{1,\dots,N_n\cdot\tilde{K}_n\} \text{ and}$$
$$(\tilde{\mathbf{w}})_{j,k,i}^{(l)} = (\mathbf{w}^{(0)})_{j,k,i}^{(l)} \text{ for } j \notin \{j_1,\dots,j_{N_n\cdot\tilde{K}_n}\}.$$

The networks defined by $\mathbf{w}^*$ and $\tilde{\mathbf{w}}$ are identical, since the weights in the outer Layer of $\mathbf{w}^{(0)}$ are all 0. On A_n we have

$$\begin{aligned}
\|\tilde{\mathbf{w}}-\mathbf{w}^{(0)}\|^2 &\leq \sum_{k=1}^{N_n\cdot\tilde{K}_n}|(\tilde{\mathbf{w}})_{1,1,j_k}^{(L)}|^2 + N_n\cdot\tilde{K}_n\cdot L\cdot(r\cdot(r+d))\cdot\epsilon_n^2\\
&\leq N_n\cdot\tilde{K}_n\cdot\left(\frac{c_{52}\cdot\tilde{K}_n^{(q+d)/d}}{N_n}\right)^2 + c_{85}\cdot N_n\cdot\tilde{K}_n\cdot\epsilon_n^2\\
&\leq c_{86}\cdot\frac{\tilde{K}_n^{2\cdot(q+d)/d+1}}{N_n} + c_{87}\cdot\frac{1}{n^2}\\
&\leq c_{88}\cdot n^{-1-\frac{2p}{2p+d^*}} + c_{87}\cdot\frac{1}{n^2} \leq c_{89}\cdot n^{-1-\frac{2p}{2p+d^*}}.
\end{aligned}$$

Here we used

$$\left(2\cdot\frac{q+d}{d}+1\right)\cdot\frac{d}{2p+d^*}-1-\frac{4p+3d}{2p+d^*} = -1-\frac{4p-2q}{2p+d^*} \leq -1-\frac{2p}{2p+d^*}.$$

We get with Theorem 2

$$\begin{aligned}
&\frac{T_{4,n}}{2}\\
&\leq \left(\frac{1}{n}\sum_{i=1}^{n}|f_{\tilde{\mathbf{w}}}(X_i)-Y_i|^2 + c_{90}\cdot(\log n)^2\cdot\|\tilde{\mathbf{w}}-\mathbf{w}^{(0)}\|^2 + c_{91}\cdot\frac{(\log n)^3}{K_n^{3/2}}\right.\\
&\qquad \left.-\frac{1}{n}\sum_{i=1}^{n}|m(X_i)-Y_i|^2\right)\cdot 1_{A_n}\\
&\leq \frac{1}{n}\sum_{i=1}^{n}|f_{\mathbf{w}^*}(X_i)-Y_i|^2 - \frac{1}{n}\sum_{i=1}^{n}|m(X_i)-Y_i|^2 + c_{92}\cdot(\log n)^2\cdot n^{-1-\frac{2p}{2p+d^*}}\\
&\qquad +\frac{1}{n}\sum_{i=1}^{n}|m(X_i)-Y_i|^2\cdot 1_{A_n^c}.
\end{aligned}$$

Using (49) and the C-S inequality, we can conclude

$$\begin{aligned}
&\mathbf{E}\left\{\frac{1}{n}\sum_{i=1}^{n}|f_{\mathbf{w}^*}(X_i)-Y_i|^2-\frac{1}{n}\sum_{i=1}^{n}|m(X_i)-Y_i|^2+\frac{1}{n}\sum_{i=1}^{n}|m(X_i)-Y_i|^2\cdot 1_{A_n^c}\right\}\\
&\leq \mathbf{E}\{|f_{\mathbf{w}^*}(X)-Y|^2\}-\mathbf{E}\left\{|m(X)-Y|^2\right\}+\frac{1}{n}\sum_{i=1}^{n}\sqrt{\mathbf{E}\{|m(X_i)-Y_i|^4\}}\cdot\sqrt{\mathbf{P}\{A_n^c\}}\\
&\leq \int|f_{\mathbf{w}*}(x)-m(x)|^2\mathbf{P}_X(dx)+c_{93}\cdot\frac{(\log n)^2}{n}.
\end{aligned}$$

Here we have used that $\|m\|_\infty \leq \kappa_n = c_3 \cdot \log n$ and that all moments of Y are bounded due to $\mathbf{E}\{\exp(c_4 \cdot Y^2)\} < \infty$. Inequality (48) implies

$$\mathbf{E}\{T_{4,n}\} \leq c_{94}\cdot\frac{1}{\tilde{K}_n^{2\cdot\frac{p}{d}}}+c_{95}\cdot\frac{(\log n)^2}{n}\leq c_{96}\cdot n^{-\frac{2p}{2p+d^*}}.$$

Combing all steps, we get for all $\epsilon > 0$

$$\mathbf{E}\int|m_n(x)-m(x)|^2\mathbf{P}_X(dx)\leq c_{97}\cdot n^{-\frac{2p}{2p+d^*}+\epsilon}.$$

□

References

[1] Alquier, P. and Kegne, W. (2025) Minimax optimality of deep neural networks on dependent data via PAC-Bayes bounds. *Electronic Journal of Statistics*, **19**, pp. 5895-5924.

[2] Andoni, A, Panigrahy, R., Valiant, G. and Zhang, L. (2014). Learning polynomials with neural networks. *Proceedings of the 31st International Conference on Machine Learning (PMLR 2014)*, **32**, Beijing, China.

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

[4] Barrera, D. and Gobet, E. (2021). Generalization bounds for nonparametric regression with β-mixing samples. Preprint, *arXiv: 2108.00997v1.*

[5] Barron, A. R. (1993). Universal approximation bounds for superpositions of a sigmoidal function. *IEEE Transactions on Information Theory* **39**, pp. 930–944.

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

[7] Bartlett, P. and Mendelson, S. (2002). Rademacher and gaussian complexities: risk bounds and structural results. *Journal of Machine Learning Research*, **3**, pp. 463-482.

[8] Bartlett, P., Foster, D. J. and Telgarsky, M. (2017). Spectrally-normalized margin bounds for neural networks. *Proceedings of the 31st Conference on Neural Information Processing Systems (NIPS 2017)*, Long Beach, CA, USA.

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

[10] Bartlett, P. L., Montanari, A., and Rakhlin, A. (2021). Deep learning: a statistical viewpoint. Preprint, *arXiv: 2103.09177v1.*

[11] 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.

[12] Berbee, H. C. P. (1979). Random walks with stationary increments and renewal theory. *Mathematical Centre tracts* **112**.

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

[14] Blum, J. R., Hanson, D. L. and Koopmans, L. H. (1963). On the strong law of large numbers for a class of stochastic processes. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, **2**, pp. 1-11.

[15] Braun, A., Kohler, M., Langer, S., and Walk, H. (2024). Convergence rates for shallow neural networks learned by gradient descent. *Bernoulli*, **30**, pp. 475-502.

[16] Cao, Y. and Gu, Q. (2019). Generalization Bounds of stochastic gradient descent for wide and deep neural networks. 33rd *Conference on Neural Information Processing Systems* (NeurIPS 2019), Vancouver, Canada. *arXiv:1905.13210.*

[17] Chen, Z., Cao, Y., Zou, D. and, Gu, Q. (2021). How Much Over-parameterization is sufficient to learn deep relu networks? *International Conference on Learning Representations* (ICLR 2021), *arXiv:1911.12360.*

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

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

[20] Doob, J. L. (1953). *Stochastic Processes*, Wiley, New-York.

[21] Doukhan, P. (1994). *Mixing. Properties and Examples*, Springer-Verlag.

[22] Drews, S. and Kohler, M. (2024). On the universal consistency of an over-parametrized deep neural network estimate learned by gradient descent. *Annals of the Institue of Statistical Mathematics* **76**, pp. 361-391.

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

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

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

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

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

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

[29] 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.

[30] Ibragimov, I. A. (1962). Some limit theorems for stationary processes. *Theory of Probability and Its Applications*, **7**, pp. 349-382.

[31] 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.

[32] Jacot, A., Gabriel, F. and Hongler, C. (2020). Neural tangent kernel: convergence and generalization in neural networks. Preprint, *arXiv: 1806.07572v4*.

[33] Jiao, Y., Shen, G., Lin, Y. and Huang, J. (2023). Deep nonparametric regression on approximate manifolds: Nonasymptotic error bounds with polynomial prefactors. *Annals of Statistics* **51**, pp. 691–716.

[34] Kawaguchi, K and Huang, J. (2019). Gradient descent finds global minima for generalizable deep neural networks of practical sizes. Preprint, *arXiv: 1908.02419v1*.

[35] Kengne, W. and Wade, M. (2025a) Robust deep learning from weakly dependent data. *Neural Networks* **185**, article 107227.

[36] Kengne, W. and Wade, M. (2025b). Deep learning from strongly mixing observations: Sparse-penalized regularization and minimax optimality. *Journal of Complexity* **185**, article 101978.

[37] Kengne, W. and Wade, M. (2025c) A general framework for deep learning. Preprint, *arXiv: 2512.23425v1.*

[38] Kim, Y. (2014). Convolutional neural networks for sentence classification. Preprint, *arXiv: 1408.5882.*

[39] Kohler, M. (2014). Optimal global rates of convergence for noiseless regression estimation problems with adaptively chosen design. *Journal of Multivariate Analysis*, **132**, pp. 197-208.

[40] Kohler, M. (2026). On the rate of convergence of deep neural network regression estimates learned by gradient descent. *IEEE Transactions on Information Theory*, **72**, pp. 1777-1797.

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

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

[43] Kohler, M. and Krzyżak, A. (2023). On the rate of convergence of an over-parametrized transformer classifier learned by gradient descent. Preprint, *arXiv: 2312.17007.*

[44] Kohler, M. and Krzyżak, A. (2025a). Analysis of the rate of convergence of an over-parametrized deep neural network estimate learned by gradient descent. *IEEE Transactions on Information Theory* **71**, pp. 6165-6182.

[45] Kohler, M. and Krzyżak, A. (2025b). Statistically guided deep learning. Preprint, *arXiv: 2504.08489v.1.*

[46] 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.

[47] Kohler, M., Langer, L. and Reif, U. (2023). Estimation of a regression function on a manifold by fully connected deep neural networks. *Journal of Statistical Planning and Inference* **222**, pp. 160-181.

[48] Krizhevsky, A., Sutskever, I. and Hinton, G. E. (2012). ImageNet classification with deep convolutional neural networks. In F. Pereira et al. (Eds.), *Advances In Neural Information Processing Systems* Red Hook, NY: Curran. **25**, pp. 1097-1105.

[49] Kurisu, D., Fukami, R. and Koike, Y. (2024). Adaptive deep learning for nonlinear time series models. Preprint, *arXiv: 2207.02546.*

[50] Langer, S. (2020). Approximating smooth functions by deep neural networks with sigmoid activation function. *Journal of Multivariate Analysis* **182**, pp. 104696.

[51] LeCun, Y., Bengio, Y. and Hinton, G. (2015). Deep learning. *Nature*, **521**, pp.436-444.

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

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

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

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

[56] Ma, M. and Safikhani, A. (2022). Theoretical analysis of deep neural networks for temporally dependent observations. In *Proceedings of the 36th Conference on Neural Information Processing Systems (NeurIPS 2022).*

[57] Mahowald, j., Bell, B. and Geyer, M. (2026). Efficient analysis of the distilled neural tangent kernel. Preprint, *arXiv: 2602:11320v2.*

[58] 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.

[59] Minaee, S., Mikolov, T., Nikzad.N., Chenaghlu, M., Socher, R. ,Amatriain, X. and Gao, J. (2025). Large language models: a survey. Preprint, *arXiv: 2402.06196v3.*

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

[61] Padilla, C., Zhang, Z., Luo, X., Padilla, O. and Wang, D. (2024). Dense ReLU Neural Networks for Temporal-spatial Model. Preprint, *arXiv: 2411.09961.*

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

[63] 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.

[64] 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.

[65] Rosenblatt, M. (1956). A central limit theorem and a strong mixing condition. *Proceedings of National Academy of Sciences U.S.A.*,**42**, pp. 43-47.

[66] Schmidt-Hieber, J. (2020). Nonparametric regression using deep neural networks with ReLU activation function (with discussion). *Annals of Statistics* **48**, pp. 1875–1897. Preprint, *arXiv:1708.06633v2.*

[67] Silver, D., Schrittwieser, J., Simonyan, K., Antonoglou, I., Huang, A., Guez, A., Huber, T., et al. (2017). Mastering the game of go without human knowledge. *Nature* **550**, pp. 354-359.

[68] Skorokhod, A. (1978). *Studies in the theory of random processes.* Translated from the Russian by Scripta Technica, Inc. Addison-Wesley, Massachusetts.

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

[70] 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.*

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

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

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

[74] Wilson, J., van der Heide, C., Hodgkinson, L. and Roosta, F. (2025). Uncertainty quantification with the empirical neural tangent kernel. Preprint, *arXiv:2502.02870v3.*

[75] Wolkonski, V. A. and Rozanov, Y. A. (1959). Some limit theorems for random functions, Part I. *Theory of Probability and Its Applications*, **4**, pp. 178-197.

[76] Wojtowytsch, S. (2020). On the convergence of gradient descent training for two-layer relu-networks in the mean field regime. Preprint, *arXiv:2005.13530v1.*

[77] Wu, Y., Schuster, M., Chen, Z., Le, Q., Norouzi, M., Macherey, W., Krikum, M., et al. (2016). Google's neural machine translation system: Bridging the gap between human and machine translation. Preprint, *arXiv: 1609.08144.*

[78] Yarotsky, D. (2017). Error bounds for approximations with deep ReLU networks. *Neural Networks*, **94**, pp. 103-114.

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

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

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

A. Proof of Lemma 9

For the proof of Lemma 9 we need the following lemma.

Lemma 12 *Let $N, d, K \in \mathbb{N}$, let $X : \Omega \to \mathbb{R}^N$, $Y : \Omega \to \mathbb{R}^d$, $Z : \Omega \to \mathbb{R}^K$ and $U : \Omega \to \mathbb{R}$ be Borel measurable random variables defined on a probability space $(\Omega, \mathcal{A}, \mathbf{P})$ such that (X, Y) and U are independent and U is uniformly distributed on $[0, 1]$.*

Then there exists a probability space $(\bar{\Omega}, \bar{\mathcal{A}}, \bar{\mathbf{P}})$, Borel measurable random variables $\bar{X} : \bar{\Omega} \to \mathbb{R}^N$, $\bar{Y}, \bar{Y}^* : \bar{\Omega} \to \mathbb{R}^d$, $\bar{Z} : \bar{\Omega} \to \mathbb{R}^K$, $\bar{U} : \bar{\Omega} \to \mathbb{R}$ and a Borel measurable function $f : \mathbb{R}^N \times \mathbb{R}^d \times \mathbb{R} \to \mathbb{R}^d$ such that

$$\mathbf{P}_{(X,Y,Z,U)} = \bar{\mathbf{P}}_{(\bar{X},\bar{Y},\bar{Z},\bar{U})}, \tag{50}$$

$$\mathbf{P}_Y = \bar{\mathbf{P}}_{\bar{Y}^*}, \tag{51}$$

$$\bar{Y}^* = f(\bar{X}, \bar{Y}, \bar{U}) \quad a.s. \tag{52}$$

$$\bar{X}, \bar{Y}^* \text{ independent} \tag{53}$$

and

$$\bar{\mathbf{P}}\{\bar{Y} \neq \bar{Y}^*\} \leq \mathbf{E}\left\{ \operatorname*{ess\,sup}_{A \in \mathcal{B}_d} |\mathbf{P}\{Y \in A|X\} - \mathbf{P}\{Y \in A\}| \right\}. \tag{54}$$

Remark 4. In the proof we will show

$$\bar{\mathbf{P}}\{\bar{Y} \neq \bar{Y}^*\} \leq \mathbf{E}\left\{ \sup_{A \in \mathcal{C}} |\mathbf{P}\{Y \in A|X\} - \mathbf{P}\{Y \in A\}| \right\}$$

for some countable subset $\mathcal{C}$ of $\mathcal{B}_d$.

Proof. The proof is based on the proof sketch of Theorem 1 in Doukhan (1994), Section 1.1.

Partition $[-2^n, 2^n]^d$ into $N_n - 1 = 4^{n \cdot d}$ many cubes $C_{1,n}, \ldots, C_{N_n-1,n}$ of side length

$$\frac{2 \cdot 2^n}{4^n} = 2 \cdot \left(\frac{1}{2}\right)^n,$$

set

$$C_{N_n,n} = \mathbb{R}^d \setminus [-2^n, 2^n]^d,$$

let $y_{C_{1,n}}, \ldots, y_{C_{N_n-1,n}}$ be the centers of $C_{1,n}, \ldots, C_{N_n-1,n}$, and set

$$y_{C_{N_n,n}} = (2^n + 1, \ldots, 2^n + 1)^T.$$

Set

$$Y_n(\omega) = y_{C_{i,n}} \quad \text{if } Y(\omega) \in C_{i,n}$$

$(i = 1, \ldots, N_n)$.

Then

$$Y_n \to^{\mathbf{P}} Y, \tag{55}$$

since for any $\epsilon > 0$ we have

$$\limsup_{n\to\infty} \mathbf{P}\{\|Y_n - Y\|_\infty > \epsilon\} \le \limsup_{n\to\infty} \mathbf{P}\{Y \in C_{N_n,n}\} = 0.$$

For $x \in \mathbb{R}^d$ we define

$$\lambda_i = \lambda_{i,n}(x) = \mathbf{P}\{Y \in C_{i,n} | X = x\} \quad \text{and} \quad \mu_i = \mu_{i,n} = \mathbf{P}\{Y \in C_{i,n}\}$$

for $i = 1, \ldots, N_n$, and reorder $C_{1,n}, \ldots, C_{N_n,n}$ (depending on x) such that

$$\lambda_1 \le \mu_1, \ldots, \lambda_k \le \mu_k, \lambda_{k+1} > \mu_{k+1}, \ldots, \lambda_{N_n} > \mu_{N_n}$$

holds for some $k = k(x) \in \{1, \ldots, N_n\}$. This means

$$\mathbf{P}\{Y \in C_{i,n} | X = x\} \le \mathbf{P}\{Y \in C_{i,n}\} \quad \text{for } i = 1, \ldots, k$$

and

$$\mathbf{P}\{Y \in C_{i,n} | X = x\} > \mathbf{P}\{Y \in C_{i,n}\} \quad \text{for } i = k+1, \ldots, N_n.$$

Given $X = x$ and $Y_n = y_{C_{i,n}}$ (the latter is equivalent to $Y \in C_{i,n}$), we choose the value of Y_n^* as follows: In case $\mathbf{P}\{Y \in C_{i,n}\} \ge \mathbf{P}\{Y \in C_{i,n} | X = x\}$ we set $Y_n^* = y_{C_{i,n}}$. And in case $\mathbf{P}\{Y \in C_{i,n}\} < \mathbf{P}\{Y \in C_{i,n} | X = x\}$ we choose the value of Y_n^* randomly as follows: With probability

$$\frac{(\mu_j - \lambda_j) \cdot (\lambda_i - \mu_i)}{a \cdot \lambda_i}$$

we set it equal to $y_{C_{j,n}}$ for $j = 1, \ldots, k$, and with probability

$$\frac{\mu_i}{\lambda_i}$$

we set it equal to $y_{C_{i,n}}$, where

$$a = a_{n,x} = \sum_{l=1}^{k} (\mu_l - \lambda_l) = \sum_{l=k+1}^{N_n} (\lambda_l - \mu_l).$$

(Here the last equality holds since $\sum_{l=1}^{N_n} \mu_l = 1 = \sum_{l=1}^{N_n} \lambda_l$.) This is possible since

$$\sum_{j=1}^{k} \frac{(\mu_j - \lambda_j) \cdot (\lambda_i - \mu_i)}{a \cdot \lambda_i} + \frac{\mu_i}{\lambda_i} = 1.$$

It is easy to see that there exists a Borel measurable function $f_n : \mathbb{R}^N \times \mathbb{R}^d \times \mathbb{R} \to \mathbb{R}$ such that

$$Y_n^* = f_n(X, Y, U) \tag{56}$$

holds, and that Y_n and Y_n^* satisfy

$$\begin{aligned}
&\mathbf{P}\{Y_n = y_{C_{i,n}}, Y_n^* = y_{C_{j,n}} | X = x\} \\
&= \begin{cases}
\mathbf{P}\{Y \in C_{i,n} | X = x\}, & i = j \text{ and } \mathbf{P}\{Y \in C_{i,n}\} \geq \mathbf{P}\{Y \in C_{i,n} | X = x\} \\
0 & i \neq j \text{ and } \mathbf{P}\{Y \in C_{i,n}\} \geq \mathbf{P}\{Y \in C_{i,n} | X = x\} \\
\frac{(\mu_j - \lambda_j) \cdot (\lambda_i - \mu_i)}{a}, & j \leq k \text{ and } \mathbf{P}\{Y \in C_{i,n}\} < \mathbf{P}\{Y \in C_{i,n} | X = x\} \\
\mathbf{P}\{Y \in C_{i,n}\}, & i = j \text{ and } \mathbf{P}\{Y \in C_{i,n}\} < \mathbf{P}\{Y \in C_{i,n} | X = x\} \\
0, & \text{else.}
\end{cases}
\end{aligned}$$

Here the third and the fourth rows in the right-hand side above do not contradict each other since $\mathbf{P}\{Y \in C_{i,n}\} < \mathbf{P}\{Y \in C_{i,n} | X = x\}$ implies $i > k$.

Next we show

$$\mathbf{P}\{Y_n^* = y_{C_{j,n}} | X = x\} = \mathbf{P}\{Y_n = y_{C_{j,n}}\} \quad \text{for all } x, \text{ all } j = 1, \ldots, N_n. \tag{57}$$

Fix $x \in \mathbb{R}^N$. If

$$\mathbf{P}\{Y \in C_{j,n}\} < \mathbf{P}\{Y \in C_{j,n} | X = x\},$$

then $j \geq k + 1$ and

$$\begin{aligned}
&\mathbf{P}\{Y_n^* = y_{C_{j,n}} | X = x\} = \sum_{i=1}^{N_n} \mathbf{P}\{Y_n = y_{C_{i,n}}, Y_n^* = y_{C_{j,n}} | X = x\} \\
&= \sum_{i=1}^{k} \mathbf{P}\{Y_n = y_{C_{i,n}}, Y_n^* = y_{C_{j,n}} | X = x\} + \mathbf{P}\{Y_n = y_{C_{j,n}}, Y_n^* = y_{C_{j,n}} | X = x\} \\
&\quad + \sum_{i=k+1,\ldots,N_n, i \neq j} \mathbf{P}\{Y_n = y_{C_{i,n}}, Y_n^* = y_{C_{j,n}} | X = x\} \\
&= 0 + \mathbf{P}\{Y \in C_{j,n}\} + 0 = \mathbf{P}\{Y_n = y_{C_{j,n}}\}.
\end{aligned}$$

If

$$\mathbf{P}\{Y \in C_{j,n}\} \geq \mathbf{P}\{Y \in C_{j,n} | X = x\},$$

then $j \leq k$ and

$$\begin{aligned}
&\mathbf{P}\{Y_n^* = y_{C_{j,n}} | X = x\} = \sum_{i=1}^{N_n} \mathbf{P}\{Y_n = y_{C_{i,n}}, Y_n^* = y_{C_{j,n}} | X = x\} \\
&= \mathbf{P}\{Y_n = y_{C_{j,n}}, Y_n^* = y_{C_{j,n}} | X = x\} + \sum_{i=1,\ldots,k, i \neq j} \mathbf{P}\{Y_n = y_{C_{i,n}}, Y_n^* = y_{C_{j,n}} | X = x\} \\
&\quad + \sum_{i=k+1}^{N_n} \mathbf{P}\{Y_n = y_{C_{i,n}}, Y_n^* = y_{C_{j,n}} | X = x\} \\
&= \mathbf{P}\{Y_n \in C_{j,n} | X = x\} + 0 + \sum_{i=k+1}^{N_n} \frac{(\mu_j - \lambda_j) \cdot (\lambda_i - \mu_i)}{a}
\end{aligned}$$

$$= \lambda_j + (\mu_j - \lambda_j) \cdot \frac{1}{a} \cdot \sum_{i=k+1}^{N_n} (\lambda_i - \mu_i)$$
$$= \lambda_j + (\mu_j - \lambda_j) \cdot \frac{1}{a} \cdot a = \mu_j = \mathbf{P}\{Y_n = y_{C_{j,n}}\}.$$

This proves (57).

From (57) we conclude

$$\begin{aligned} \mathbf{P}\{Y_n^* = y_{C_{j,n}}\} &= \int \mathbf{P}\{Y_n^* = y_{C_{j,n}} | X = x\} \mathbf{P}_X(dx) = \int \mathbf{P}\{Y_n = y_{C_{j,n}}\} \mathbf{P}_X(dx) \\ &= \mathbf{P}\{Y_n = y_{C_{j,n}}\} \end{aligned}$$

for $j = 1, \ldots, N_n$, hence

$$\mathbf{P}_{Y_n^*} = \mathbf{P}_{Y_n}. \tag{58}$$

Furthermore, for $j \in \{1, \ldots, N_n\}$ and $A \in \mathcal{B}_N$ we can conclude from (57) and (58)

$$\begin{aligned} \mathbf{P}\{Y_n^* = y_{C_{j,n}}, X \in A\} &= \int_A \mathbf{P}\{Y_n^* = y_{C_{j,n}} | X = x\} \mathbf{P}_X(dx) \\ &= \int_A \mathbf{P}\{Y_n = y_{C_{j,n}}\} \mathbf{P}_X(dx) \\ &= \mathbf{P}\{Y_n = y_{C_{j,n}}\} \cdot \mathbf{P}\{X \in A\} \\ &= \mathbf{P}\{Y_n^* = y_{C_{j,n}}\} \cdot \mathbf{P}\{X \in A\}, \end{aligned}$$

hence

$$Y_n^* \text{ and } X \text{ are independent.} \tag{59}$$

Set

$$\mathcal{C}_n = \{\cup_{j \in J} C_{j,n} \,:\, J \subseteq \{1, \ldots, N_n\}\} \quad \text{and} \quad \mathcal{C} = \cup_{n \in \mathbb{N}} \mathcal{C}_n.$$

By the definition of Y_n^* we get

$$\begin{aligned} &\mathbf{P}\{Y_n \neq Y_n^* | X = x\} \\ &= \sum_{1 \leq i,j \leq N_n, i \neq j} \mathbf{P}\{Y_n = y_{C_{i,n}}, Y_n^* = y_{C_{j,n}} | X = x\} \\ &= \sum_{j=1}^{k} \sum_{i=k+1}^{N_n} \frac{(\mu_j - \lambda_j) \cdot (\lambda_i - \mu_i)}{a} \\ &= \frac{1}{a} \cdot a \cdot a = a \\ &= \sum_{l=1}^{k} (\mathbf{P}\{Y \in C_{l,n}\} - \mathbf{P}\{Y \in C_{l,n} | X = x\}) \\ &= \mathbf{P}\{Y \in \cup_{l=1}^{k} C_{l,n}\} - \mathbf{P}\{Y \in \cup_{l=1}^{k} C_{l,n} | X = x\} \\ &\leq \sup_{C \in \mathcal{C}_n} |\mathbf{P}\{Y \in C\} - \mathbf{P}\{Y \in C | X = x\}| \end{aligned}$$

where we have used

$$\cup_{l=1}^{k} C_{l,n} \in \mathcal{C}_n.$$

(Here the sets in the union above depend on x and not only on l and n). Hence

$$\begin{aligned}
\mathbf{P}\{Y_n \neq Y_n^*\} &= \int \mathbf{P}\{Y_n \neq Y_n^* | X = x\} \mathbf{P}_X(dx) \\
&\leq \int \sup_{C \in \mathcal{C}_n} |\mathbf{P}\{Y \in C\} - \mathbf{P}\{Y \in C | X = x\}| \, \mathbf{P}_X(dx) \\
&= \mathbf{E}\left\{ \sup_{C \in \mathcal{C}_n} |\mathbf{P}\{Y \in C\} - \mathbf{P}\{Y \in C | X\}| \right\} \\
&\leq \mathbf{E}\left\{ \sup_{C \in \mathcal{C}} |\mathbf{P}\{Y \in C\} - \mathbf{P}\{Y \in C | X\}| \right\}. \qquad (60)
\end{aligned}$$

For $M \in \mathbb{N}$ we have

$$\begin{aligned}
&\sup_{n \in \mathbb{N}} \mathbf{P}\left\{ (X, Y, Y_n, Y_n^*, Z, U) \notin [-M, M]^{N+3d+K+1} \right\} \\
&\leq \mathbf{P}\{X \notin [-M, M]^N\} + \mathbf{P}\{Y \notin [-M, M]^d\} + \sup_{n \in \mathbb{N}} \mathbf{P}\{Y_n \notin [-M, M]^d\} \\
&\quad + \sup_{n \in \mathbb{N}} \mathbf{P}\{Y_n^* \notin [-M, M]^d\} + \mathbf{P}\{Z \notin [-M, M]^K\} + \mathbf{P}\{U \notin [-M, M]\} \\
&\to 0 \quad (M \to \infty),
\end{aligned}$$

since (58) and the definition of Y_n imply

$$\begin{aligned}
&\sup_{n \in \mathbb{N}} \mathbf{P}\{Y_n^* \notin [-M, M]^d\} = \sup_{n \in \mathbb{N}} \mathbf{P}\{Y_n \notin [-M, M]^d\} \\
&\leq \sup_{n \in \mathbb{N}: 2^n + 1 \leq M} \mathbf{P}\{Y_n \notin [-M, M]^d\} + \sup_{n \in \mathbb{N}: M < 2^n + 1} \mathbf{P}\{Y_n \notin [-M, M]^d\} \\
&= 0 + \mathbf{P}\{Y \notin [-M, M]^d\} \to 0 \quad (M \to \infty).
\end{aligned}$$

From this we get that

$$\left(\mathbf{P}_{(X, Y, Y_n, Y_n^*, Z, U)}\right)_{n \in \mathbb{N}}$$

is tight, hence the theorem of Prokhorov implies that there exists a subsequence $(n_k)_k$ of $(n)_n$ and a probability measure

$$\mathbf{P}_{(\tilde{X}, \tilde{Y}, \tilde{\tilde{Y}}, \tilde{Y}^*, \tilde{Z}, \tilde{U})}$$

such that

$$\mathbf{P}_{(X, Y, Y_{n_k}, Y_{n_k}^*, Z, U)} \to \mathbf{P}_{(\tilde{X}, \tilde{Y}, \tilde{\tilde{Y}}, \tilde{Y}^*, \tilde{Z}, \tilde{U})} \quad \text{weakly.} \qquad (61)$$

By the Portmanteau theorem we can conclude from this

$$\mathbf{P}\{\tilde{Y} \neq \tilde{\tilde{Y}}\}\} = \limsup_{\epsilon \to 0} \mathbf{P}\{|\tilde{Y} - \tilde{\tilde{Y}}| > \epsilon\} \leq \limsup_{\epsilon \to 0} \liminf_{k \to \infty} \mathbf{P}\{|Y - Y_{n_k}| > \epsilon\} = 0,$$

where the last equality follows from (56). Hence we can conclude from (61)

$$\mathbf{P}_{(X,Y,Y_{n_k},Y^*_{n_k},Z,U)} \to \mathbf{P}_{(\tilde{X},\tilde{Y},\tilde{Y},\tilde{Y}^*,\tilde{Z},\tilde{U})} \quad \text{weakly}, \tag{62}$$

By the continuous mapping theorem this implies

$$\mathbf{P}_{(X,Y,Z,U)} \to \mathbf{P}_{(\tilde{X},\tilde{Y},\tilde{Z},\tilde{U})} \quad \text{weakly}.$$

and because of the uniqueness of the limit distribution for weak convergence we get

$$\mathbf{P}_{(X,Y,Z,U)} = \mathbf{P}_{(\tilde{X},\tilde{Y},\tilde{Z},\tilde{U})}. \tag{63}$$

By the proof of Skorokhod's representation theorem in Skorokhod (1978) (cf., Lemma 13 below) we can conclude from (62) that there exists a probability space $(\bar{\Omega},\bar{\mathcal{A}},\bar{\mathbf{P}})$ and random variables $(\bar{X},\bar{Y},\bar{Y}_{n_k},\bar{Y}^*_{n_k},\bar{Z},\bar{U})$ and $(\bar{X},\bar{Y},\bar{Y},\bar{\bar{Y}}^*,\bar{Z},\bar{U})$ such that

$$\mathbf{P}_{(X,Y,Y_{n_k},Y^*_{n_k},Z,U)} = \bar{\mathbf{P}}_{(\bar{X},\bar{Y},\bar{Y}_{n_k},\bar{Y}^*_{n_k},\bar{Z},\bar{U})}, \tag{64}$$

$$\mathbf{P}_{(\tilde{X},\tilde{Y},\tilde{Y},\tilde{Y}^*,\tilde{Z},\tilde{U})} = \bar{\mathbf{P}}_{(\bar{X},\bar{Y},\bar{\bar{Y}},\bar{Y}^*,\bar{Z},\bar{U})} \tag{65}$$

and

$$(\bar{X},\bar{Y},\bar{Y}_{n_k},\bar{Y}^*_{n_k},\bar{Z},U) \to (\bar{X},\bar{Y},\bar{\bar{Y}},\bar{Y}^*,\bar{Z},\bar{U}) \quad a.s. \tag{66}$$

From (65) we get

$$\bar{\mathbf{P}}\{\bar{Y} \neq \bar{\bar{Y}}\} = \mathbf{P}\{\tilde{Y} \neq \tilde{Y}\} = 0,$$

hence (65) and (66) imply

$$\mathbf{P}_{(\tilde{X},\tilde{Y},\tilde{Y},\tilde{Y}^*,\tilde{Z},\tilde{U})} = \bar{\mathbf{P}}_{(\bar{X},\bar{Y},\bar{Y},\bar{Y}^*,\bar{Z},\bar{U})} \tag{67}$$

and

$$(\bar{X},\bar{Y},\bar{Y}_{n_k},\bar{Y}^*_{n_k},\bar{Z},U) \to (\bar{X},\bar{Y},\bar{Y},\bar{Y}^*,\bar{Z},\bar{U}) \quad a.s. \tag{68}$$

We show next that $\bar{X},\bar{Y},\bar{Y}^*,\bar{Z},\bar{U}$ satisfy (50)-(54).

Identity (50) follows from (67) and (63), because these two equations imply

$$\bar{\mathbf{P}}_{(\bar{X},\bar{Y},\bar{Z},\bar{U})} = \mathbf{P}_{(\tilde{X},\tilde{Y},\tilde{Z},\tilde{U})} = \mathbf{P}_{(X,Y,Z,U)}.$$

Identity (51) follows from

$$\bar{\mathbf{P}}_{\bar{Y}_{n_k}} \to \bar{\mathbf{P}}_{\bar{Y}} \quad \text{weakly} \tag{69}$$

(cf., (68)),

$$\bar{\mathbf{P}}_{\bar{Y}^*_{n_k}} \to \bar{\mathbf{P}}_{\bar{Y}^*} \quad \text{weakly} \tag{70}$$

(cf., (68)) and

$$\bar{\mathbf{P}}_{\bar{Y}^*_{n_k}} = \mathbf{P}_{Y^*_{n_k}} = \mathbf{P}_{Y_{n_k}} = \bar{\mathbf{P}}_{\bar{Y}_{n_k}} \tag{71}$$

(cf., (64) and (58)), which imply

$$\bar{\mathbf{P}}_{\bar{Y}_{n_k}} \to \bar{\mathbf{P}}_{\bar{Y}^*} \quad \text{weakly}. \tag{72}$$

Using the uniqueness of the limit distribution we get (51) from (72), (69) and (50).

In order to show (52) we observe that

$$\bar{\mathbf{P}}\{\bar{Y}^*_{n_k} = f_{n_k}(\bar{X}, \bar{Y}, \bar{U})\} = \mathbf{P}\{Y^*_{n_k} = f_{n_k}(X, Y, U)\} = 1$$

(cf., (64) and (56)) and (68) imply that we have with probability one

$$\bar{Y}^* = \lim_{k\to\infty} \bar{Y}^*_{n_k} = \lim_{k\to\infty} f_{n_k}(\bar{X}, \bar{Y}, \bar{U}).$$

Set

$$f(x, y, u) = \limsup_{k\to\infty} f_{n_k}(x, y, u) \cdot 1_{\{\limsup_{k\to\infty} f_{n_k}(x,y,u)<\infty\}}.$$

Then $\bar{Y}^* < \infty$ *a.s.* implies

$$\bar{Y}^* = \lim_{k\to\infty} f_{n_k}(\bar{X}, \bar{Y}, \bar{U}) = f(\bar{X}, \bar{Y}, \bar{U}),$$

which proves (52).

Next we prove (53). Let $\varphi_{(\bar{X},\bar{Y}^*)}$, $\varphi_{\bar{X}}$ and $\varphi_{\bar{Y}^*}$ be the characteristic functions of $(\bar{X}, \bar{Y}^*)$, $\bar{X}$ and $\bar{Y}^*$, resp. Then the continuity theorem of Levy-Cramer together with (67), (62), (59) and (64) imply

$$\begin{aligned}
&\varphi_{(\bar{X},\bar{Y}^*)}(u, v) = \varphi_{(\tilde{X},\tilde{Y}^*)}(u, v) = \lim_{k\to\infty} \varphi_{(X,Y^*_{n_k})}(u, v) = \lim_{k\to\infty} \varphi_X(u) \cdot \varphi_{Y_{n_k}}(v) \\
&= \varphi_X(u) \cdot \varphi_{\tilde{Y}^*}(v) = \varphi_X(u) \cdot \varphi_{\bar{Y}^*}(v),
\end{aligned}$$

hence (53) holds.

We finish the proof by showing (54). To do this we conclude from

$$\bar{\mathbf{P}}_{(\bar{Y}_{n_k},\bar{Y}^*_{n_k})} \to \bar{\mathbf{P}}_{(\bar{Y},\bar{Y}^*)} \quad \text{weakly}$$

(cf., (68)), the Portmanteau theorem, (64) and (60)

$$\begin{aligned}
\bar{\mathbf{P}}\{\bar{Y} \neq \bar{Y}^*\} &\leq \liminf_{k\to\infty} \bar{\mathbf{P}}\{\bar{Y}_{n_k} \neq \bar{Y}^*_{n_k}\} \\
&= \liminf_{k\to\infty} \mathbf{P}\{Y_{n_k} \neq Y^*_{n_k}\} \\
&\leq \mathbf{E}\left\{\sup_{C\in\mathcal{C}} |\mathbf{P}\{Y \in C\} - \mathbf{P}\{Y \in C|X\}|\right\},
\end{aligned}$$

hence (54) holds. □

In the proof above we have used the following modification of the classical representation theorem of Skorokhod.

Lemma 13 *Let $M, N \in \mathbb{N}$, let $(\Omega, \mathcal{A}, \mathbf{P})$ be a probability space, and let $X : \bar{\Omega} \to \mathbb{R}^M$, $Y, Y_n : \bar{\Omega} \to \mathbb{R}^N$ be Borel measurable random variables with*

$$\mathbf{P}_{(X,Y_n)} \to \mathbf{P}_{(X,Y)} \quad \textit{weakly.}$$

Then there exists a probability space $(\bar{\Omega}, \bar{\mathcal{A}}, \bar{\mathbf{P}})$ *and Borel measurable random variables* $\bar{X} : \Omega \to \mathbb{R}^M$, $\bar{Y}, \bar{Y}_n : \Omega \to \mathbb{R}^N$ *such that*

$$\mathbf{P}_{(X,Y_n)} = \bar{\mathbf{P}}_{(\bar{X},\bar{Y}_n)} \quad (n \in \mathbb{N}), \tag{73}$$

$$\mathbf{P}_{(X,Y)} = \bar{\mathbf{P}}_{(\bar{X},\bar{Y})} \tag{74}$$

and

$$(\bar{X}, \bar{Y}_n) \to (\bar{X}, \bar{Y}) \quad a.s. \tag{75}$$

Proof. The assertion follows from Skorokhod (1978), pp. 10-11, if we use the construction there to define $S_{i_1,\ldots,i_k}$ and $\tilde{S}_{j_1,\ldots,j_k}$ for $\mathbb{R}^M$ and $\mathbb{R}^N$, resp., use then as there

$$S_{i_1,\ldots,i_k} \times \tilde{S}_{j_1,\ldots,j_k}$$

to construct $(\bar{X}_n^{(k)}, \bar{Y}_n^{(k)})$ from (X, Y), and use that on

$$\Delta^{(n)}_{i_1,\ldots,i_k,j_1,\ldots,j_k}$$

$\bar{X}_n^{(k)}$ has the same value for all $j_1, \ldots, j_k$. Let λ denote the Lebesgue measure. Since we have that

$$\begin{aligned} \sum_{j_1,\ldots,j_k} \lambda(\Delta^{(n)}_{i_1,\ldots,i_k,j_1,\ldots,j_k}) &= \sum_{j_1,\ldots,j_k} \mathbf{P}\{X \in S_{i_1,\ldots,i_k}, Y_n \in \tilde{S}_{j_1,\ldots,j_k}\} \\ &= \mathbf{P}\{X \in S_{i_1,\ldots,i_k}\} \end{aligned}$$

does not depend on n and hence

$$\bar{X}_n(\omega) = \lim_{k\to\infty} \bar{X}_n^{(k)}(\omega)$$

does not depend on n for all ω, the construction in the proof of Skorokhod (1978), pp. 10-11, gives us that the random variables $(\bar{X}_n, \bar{Y}_n)$ constructed there with the property

$$(\bar{X}_n, \bar{Y}_n) \to (\bar{X}, \bar{Y}) \quad a.s.$$

satisfy

$$\bar{X}_n = \bar{X}$$

for all $n \in \mathbb{N}$. □

Now we are ready to prove Lemma 9.

Proof of Lemma 9. It suffices to show that for every $l \in \{1, \ldots, n\}$ there exists a probability space $(\bar{\Omega}, \bar{\mathcal{A}}, \bar{\mathbf{P}})$ and random variables $\bar{Z}_0$, $\bar{Z}_1$, …, $\bar{Z}_n$, $\bar{Z}_1^*$, …, $\bar{Z}_l^*$, $\bar{U}_1$, …, $\bar{U}_l$ defined on this probability space such that

$$\mathbf{P}_{(Z_0,Z_1,\ldots,Z_n)} = \bar{\mathbf{P}}_{(\bar{Z}_0,\bar{Z}_1,\ldots,\bar{Z}_n)}, \tag{76}$$

$$\bar{Z}_0, \bar{Z}_1^*, \ldots, \bar{Z}_l^* \quad \text{are } i.i.d., \tag{77}$$

$$\bar{Z}_k^* = f_k(\bar{U}_1, \ldots, \bar{U}_{k-1}, \bar{Z}_0, \bar{Z}_1, \ldots, \bar{Z}_{k-1}, \bar{Z}_1^*, \ldots, \bar{Z}_{k-1}^*, \bar{Z}_k, \bar{U}_k) \tag{78}$$

for some Borel measurable function f_k and all $k \in \{1, \ldots, l\}$,

$$(\bar{Z}_0, \bar{Z}_1, \ldots, \bar{Z}_n) \text{ and } (\bar{U}_1, \ldots, \bar{U}_l) \text{ are independent,} \tag{79}$$

and

$$\begin{aligned} &\bar{\mathbf{P}}\{\exists k \in \{1, \ldots, l\} : \bar{Z}_k \neq \bar{Z}_k^*\} \\ &\leq (l-1) \cdot \max_{s \in \{2,\ldots,l\}} \mathbf{E}\left\{ \sup_{C \in \mathcal{C}} |\mathbf{P}\{Z_s \in C | Z_1, \ldots, Z_{s-1}\} - \mathbf{P}\{Z_s \in C\}| \right\}. \end{aligned} \tag{80}$$

We show this by induction on l. The assertion trivially holds for $l = 1$ if we choose

$$(\bar{\Omega}, \bar{\mathcal{A}}, \bar{\mathbf{P}}) = (\Omega, \mathcal{A}, \mathbf{P}), \quad \bar{Z}_i = Z_i \quad (i = 0, \ldots, n) \quad \text{and} \quad \bar{Z}_1^* = Z_1.$$

So assume next that the assertion holds for some $l \in \{1, \ldots, n-1\}$. We extend the probability space $(\bar{\Omega}, \bar{\mathcal{A}}, \bar{\mathbf{P}})$ such that there exists a random variable U_{l+1} uniformly distributed on $[0, 1]$, which is independent from $\bar{Z}_0$, $\bar{Z}_1$, …, $\bar{Z}_n$, $\bar{Z}_1^*$, …, $\bar{Z}_l^*$, $\bar{U}_1$, …, $\bar{U}_l$. To simplify the notation we write in the sequel $\mathbf{P}$ and Z_0, Z_1, …, Z_n, Z_1^*, …, Z_l^*, U_1, …, U_l instead of $\bar{\mathbf{P}}$ and $\bar{Z}_0$, $\bar{Z}_1$, …, $\bar{Z}_n$, $\bar{Z}_1^*$, …, $\bar{Z}_l^*$, $\bar{U}_1$, …, $\bar{U}_l$. We will apply Lemma 12 in order to show the assertion for $l+1$. To do this, we set

$$X = (U_1, \ldots, U_l, Z_0, Z_1, \ldots, Z_l, Z_1^*, \ldots, Z_l^*),$$

$$Y = Z_{l+1},$$

$$Z = (Z_{l+2}, \ldots, Z_n)$$

and

$$U = U_{l+1}$$

Application of Lemma 12 yields

$$\bar{X} = (\bar{U}_1, \ldots, \bar{U}_l, \bar{Z}_0, \bar{Z}_1, \ldots, \bar{Z}_l, \bar{Z}_1^*, \ldots, \bar{Z}_l^*),$$

$$\bar{Y} = \bar{Z}_{l+1},$$

$$\bar{Y}^* = \bar{Z}_{l+1}^*,$$

$$\bar{Z} = (\bar{Z}_{l+2}, \ldots, \bar{Z}_n)$$

and

$$\bar{U} = \bar{U}_{l+1}$$

such that (50)–(53) hold. We show next that $\bar{Z}_0$, $\bar{Z}_1$, …, $\bar{Z}_n$, $\bar{Z}_1^*$, …, $\bar{Z}_{l+1}^*$, $\bar{U}_1$, …, $\bar{U}_{l+1}$ satisfy (76)–(80) for $l+1$ instead of l.

(76) directly follows from (50) and our choice of X, Y and Z.

Because of the induction hypothesis and (76), (77) holds for l, and using (51) and (53) we can conclude from this that (77) also holds for $l+1$.

For $k \leq l$ the induction hypothesis and (50) imply (78). For $k = l+1$ identity (78) is a direct consequence of (52).

(79) follows from the induction hypothesis (which together with (50) implies that (79) holds for l) and U_{l+1} independent from $Z_0, \ldots, Z_n, U_1, \ldots, U_l$ (which together with (50) implies $\bar{U}_{l+1}$ independent from $\bar{Z}_0, \ldots, \bar{Z}_n, \bar{U}_1, \ldots, \bar{U}_l$).

So it remains to show (80) for $l+1$. Because of the induction hypothesis, (76) and the union bound it suffices to show

$$\bar{\mathbf{P}}\{\bar{Z}_{l+1} \neq \bar{Z}^*_{l+1}\} \leq \mathbf{E}\left\{\sup_{C\in\mathcal{C}} |\mathbf{P}\{Z_{l+1} \in C | Z_1, \ldots, Z_l\} - \mathbf{P}\{Z_{l+1} \in C\}|\right\}.$$

(54) implies

$$\begin{aligned}
&\bar{\mathbf{P}}\{\bar{Z}_{l+1} \neq \bar{Z}^*_{l+1}\} = \bar{\mathbf{P}}\{\bar{Y} \neq \bar{Y}^*\} \leq \mathbf{E}\left\{\sup_{C\in\mathcal{C}} |\mathbf{P}\{Y \in C|X\} - \mathbf{P}\{Y \in C\}|\right\} \\
&= \mathbf{E}\left\{\sup_{C\in\mathcal{C}} |\mathbf{P}\{Z_{l+1} \in C | U_1, \ldots, U_l, Z_0, Z_1, \ldots, Z_l, Z^*_1, \ldots, Z^*_l\} - \mathbf{P}\{Z_{l+1} \in C\}|\right\}
\end{aligned}$$

By (78) we get that the σ field generated by $U_1, \ldots, U_l, Z_0, Z_1, \ldots, Z_l, Z^*_1, \ldots, Z^*_l$ is the same as the σ-field generated by $U_1, \ldots, U_l, Z_0, Z_1, \ldots, Z_l$, hence

$$\begin{aligned}
&\mathbf{P}\{Z_{l+1} \in C | U_1, \ldots, U_l, Z_0, Z_1, \ldots, Z_l, Z^*_1, \ldots, Z^*_l\} \\
&= \mathbf{P}\{Z_{l+1} \in C | U_1, \ldots, U_l, Z_0, Z_1, \ldots, Z_l\}.
\end{aligned}$$

Because of $(U_1, \ldots, U_l)$ independent from $(Z_{l+1}, Z_0, Z_1, \ldots, Z_l)$ the last conditional probability is equal to

$$\mathbf{P}\{Z_{l+1} \in C | Z_0, Z_1, \ldots, Z_l\},$$

from which we can conclude

$$\begin{aligned}
&\bar{\mathbf{P}}\{\bar{Z}_{l+1} \neq \bar{Z}^*_{l+1}\} \\
&\leq \mathbf{E}\left\{\sup_{C\in\mathcal{C}} |\mathbf{P}\{Z_{l+1} \in C | U_1, \ldots, U_l, Z_0, Z_1, \ldots, Z_l, Z^*_1, \ldots, Z^*_l\} - \mathbf{P}\{Z_{l+1} \in C\}|\right\} \\
&= \mathbf{E}\left\{\sup_{C\in\mathcal{C}} |\mathbf{P}\{Z_{l+1} \in C | Z_0, Z_1, \ldots, Z_l\} - \mathbf{P}\{Z_{l+1} \in C\}|\right\},
\end{aligned}$$

which proves (80). □