ETH
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Seminar for Applied Mathematics SAM

Analysis of the generalization error: Empirical risk minimization over deep artificial neural networks overcomes the curse of dimensionality in the numerical approximation of Black-Scholes partial differential equations

J. Berner and Ph. Grohs and A. Jentzen

Research Report No. 2018-31
September 2018

Seminar für Angewandte Mathematik
Eidgenössische Technische Hochschule
CH-8092 Zürich
Switzerland

Analysis of the generalization error: Empirical risk minimization over deep artificial neural networks overcomes the curse of dimensionality in the numerical approximation of Black-Scholes partial differential equations

Julius Berner[1], Philipp Grohs[2], and Arnulf Jentzen[3]

[1]Faculty of Mathematics, University of Vienna,
Austria, e-mail: julius.berner@univie.ac.at

[2]Faculty of Mathematics and Research Platform DataScience@UniVienna,
University of Vienna, Austria, e-mail: philipp.grohs@univie.ac.at

[3]Department of Mathematics, ETH Zürich,
Switzerland, e-mail: arnulf.jentzen@sam.math.ethz.ch

September 11, 2018

Abstract

The development of new classification and regression algorithms based on empirical risk minimization (ERM) over deep neural network hypothesis classes, coined Deep Learning, revolutionized the area of artificial intelligence, machine learning, and data analysis. More recently, these methods have been applied to the numerical solution of high dimensional partial differential equations (PDEs) with great success. In particular, recent simulations indicate that deep learning based algorithms are capable of overcoming the curse of dimensionality for the numerical solution of linear Kolmogorov PDEs. Kolmogorov PDEs have been widely used in models from engineering, finance, and the natural sciences. In particular Kolmogorov PDEs are highly employed in models for the approximative pricing of financial derivatives. Nearly all approximation methods for Kolmogorov PDEs in the literature suffer under the curse of dimensionality. By contrast, in recent work by some of the authors it was shown that deep ReLU neural networks are capable of approximating solutions of Kolmogorov PDEs without incurring the curse of dimensionality. The present paper considerably strengthens these results by providing an analysis of the generalization error. In particular we show that for Kolmogorov PDEs with affine drift and diffusion coefficients and a given accuracy $\varepsilon > 0$, ERM over deep neural network hypothesis classes of size scaling polynomially in the dimension d and ε^{-1} and with a number of training samples scaling polynomially in the dimension d and ε^{-1} approximates the solution of the Kolmogorov PDE to within accuracy ε with high probability. We conclude that ERM over deep neural network hypothesis classes breaks the curse of dimensionality for the numerical solution of linear Kolmogorov PDEs with affine drift and diffusion coefficients. To the best of our knowledge this is the first rigorous mathematical result that proves the efficiency of deep learning methods for high dimensional problems.

Contents

1 Introduction

1.1 Problem Statement

Suppose we need to numerically approximate the end value $[u,v]^d \ni x \mapsto F_d(T,x)$ of the solution $F_d \in C([0,T]\times\mathbb{R}^d,\mathbb{R})$ of a *linear Kolmogorov equation* which for an initial value $\varphi_d \in C(\mathbb{R}^d,\mathbb{R})$, diffusion coefficient $\sigma_d \in C(\mathbb{R}^d,\mathbb{R}^{d\times d})$ and drift $\mu_d \in C(\mathbb{R}^d,\mathbb{R}^d)$ is defined as

$$\begin{cases} \frac{\partial F_d}{\partial t}(t,x) = \frac{1}{2}\mathrm{Trace}\big(\sigma_d(x)[\sigma_d(x)]^*(\mathrm{Hess}_x F_d)(t,x)\big) + \big\langle \mu_d(x), (\nabla_x F_d)(t,x)\big\rangle_{\mathbb{R}^d} \\ F_d(0,x) = \varphi_d(x) \end{cases} \tag{1}$$

for every $(t,x) \in [0,T]\times\mathbb{R}^d$. Important special cases include the heat equation or the Black-Scholes equation from computational finance where typically the functions σ_d, μ_d are affine and the initial values φ_d can be represented as a composition of multivariate minima, maxima and linear combinations such as

$$\varphi_d(x) = \min\left\{\max\left\{\mathfrak{D} - \sum_{i=1}^{d} c_{d,i}x_i, 0\right\}, \mathfrak{D}\right\} \tag{2}$$

with suitable coefficients $\mathfrak{D}, c_{d,i} \in (0,\infty)$, $d \in \mathbb{N}$, $i \in \{1,2,\dots,d\}$, in the case of a European Put option pricing problem. It is well-known that most standard numerical algorithms for this problem suffer from the curse of dimensionality, meaning that their computational complexity grows exponentially in the dimension d [8].

If the goal is simply to evaluate $F_d(T,\cdot)$ *at a single value* $\xi \in \mathbb{R}^d$, then under suitable assumptions Monte-Carlo sampling methods are known to not suffer from the curse of dimensionality. These methods are based on the integral representation (Feynman-Kac formula)

$$F_d(T,\xi) = \mathbb{E}\big[\varphi_d(S_T^\xi)\big] \tag{3}$$

with $(S_t^\xi)_{t\in[0,T]}$ a stochastic process satisfying the stochastic differential equation

$$dS_t^\xi = \sigma_d(S_t^\xi)dB_t^d + \mu_d(S_t^\xi)dt \quad \text{and} \quad S_0^\xi = \xi$$

on some probability space $(\Omega, \mathcal{G}, \mathbb{P})$. The evaluation of $F_d(T,\xi)$ can then be computed by approximating the expectation in (3) by Monte-Carlo integration, that is, by simulating i.i.d. samples $(S^{(i)})_{i=1}^m$ with $S^{(1)} \sim S_T^\xi$ and by approximating $F_d(T,\xi)$ with the empirical average

$$\frac{1}{m}\sum_{i=1}^m \varphi_d(S^{(i)}).$$

It is well known that the number of samples m needed to obtain a desired accuracy ε depends only polynomially on the dimension d and ε^{-1}, implying that Monte-Carlo sampling does indeed not suffer from the curse of dimensionality.

If the goal is however to approximate $F_d(T,\cdot)$ not only at a single value but, for example, on a full hypercube $[u,v]^d$, there has been no known method not suffering from the curse of dimensionality. In particular, there has been no known method that can provably be applied efficiently in high dimensions, say, $d >> 100$.

The present paper introduces and analyzes deep learning based algorithms for the numerical approximation of $F_d(T,\cdot)$ on a full hypercube $[u,v]^d$. We will prove that the resulting algorithms overcome the curse of dimensionality and can consequently be efficiently applied even in high dimensions. Our proofs will be based on tools from statistical learning theory [13] and the following key properties of linear Kolmogorov equations:

- P.1 the fact that one can reformulate (1) as a mathematical learning problem (see Lemma 3.2 below),
- P.2 the fact that typical initial conditions arising from problems in computational finance, such as for example (2), are either exactly representable as neural networks with ReLU activation function or can be approximated by such neural networks without incurring the curse of dimensionality (see [23, Section 4])
- P.3 and the fact that Property P.2 is preserved under the evolution of linear Kolmogorov equations (1) with affine diffusion- and drift coefficients which implies that $F_d(T,\cdot)$ can be approximated by neural networks with ReLU activation function without incurring the curse of dimensionality (see Theorem 3.3 below).

1.2 Deep Learning and Statistical Learning Theory

In their most basic incarnation, deep learning based algorithms start with training data

$$((X_d^{(i)}, Y_d^{(i)}))_{i=1}^m \colon \Omega \to ([u,v]^d \times [-\mathfrak{D},\mathfrak{D}])^m.$$

To give a concrete example, $X_d^{(i)}$ may consist of different 28×28 pixel grayscale images of handwritten digits and $Y_d^{(i)}$ may consist of corresponding probabilities describing the likelihood of a certain digit to be shown in image $X_d^{(i)}$ [36]. The goal is then to find a functional relation between images and labels and use it for predictive purposes.

Empirical risk minimization (ERM) attempts to solve this prediction problem by minimizing the empirical risk

$$f \mapsto \mathcal{E}_{d,m}(f) := \frac{1}{m}\sum_{i=1}^m \left(f(X_d^{(i)}) - Y_d^{(i)}\right)^2 \tag{4}$$

over a compact hypothesis class $\mathcal{H} \subseteq C([u,v]^d, \mathbb{R})$, resulting in a predictor

$$\widehat{f}_{d,m,\mathcal{H}} \in \operatorname*{argmin}_{f \in \mathcal{H}} \mathcal{E}_{d,m}(f) \tag{5}$$

that is hoped to provide a good approximation of the desired functional relation in the training data. In deep learning, these hypothesis classes consist of deep neural networks which are typically defined via an activation function $\rho \in C(\mathbb{R}, \mathbb{R})$, a number of (hidden) layers $l \in \mathbb{N}$ and an architecture $\mathbf{a} = (a_0, a_1, a_2, \ldots, a_l, a_{l+1}) \in \mathbb{N}^{l+2}$ with $a_0 = d$, $a_{l+1} = 1$ by

$$\mathcal{F}_{\rho,\mathbf{a}}(\boldsymbol{\theta}, x) := \mathcal{A}_{W_l,B_l} \circ \rho_{a_l} \circ \mathcal{A}_{W_{l-1},B_{l-1}} \circ \rho_{a_{l-1}} \circ \cdots \circ \rho_{a_1} \circ \mathcal{A}_{W_0,B_0}(x),$$

where for $k, n \in \mathbb{N}$, $x = (x_i)_{i=1}^n \in \mathbb{R}^n$, a weight matrix $W \in \mathbb{R}^{k \times n}$ and a bias vector $B \in \mathbb{R}^k$, we let $\mathcal{A}_{W,B}(x) := Wx + B$,

$$\boldsymbol{\theta} = ((W_i, B_i))_{i=0}^l \in \bigtimes_{i=0}^{l} \left(\mathbb{R}^{a_{i+1} \times a_i} \times \mathbb{R}^{a_{i+1}}\right)$$

and

$$\rho_n(x) = (\rho(x_i))_{i=1}^n \in \mathbb{R}^n,$$

so that corresponding hypothesis classes are of the form[1]

$$\mathcal{N}_{\rho,\mathbf{a},R}^{u,v} = \left\{ ([u,v]^{a_0} \ni x \mapsto \mathcal{F}_{\rho,\mathbf{a}}(\boldsymbol{\theta}, x)) : \ \boldsymbol{\theta} \in \bigtimes_{i=0}^{l} \left(\mathbb{R}^{a_{i+1} \times a_i} \times \mathbb{R}^{a_{i+1}}\right), \ \|\boldsymbol{\theta}\|_\infty \leq R \right\}. \tag{6}$$

Despite the great practical successes of the "deep learning paradigm" as just described, a theoretical analysis that specifies useful bounds on the number of samples m and the network size (described by $\mathcal{P}(\mathbf{a}) := \sum_{i=0}^l a_{i+1}a_i + a_{i+1}$, the number of free parameters) is far out of reach.

Theoretical tools for achieving such results have been developed within the field of statistical learning theory where it is typically postulated that $((X_d^{(i)}, Y_d^{(i)}))_{i=1}^m$ are i.i.d. samples drawn from the distribution of some (unknown!) data (X_d, Y_d) and that the optimal functional relation between X_d and Y_d is given by the regression function

$$\widehat{f}_d : \left\{ \begin{array}{ccc} [u,v]^d & \to & \mathbb{R} \\ x & \mapsto & \mathbb{E}\left[Y_d \middle| X_d = x\right] \end{array} \right. ,$$

which minimizes the risk

$$f \mapsto \mathcal{E}_d(f) := \mathbb{E}\Big[\left(f(X_d) - Y_d\right)^2 \Big].$$

The minimization of functionals of the form $\mathcal{E}_d$ defined via the probability distribution of (X_d, Y_d) is commonly referred to as a

mathematical learning problem with data (X_d, Y_d) and quadratic loss function,

see, for instance, [13]. Under strong regularity assumptions on the regression function $\widehat{f}_d$ (in the sense that $\widehat{f}_d$ can be well approximated by the hypothesis class $\mathcal{H}$, see [18, 19, 46, 48, 14, 30, 12, 3, 11, 4, 39, 38, 40, 22, 31, 47, 10, 15, 42, 9, 44, 51, 52, 45] for corresponding results with $\mathcal{H}$ consisting of neural networks and $\widehat{f}_d$ satisfying various smoothness assumptions) and the law of (X_d, Y_d) it is then possible to obtain bounds on the sample size m and the number $\mathcal{P}(\mathbf{a})$ of neural network parameters of

$$\mathcal{H} = \mathcal{N}_{\rho,\mathbf{a},R}^{u,v}$$

[1]For a finite index set I and $M \in \mathbb{R}^I$ we define $\|M\|_\infty := \max_{i \in I} |M_i|$ and $\|M\|_2 := \sqrt{\sum_{i \in I} |M_i|^2}$.

to achieve, with high probability, an error

$$\mathbb{E}\left[\left(\widehat{f}_{d,m,\mathcal{H}}(X_d) - \widehat{f}_d(X_d)\right)^2\right] \leq \varepsilon, \tag{7}$$

see for example [13, 34, 37, 50, 24, 5].

Unfortunately it is not clear to what extent these classical techniques are useful for the analysis of real world applications of deep learning methods for (at least) the following reasons:

1. The crucial assumption that the training data consists of i.i.d. samples of an underlying probability distribution is typically not satisfied.
2. Even if the assumption were satisfied, the underlying distribution of (X_d, Y_d) is typically unknown. This implies that it is impossible to verify a priory the regularity assumptions on $\widehat{f}_d$ that are needed for the theory of [13] to be applicable.
3. Even if the the theory of [13] were applicable, since the distribution of X_d is unknown, it is not clear how the quantity $\mathbb{E}\left[\left(\widehat{f}_{d,m,\mathcal{H}}(X_d) - \widehat{f}_d(X_d)\right)^2\right]$ of (7) can be interpreted.
4. The theory developed in [13] operates in an asymptotic regime where the number m of training samples exceeds the "dimension" $\mathcal{P}(\mathbf{a})$ of the hypothesis class $\mathcal{H}$. However, in many applications the number of training samples is fixed and it is not possible to generate more training data at will. Moreover, many successful deep learning applications operate in a regime where there is far less training data available, see also [53].

1.3 Kolmogorov Equations as Learning Problem

We will reformulate the numerical solution of linear Kolmogorov equations as a classical statistical learning problem and demonstrate that in this specific case none of the aforementioned problems appears. The Feynman-Kac formula (3) directly implies that the numerical approximation of $F_d(T, \cdot)$ can be restated as a classical learning problem in the sense of [13] as follows. Let

$$X_d \sim \mathcal{U}([u, v]^d),$$

the uniform distribution on $[u, v]^d$, and let

$$Y_d := \varphi_d(S_T^{X_d})$$

with $(S_t^{X_d})_{t\in[0,T]}$ a stochastic process satisfying the stochastic differential equation

$$dS_t^{X_d} = \sigma_d(S_t^{X_d})dB_t^d + \mu_d(S_t^{X_d})dt \quad \text{and} \quad S_0^{X_d} = X_d. \tag{8}$$

Under suitable conditions it then follows from (3) that $F_d(T, x)$ is the minimizer of the risk functional $\mathcal{E}_d(f) := \mathbb{E}\left[(f(X_d) - Y_d)^2\right]$, that is,

$$\widehat{f}_d(x) = F_d(T, x)$$

for a.e. $x \in [u, v]^d$, see Lemma 3.2 and [6, Proposition 2.7]. As outlined in Subsection 1.2, we thus have that

> *$F_d(T, \cdot)$ is the solution of the mathematical learning problem with data (X_d, Y_d) and quadratic loss function.*

Having reformulated the numerical approximation of $F_d(T,\cdot)$ as a learning problem, a natural next step is to apply the deep learning paradigm, that is, for $m \in \mathbb{N}$ and for i.i.d. samples

$$((X_d^{(i)}, Y_d^{(i)}))_{i=1}^m$$

with $(X_d^{(1)}, Y_d^{(1)}) \sim (X_d, Y_d)$ to minimize the empirical risk (4) over a class $\mathcal{H} = \mathcal{N}_{\rho,\mathbf{a},R}^{u,v}$ of neural networks of a given architecture $\mathbf{a}$, see Equation (6), and to compute $\widehat{f}_{d,m,\mathcal{H}} \in \mathcal{H}$ as in (5).

In [6] this idea has been implemented with suitable classes of deep neural networks of a given architecture as hypothesis class $\mathcal{H}$. In extensive numerical simulations it was observed that the algorithm introduced in [6] is efficient even in very high dimensions. In particular, the simulations carried out in [6] suggest that this algorithm does not suffer from the curse of dimensionality. Related work with similar conclusions can be found in [49, 21, 21, 27, 32, 17, 7, 26, 16]. We emphasize that all these works are purely empirical. Prior to this work no theoretical results confirming the efficiency of deep learning methods applied to high dimensional problems existed.

Two main parameters influence the complexity of the algorithm described above: the "size" of the hypothesis class $\mathcal{H}_{d,\varepsilon}$ (in the case of deep neural networks: the number $\mathcal{P}(\mathbf{a}_{d,\varepsilon})$ of network parameters that need to be optimized) as well as the number of training samples $m_{d,\varepsilon}$ needed to guarantee that, with high probability, the estimate

$$\begin{aligned} &\frac{1}{(v-u)^d} \left\| \widehat{f}_{d,m_{d,\varepsilon},\mathcal{H}_{d,\varepsilon}}(\cdot) - F_d(T,\cdot) \right\|_{L^2[u,v]^d}^2 \\ &= \mathbb{E}\left[\left(\widehat{f}_{d,m_{d,\varepsilon},\mathcal{H}_{d,\varepsilon}}(X_d) - F_d(T,X_d) \right)^2 \right] \leq \varepsilon \end{aligned} \tag{9}$$

holds true. We are interested in their scaling with respect to the precision ε and the dimension d.

Observe that contrary to conventional learning problems, the data distribution (X_d, Y_d) is now explicitly known. Moreover, i.i.d. samples of this distribution can be efficiently simulated as needed (X_d is a simple uniform distribution that can be efficiently simulated using a suitable random number generator and $S_T^{X_d}$ can be simulated by any numerical solver for the stochastic differential equation (8)). In particular, in the mathematical learning problem that arises from our reformulation of the Kolmogorov equation, none of the Problems 1.- 4. described in Subsection 1.2 occurs! We will therefore be able to use tools from statistical learning theory to obtain bounds on the quantities $m_{d,\varepsilon}$, $\mathcal{P}(\mathbf{a}_{d,\varepsilon})$ above.

1.4 Contribution

We show that, whenever for all $d \in \mathbb{N}$ both σ_d and μ_d are affine functions (this includes the important case of the Black-Scholes equation in option pricing), and if the initial conditions $(\varphi_d)_{d\in\mathbb{N}}$ can be approximated by deep neural networks without curse of dimensionality (this can easily shown to be true for a large number of relevant options such as (capped) Basket Call -, Basket Put -, Call on max -, and Call on min - options, see [23, Section 4]), there exists a polynomial $p\colon \mathbb{R}^2 \to \mathbb{R}$ such that for every $d \in \mathbb{N}$ and $\varepsilon \in (0,1)$ it holds that

$$\max\{m_{d,\varepsilon}, \mathcal{P}(\mathbf{a}_{d,\varepsilon})\} \leq p(\varepsilon^{-1}, d).$$

We conclude that the aforementioned deep learning based algorithm *does not suffer from the curse of dimensionality*[2].

[2]Our analysis does not consider the computational cost of solving the ERM problem (5). The latter is typically achieved by stochastic first order optimization methods whose theoretical analysis is completely open to date.

We briefly describe our proof strategy for bounding the error

$$\frac{1}{(v-u)^d}\left\|\widehat{f}_{d,m,\mathcal{H}}(\cdot)-F_d(T,\cdot)\right\|^2_{L^2[u,v]^d} \tag{10}$$

as in (9) with high probability. Let a best approximation in our hypothesis class be defined by

$$\widehat{f}_{d,\mathcal{H}} \in \operatorname*{argmin}_{f\in\mathcal{H}} \frac{1}{(v-u)^d}\left\|f(\cdot)-F_d(T,\cdot)\right\|^2_{L^2[u,v]^d}.$$

By the so-called Bias-Variance Decomposition (see Lemma 2.8) we can decompose the error (10) according to

$$\begin{aligned}&\frac{1}{(v-u)^d}\left\|\widehat{f}_{d,m,\mathcal{H}}(\cdot)-F_d(T,\cdot)\right\|^2_{L^2[u,v]^d}\\ &\qquad=\underbrace{\frac{1}{(v-u)^d}\left\|\widehat{f}_{d,\mathcal{H}}-F_d(T,\cdot)\right\|^2_{L^2[u,v]^d}}_{\text{approximation error}}+\underbrace{\mathcal{E}_d(\widehat{f}_{d,m,\mathcal{H}})-\mathcal{E}_d(\widehat{f}_{d,\mathcal{H}})}_{\text{generalization error}}\end{aligned}$$

and obtain separate bounds on the approximation and generalization error.

For the important case that the functions σ_d and μ_d are affine, which will be one of our main assumptions, the approximation error of neural network hypothesis classes with ReLU activation function

$$\rho(x)=\mathrm{ReLU}(x):=\max\{x,0\}$$

has been analyzed in [23]. By the results of [23] neural network hypothesis classes with ReLU activation function are capable of approximating the solutions $(F_d(T,\cdot))_{d\in\mathbb{N}}$ without incurring the curse of dimensionality whenever the same is true for the initial conditions $(\varphi_d)_{d\in\mathbb{N}}$. By a lucky coincidence, most initial conditions that come from applications in financial engineering are in fact exactly representable by small neural networks with ReLU activation function so that the latter is always satisfied, see for example Subsection 3.4. In Section 3 we extend these approximation results to neural network hypothesis classes with ReLU activation functions whose maximal coefficient size scales at most polynomially in the size of the neural network, see Theorem 3.3. We then leverage these approximation results as well as tools from [13] to obtain probabilistic estimates of the generalization error. These tools require sharp bounds on the covering numbers of hypothesis classes consisting of deep neural networks and we provide such bounds in Section 4 under the condition that the maximal coefficient size scales at most polynomially in the size of the neural network.

The results of [13] require that the regression function $\widehat{f}_d$ as well as all functions in $\mathcal{H}$ are uniformly bounded. This forces us to require that the initial conditions φ_d are uniformly bounded which by (3) implies that also the functions $\widehat{f}_d = F_d(T,\cdot)$ are uniformly bounded. Furthermore, we introduce hypothesis classes of "clipped" neural networks which are defined by

$$\mathcal{N}^{u,v}_{\rho,\mathbf{a},R,\mathfrak{D}}=\left\{\left([u,v]^{a_0}\ni x\mapsto \mathcal{C}_{a_{l+1},\mathfrak{D}}\circ\mathcal{F}_{\rho,\mathbf{a}}(\boldsymbol{\theta},x)\right):\boldsymbol{\theta}\in\bigtimes_{i=0}^{l}\left(\mathbb{R}^{a_{i+1}\times a_i}\times\mathbb{R}^{a_{i+1}}\right),\|\boldsymbol{\theta}\|_\infty\le R\right\}$$

with

$$\mathcal{C}_{a_{l+1},\mathfrak{D}}(x)=\left(\min\{|x_i|,\mathfrak{D}\}\operatorname{sgn}(x_i)\right)_{i=1}^{a_{l+1}}$$

for $x\in\mathbb{R}^{a_{l+1}}$ denoting a clipping function with clipping amplitude $\mathfrak{D}\in(0,\infty)$. In Subsection 2.2 we show that the clipping function $\mathcal{C}_{a_{l+1},\mathfrak{D}}$ can be represented as a small neural network with ReLU activation function so that clipped neural networks with ReLU activation function are in fact standard neural networks with ReLU activation function.

We are now ready to formulate a first specific result of this paper as an appetizer.

Theorem 1.1 (Pricing of European Put Option without Curse of Dimensionality)**.** *Let $(\Omega, \mathcal{G}, \mathbb{P}, (\mathcal{G}_t)_{t\in[0,T]})$ be a filtered probability space which fulfills the usual conditions, let $T, L \in (0,\infty)$, $\mathfrak{D} \in [1,\infty)$, $u \in \mathbb{R}$, $v \in (u,\infty)$ and for all $d \in \mathbb{N}$ let $c_{d,i} \in (0,\infty)$, $i \in \{1,2,\ldots,d\}$, be real numbers which satisfy that $\sum_{i=1}^{d} c_{d,i} = 1$, let $\varphi_d \in C(\mathbb{R}^d, \mathbb{R})$ be given by*

$$\varphi_d(x) = \min\left\{\max\left\{\mathfrak{D} - \sum_{i=1}^{d} c_{d,i} x_i, 0\right\}, \mathfrak{D}\right\}$$

for every $x = (x_i)_{i=1}^d \in \mathbb{R}^d$, let $\sigma_d \in C(\mathbb{R}^d, \mathbb{R}^{d\times d})$, $\mu_d \in C(\mathbb{R}^d, \mathbb{R}^d)$ be affine linear functions which satisfy for every $x \in \mathbb{R}^d$ that

$$\|\sigma_d(x)\|_2 + \|\mu_d(x)\|_2 \leq L(1 + \|x\|_2),$$

let $F_d \in C([0,T]\times\mathbb{R}^d, \mathbb{R})$ be the unique at most polynomially growing viscosity solution[3] of the d-dimensional Kolmogorov PDE

$$\begin{cases} \frac{\partial F_d}{\partial t}(t,x) = \frac{1}{2}\mathrm{Trace}\big(\sigma_d(x)[\sigma_d(x)]^*(\mathrm{Hess}_x F_d)(t,x)\big) + \big\langle \mu_d(x), (\nabla_x F_d)(t,x)\big\rangle_{\mathbb{R}^d} \\ F_d(0,x) = \varphi_d(x) \end{cases}$$

for every $(t,x) \in (0,T)\times\mathbb{R}^d$, let B^d be a d-dimensional $(\mathcal{G}_t)$-Brownian motion, let $X_d \sim \mathcal{U}([u,v]^d)$ be $\mathcal{G}_0$-measurable, let $(S_t^{X_d})_{t\in[0,T]}$ be an adapted stochastic process with continuous sample paths satisfying the stochastic differential equation

$$dS_t^{X_d} = \sigma_d(S_t^{X_d})dB_t^d + \mu_d(S_t^{X_d})dt \quad \text{and} \quad S_0^{X_d} = X_d,$$

let $Y_d := \varphi_d(S_T^{X_d})$ and let $((X_d^{(i)}, Y_d^{(i)}))_{i\in\mathbb{N}}$ be i.i.d. random variables with $(X_d^{(1)}, Y_d^{(1)}) \sim (X_d, Y_d)$. Then there exists $\mathbf{C} \in (0,\infty)$ such that for every $\varepsilon, \varrho \in (0,1)$ and every $d \in \mathbb{N}$ there exist

$$\mathbf{a} = (d, a_1, a_2, 1) \in \mathbb{N}^4,$$

$R \in [1,\infty)$ and $m \in \mathbb{N}$ so that with $\mathcal{H} := \mathcal{N}^{u,v}_{\mathrm{ReLU},\mathbf{a},R,\mathfrak{D}}$ and[4]

$$\widehat{f}_{d,m,\mathcal{H}} \in \operatorname*{argmin}_{f\in\mathcal{H}} \frac{1}{m}\sum_{i=1}^{m}\Big(f(X_d^{(i)}) - Y_d^{(i)}\Big)^2$$

it holds

(i) that

$$\mathcal{P}(\mathbf{a}) \leq \mathbf{C}d^2\varepsilon^{-2}$$

(ii) that

$$\max\{a_1, a_2\} \leq \mathbf{C}d^{3/2}\varepsilon^{-1}$$

(iii) that

$$R \leq \mathbf{C}d^2\varepsilon^{-1}$$

(iv) that

$$m \leq \mathbf{C}d^2\varepsilon^{-4}\big(1 + \ln(d\varepsilon^{-1}\varrho^{-1})\big)$$

(v) and that

$$\mathbb{P}\left[\frac{1}{(v-u)^d}\left\|\widehat{f}_{d,m,\mathcal{H}} - F_d(T,\cdot)\right\|^2_{L^2[u,v]^d} \leq \varepsilon\right] \geq 1 - \varrho.$$

[3]We refer the interested reader to [25] for the definition and properties of viscosity solutions.

[4]More precisely, for every outcome $\omega \in \Omega$ we pick $\widehat{f}_{d,m,\mathcal{H}}(\omega) \in \operatorname{argmin}_{f\in\mathcal{H}} \mathcal{E}_{d,m}(f)(\omega)$ such that the mapping $\omega \to \widehat{f}_{d,m,\mathcal{H}}(\omega)$ is measurable, see Lemma 2.2.

Theorem 1.1 demonstrates that deep learning based ERM succeeds in solving the option pricing problem for European Put options without incurring the curse of dimensionality. A proof will be given in Section 3 below. We hasten to add that the scope of this paper is much wider and that analogous results will be shown in a much more general context, see Theorem 3.7 which states that a result analogous to Theorem 1.1 holds true whenever the initial conditions $(\varphi_d)_{d\in\mathbb{N}}$ can be approximated by neural networks with ReLU activation function without curse of dimensionality. Theorem 3.7, in conjunction with the results of [23], can then be applied to prove the absence of the curse of dimensionality in the pricing of (capped) Basket Call -, Basket Put -, Call on max -, and Call on min - options.

1.5 Outline

The outline is as follows. In Section 2 we present our main result related to the generalization of clipped ReLU networks in a rather general setting. Whenever the regression functions $(\widehat{f}_d)_{d\in\mathbb{N}}$ can be approximated without curse of dimensionality by clipped neural networks, we show that also the number m of required training samples to achieve a desired accuracy ε with high probability does not suffer from the curse of dimensionality. This result is proven in Section 5 using tools from statistical learning theory and the covering number estimates of the hypothesis class of clipped neural networks from Section 4. In Section 3 we extend a result of [23] claiming that the end value of the solution to certain Kolmogorov PDEs can indeed by approximated by clipped neural networks without the curse of dimensionality and therefore our results from Section 2 apply. This gives rise to the quantitative polynomial bounds on the number of samples and the network architecture in Theorem 3.7. As an example we prove that the complexity of pricing European Put options is only growing polynomially in the dimension.

2 Results in Statistical Learning Theory

The present section develops generalization bounds for ERM problems in the spirit of [13]. In Subsection 2.1 we present the basic setting in which we operate, as well as the definition of clipped neural network hypothesis classes. After expanding on these hypothesis classes in Subsection 2.2 we present the main result related to the generalization of clipped ReLU networks in Subsection 2.3. This result, Corollary 2.11, states that approximation results that are free of the curse of dimensionality can be leveraged to generalization results that are free of the curse of dimensionality.

2.1 Basic Setting

This subsection summarizes the basic setting for our main results. The following Setting 2.1 describes a standard statistical mathematical learning problem as defined, for example, in [13].

Setting 2.1 (Mathematical Learning Problem)**.** *Let $u \in \mathbb{R}$, $v \in (u,\infty)$, $\mathfrak{D} \in [1,\infty)$, let $(\Omega, \mathcal{G}, \mathbb{P})$ be a probability space, for all $d \in \mathbb{N}$ let*

$$X_d : \Omega \to [u,v]^d$$

*(*input data*) and*

$$Y_d : \Omega \to [-\mathfrak{D}, \mathfrak{D}]$$

*(*label*) be random variables, let*

$$(X_d^{(i)}, Y_d^{(i)}) \colon \Omega \to [u,v]^d \times [-\mathfrak{D},\mathfrak{D}], \quad i \in \mathbb{N},$$

be i.i.d. random variables with $(X_d^{(1)}, Y_d^{(1)}) \sim (X_d, Y_d)$ *(*training data*), for* $d, m \in \mathbb{N}$ *and a Borel measurable function* $f\colon [u,v]^d \to \mathbb{R}$ *let*

$$\mathcal{E}_d(f) := \int_\Omega \big(f(X_d) - Y_d\big)^2 d\mathbb{P} = \mathbb{E}\Big[\big(f(X_d) - Y_d\big)^2\Big]$$

be the risk *and let*

$$\mathcal{E}_{d,m}(f) := \frac{1}{m}\sum_{i=1}^{m}\Big(f(X_d^{(i)}) - Y_d^{(i)}\Big)^2$$

be the empirical risk*, for every* $d \in \mathbb{N}$ *let* $\mathbb{P}_{X_d}$ *be the image measure of* X_d *on the hypercube* $[u,v]^d$ *and let* $\widehat{f}_d \in L^2(\mathbb{P}_{X_d})$[5] *be the* regression function *defined by*

$$\widehat{f}_d : \left\{ \begin{array}{ccc} [u,v]^d & \to & \mathbb{R} \\ x & \mapsto & \mathbb{E}\left[Y_d \middle| X_d = x\right] \end{array} \right.$$

and for $d, m \in \mathbb{N}$*,* $\omega \in \Omega$ *and compact* $\mathcal{H} \subseteq C([u,v]^d, \mathbb{R})$ *(*hypothesis class*) let*

$$\widehat{f}_{d,\mathcal{H}} \in \operatorname*{argmin}_{f\in\mathcal{H}} \mathcal{E}_d(f) \tag{11}$$

be a best approximation *and let*

$$\widehat{f}_{d,m,\mathcal{H}}(\omega) \in \operatorname*{argmin}_{f\in\mathcal{H}} \mathcal{E}_{d,m}(f)(\omega) \tag{12}$$

be an empirical regression function *such that the mapping* $\Omega \ni \omega \mapsto \widehat{f}_{d,m,\mathcal{H}}(\omega)$ *is measurable.*

We want to emphasize that the minima in (11) and (12) will be attained due to the compactness of our hypothesis class but they need not be unique. For the probability in our generalization error bound (Theorem 2.10) to be well-defined one needs the measurability of the mapping

$$\Omega \ni \omega \mapsto \widehat{f}_{d,m,\mathcal{H}}(\omega).$$

While this technical assumption is often not explicitly stated in the literature on statistical learning theory it is actually crucial for analyzing the generalization error. We prove that in our setting (by choosing a suitable minimizer) measurability can indeed by satisfied.

Lemma 2.2 (Measurability of the Empirical Regression Function)**.** *Assume Setting 2.1, let* $d, m \in \mathbb{N}$ *and let* $\mathcal{H} \subseteq C([u,b]^d, \mathbb{R})$ *be compact. For every* $\omega \in \Omega$ *one can choose the empirical regression function*

$$\widehat{f}_{d,m,\mathcal{H}}(\omega) \in \operatorname*{argmin}_{f\in\mathcal{H}} \frac{1}{m}\sum_{i=1}^{m}\Big(f(X_d^{(i)}(\omega)) - Y_d^{(i)}(\omega)\Big)^2$$

in a way, such that it holds that

(i) the mapping

$$\Omega \ni \omega \mapsto \widehat{f}_{d,m,\mathcal{H}}(\omega)$$

is $\mathcal{G}/\mathcal{B}(\mathcal{H})$*-measurable*

[5] We define the Hilbert-Space $L^2(\mathbb{P}_{X_d})$ as the space of all Borel measurable functions $f \in [u,v]^d \to \mathbb{R}$ with finite norm $\|f\|_{L^2(\mathbb{P}_{X_d})} = \big(\int_{[u,v]^d} f^2\, d\mathbb{P}_{X_d}\big)^{1/2} = \mathbb{E}\big[(f(X_d))^2\big]^{1/2} < \infty$ where functions which coincide $\mathbb{P}_{X_d}$-a.s. are identified as usual.

(ii) and the mapping

$$\Omega \ni \omega \mapsto \mathcal{E}_d\big(\widehat{f}_{d,m,\mathcal{H}}(\omega)\big)$$

is $\mathcal{G}/\mathcal{B}(\mathbb{R})$-measurable.

Proof of Lemma 2.2. First observe that $\mathcal{H}$ is a separable metric space induced by the uniform norm and that for every $f \in \mathcal{H}$ the mapping

$$\Omega \ni \omega \mapsto \mathcal{E}_{d,m}(f)(\omega)$$

is $\mathcal{G}/\mathcal{B}(\mathbb{R})$-measurable. By the reverse triangle inequality we get that

$$\begin{aligned}\Big|\mathcal{E}_{d,m}(f)^{1/2} - \mathcal{E}_{d,m}(g)^{1/2}\Big| &= \frac{1}{\sqrt{m}}\left|\left\|(f(X_d^{(i)}) - Y_d^{(i)})_{i=1}^m\right\|_2 - \left\|(g(X_d^{(i)}) - Y_d^{(i)})_{i=1}^m\right\|_2\right| \\ &\leq \frac{1}{\sqrt{m}}\left\|(f(X_d^{(i)}) - g(X_d^{(i)}))_{i=1}^m\right\|_2 \leq \sup_{x\in[u,v]^d}|f(x) - g(x)|\end{aligned}$$

for every $f, g \in \mathcal{H}$. This shows that for every $\omega \in \Omega$ the function

$$\mathcal{H} \ni f \mapsto \mathcal{E}_{d,m}(f)(\omega)$$

is continuous which implies that the function

$$\Omega \times \mathcal{H} \ni (\omega, f) \mapsto \mathcal{E}_{d,m}(f)(\omega)$$

is a Carathéodory function. The Measurable Maximum Theorem in [1, Theorem 18.19 with $(S,\Sigma) \leftarrow (\Omega,\mathcal{G})$, $X \leftarrow \mathcal{H}$, $f(s,x) \leftarrow -\mathcal{E}_{d,m}(x)(s)$ and $\varphi(s) = \mathcal{H}$ for every $s \in S$] assures that the set-valued function of minimizers of

$$\min_{f\in\mathcal{H}} \mathcal{E}_{d,m}(f)$$

admits a measurable selector. That is to say, there exists a $\mathcal{G}/\mathcal{B}(\mathcal{H})$-measurable mapping $\widehat{f}_{d,m,\mathcal{H}}\colon \Omega \to \mathcal{H}$ such that for every $\omega \in \Omega$ it holds that

$$\widehat{f}_{d,m,\mathcal{H}}(\omega) \in \operatorname*{argmin}_{f\in\mathcal{H}} \mathcal{E}_{d,m}(f)(\omega).$$

This establishes item (i). For the proof of the second item observe that the risk $\mathcal{E}_d\colon \mathcal{H} \to \mathbb{R}$ is continuous and thus $\mathcal{B}(\mathcal{H})/\mathcal{B}(\mathbb{R})$-measurable. Indeed, an analogous computation as for the empirical risk above shows that for $f, g \in \mathcal{H}$ it holds that

$$\begin{aligned}\Big|\mathcal{E}_d(f)^{1/2} - \mathcal{E}_d(g)^{1/2}\Big| &= \Big|\|f(X_d) - Y_d\|_{L^2(\mathbb{P})} - \|g(X_d) - Y_d\|_{L^2(\mathbb{P})}\Big| \\ &\leq \|f(X_d) - g(X_d)\|_{L^2(\mathbb{P})} \leq \sup_{x\in[u,v]^d}|f(x) - g(x)|.\end{aligned}$$

This yields the claim as compositions of measurable functions are again measurable. □

For our theory we assume in Setting 2.1 that the empirical regression function is chosen in the sense of Lemma 2.2. This allows us to view the risk of the empirical regression function as a random variable

$$\Omega \ni \omega \mapsto \mathcal{E}_d\big(\widehat{f}_{d,m,\mathcal{H}}(\omega)\big)$$

which is necessary for bounding the generalization error.

The following setting describes suitable hypothesis classes based on deep artificial neural networks. Our definition of (artificial feedforward) neural networks with ReLU activation function is completely standard, except for the composition with a "clipping function" $\mathcal{C}_{\mathfrak{D},n}$ in (13) below which will be clarified in Subsection 2.2. From now on we will only consider neural networks with ReLU activation function and therefore we will omit writing the index $\rho = \mathrm{ReLU}$ in our notation.

Setting 2.3 (Neural Networks)**.** *For $\mathfrak{D} \in (0,\infty)$, $k,n \in \mathbb{N}$, $W \in \mathbb{R}^{k\times n}$, $B \in \mathbb{R}^k$ let $\mathcal{A}_{W,B} \in C(\mathbb{R}^n,\mathbb{R}^k)$ be the affine linear mapping given by*

$$\mathcal{A}_{W,B}(x) = Wx + B,$$

let $\mathrm{ReLU}_n \in C(\mathbb{R}^n,\mathbb{R}^n)$ *be the n-dimensional* Rectified Linear Unit function *given by*

$$\mathrm{ReLU}_n(x) = \big(\max\{x_i, 0\}\big)_{i=1}^n$$

and let $\mathcal{C}_{n,\mathfrak{D}} \in C(\mathbb{R}^n,\mathbb{R}^n)$ be the n-dimensional clipping function *given by*

$$\mathcal{C}_{n,\mathfrak{D}}(x) = \big(\min\{|x_i|,\mathfrak{D}\}\operatorname{sgn}(x_i)\big)_{i=1}^n$$

for every $x = (x_i)_{i=1}^n \in \mathbb{R}^n$. For $l \in \mathbb{N}_0$, $\mathfrak{D} \in (0,\infty)$ and a network architecture

$$\mathbf{a} = (a_0, a_1, \ldots, a_l, a_{l+1}) \in \mathbb{N}^{l+2}$$

let the number of hidden layers $\mathcal{L}(\mathbf{a})$ *be given by*

$$\mathcal{L}(\mathbf{a}) = l,$$

the number of parameters $\mathcal{P}(\mathbf{a})$ *be given by*

$$\mathcal{P}(\mathbf{a}) = \sum_{i=0}^{l} a_{i+1}a_i + a_{i+1}$$

and for $x \in \mathbb{R}^{a_0}$ and parameters

$$\boldsymbol{\theta} = \big((W_i,B_i)\big)_{i=0}^l \in \bigtimes_{i=0}^{l} \big(\mathbb{R}^{a_{i+1}\times a_i} \times \mathbb{R}^{a_{i+1}}\big) \simeq \mathbb{R}^{\mathcal{P}(\mathbf{a})}$$

let the neural network $\mathcal{F}_{\mathbf{a}} : \mathbb{R}^{\mathcal{P}(\mathbf{a})} \times \mathbb{R}^{a_0} \to \mathbb{R}^{a_{l+1}}$ *be defined as*

$$\mathcal{F}_{\mathbf{a}}(\boldsymbol{\theta},x) = \mathcal{A}_{W_l,B_l} \circ \mathrm{ReLU}_{a_l} \circ \mathcal{A}_{W_{l-1},B_{l-1}} \circ \mathrm{ReLU}_{a_{l-1}} \circ \cdots \circ \mathrm{ReLU}_{a_1} \circ \mathcal{A}_{W_0,B_0}(x),$$

and let the clipped neural network $\mathcal{F}_{\mathbf{a},\mathfrak{D}} : \mathbb{R}^{\mathcal{P}(\mathbf{a})} \times \mathbb{R}^{a_0} \to \mathbb{R}^{a_{l+1}}$ *be defined as*

$$\mathcal{F}_{\mathbf{a},\mathfrak{D}}(\boldsymbol{\theta},x) = \mathcal{C}_{a_{l+1},\mathfrak{D}} \circ \mathcal{F}_{\mathbf{a}}(\boldsymbol{\theta},x).$$

For $l \in \mathbb{N}_0$, $R,\mathfrak{D} \in (0,\infty)$, $u \in \mathbb{R}$, $v \in (u,\infty)$, $\mathbf{a} = (a_0,a_1,\ldots,a_l,a_{l+1}) \in \mathbb{N}^{l+2}$ let

$$\mathcal{N}_{\mathbf{a},R}^{u,v} = \left\{ \big([u,v]^{a_0} \ni x \mapsto \mathcal{F}_{\mathbf{a}}(\boldsymbol{\theta},x)\big) :\ \boldsymbol{\theta} \in \bigtimes_{i=0}^{l} \big(\mathbb{R}^{a_{i+1}\times a_i} \times \mathbb{R}^{a_{i+1}}\big),\ \|\boldsymbol{\theta}\|_\infty \le R \right\}$$

*(*hypothesis class of neural networks*) and let*

$$\mathcal{N}_{\mathbf{a},R,\mathfrak{D}}^{u,v} = \left\{ \big([u,v]^{a_0} \ni x \mapsto \mathcal{F}_{\mathbf{a},\mathfrak{D}}(\boldsymbol{\theta},x)\big) :\ \boldsymbol{\theta} \in \bigtimes_{i=0}^{l} \big(\mathbb{R}^{a_{i+1}\times a_i} \times \mathbb{R}^{a_{i+1}}\big),\ \|\boldsymbol{\theta}\|_\infty \le R \right\} \quad (13)$$

*(*hypothesis class of clipped neural networks*) and for $d \in \mathbb{N}$ let*

$$\mathbf{A}_d = \bigcup_{l\in\mathbb{N}_0} \{(a_0,a_1,\ldots,a_l,a_{l+1}) \in \mathbb{R}^{l+2} :\ a_0 = d,\ a_{l+1} = 1\}$$

*(*admissible network architectures*).*

2.2 Clipped Neural Networks are Standard Neural Networks

In this subsection we clarify the role of the function $\mathcal{C}_{a_{l+1},\mathfrak{D}}$ in the definition of $\mathcal{N}^{u,v}_{\mathbf{a},R,\mathfrak{D}}$ in Setting 2.3. The neural network classes $\mathcal{N}^{u,v}_{\mathbf{a},R,\mathfrak{D}}$ are somewhat non-standard in the sense that the function $\mathcal{C}_{a_{l+1},\mathfrak{D}}$ is applied to the output of a neural network, see Setting 2.3. The reason for our choice of this definition is that our main results will require that the set of neural networks over which the empirical risk minimization problem is solved consists of uniformly bounded functions (such boundedness assumptions are in fact standard in statistical learning theory, see [13]). Lemma 2.6 shows that the clipping function $\mathcal{C}_{a_{l+1},\mathfrak{D}}$ can be represented as a small neural network which implies that the seemingly non-standard classes $\mathcal{N}^{u,v}_{\mathbf{a},R,\mathfrak{D}}$ are actually conventional neural network classes that can be trained with standard methods [35, 28, 33, 43].

Lemma 2.4. *Assume Setting 2.3, let $n \in \mathbb{N}$ and $\mathfrak{D} \in (0,\infty)$. Then for all $x \in \mathbb{R}^n$ it holds that*

$$\mathcal{C}_{n,\mathfrak{D}}(x) = -\operatorname{ReLU}_n((\mathfrak{D})_{i=1}^n - \operatorname{ReLU}_n(x)) + \operatorname{ReLU}_n(-\operatorname{ReLU}_n(-x) + (\mathfrak{D})_{i=1}^n).$$

Proof of Lemma 2.4. Without loss of generality we may assume that $n = 1$. We distinguish four cases:
$\mathbf{x < -\mathfrak{D}}$. In this case it holds that $\operatorname{ReLU}_1(x) = 0$, $\operatorname{ReLU}_1(-x) = -x$ and $\operatorname{ReLU}_1(x+\mathfrak{D}) = 0$. Therefore it holds that

$$\begin{aligned}\mathcal{C}_{1,\mathfrak{D}}(x) &= -\mathfrak{D} = -\operatorname{ReLU}_1(\mathfrak{D}) + \operatorname{ReLU}_1(x+\mathfrak{D}) \\ &= -\operatorname{ReLU}_1(\mathfrak{D} - \operatorname{ReLU}_1(x)) + \operatorname{ReLU}_1(-\operatorname{ReLU}_1(-x) + \mathfrak{D}).\end{aligned}$$

$\mathbf{x > \mathfrak{D}}$. In this case it holds that $\operatorname{ReLU}_1(x) = x$, $\operatorname{ReLU}_1(-x) = 0$ and $\operatorname{ReLU}_1(\mathfrak{D}-x) = 0$. Therefore it holds that

$$\begin{aligned}\mathcal{C}_{1,\mathfrak{D}}(x) &= \mathfrak{D} = -\operatorname{ReLU}_1(\mathfrak{D} - x) + \operatorname{ReLU}_1(\mathfrak{D}) \\ &= -\operatorname{ReLU}_1(\mathfrak{D} - \operatorname{ReLU}_1(x)) + \operatorname{ReLU}_1(-\operatorname{ReLU}_1(-x) + \mathfrak{D}).\end{aligned}$$

$\mathbf{x \leq \mathfrak{D}}$ and $\mathbf{x \geq 0}$. In this case it holds that $\operatorname{ReLU}_1(x) = x$, $\operatorname{ReLU}_1(-x) = 0$ and $\operatorname{ReLU}_1(\mathfrak{D} - x) = \mathfrak{D} - x$. Therefore it holds that

$$\begin{aligned}\mathcal{C}_{1,\mathfrak{D}}(x) &= x = -\operatorname{ReLU}_1(\mathfrak{D} - x) + \operatorname{ReLU}_1(\mathfrak{D}) \\ &= -\operatorname{ReLU}_1(\mathfrak{D} - \operatorname{ReLU}_1(x)) + \operatorname{ReLU}_1(-\operatorname{ReLU}_1(-x) + \mathfrak{D}).\end{aligned}$$

$\mathbf{x \geq -\mathfrak{D}}$ and $\mathbf{x \leq 0}$. In this case it holds that $\operatorname{ReLU}_1(x) = 0$, $\operatorname{ReLU}_1(-x) = -x$ and $\operatorname{ReLU}_1(x+\mathfrak{D}) = x + \mathfrak{D}$. Therefore it holds that

$$\begin{aligned}\mathcal{C}_{1,\mathfrak{D}}(x) &= x = -\operatorname{ReLU}_1(\mathfrak{D}) + \operatorname{ReLU}_1(x+\mathfrak{D}) \\ &= -\operatorname{ReLU}_1(\mathfrak{D} - \operatorname{ReLU}_1(x)) + \operatorname{ReLU}_1(-\operatorname{ReLU}_1(-x) + \mathfrak{D}).\end{aligned}$$

This proves the lemma. □

The following contraction property will be useful in several places later on.

Corollary 2.5 (Contraction Property of the Clipping Function)**.** *Assume Setting 2.3, let $n \in \mathbb{N}$ and $\mathfrak{D} \in (0,\infty)$. Then for all $x, y \in \mathbb{R}^n$ it holds that*

$$\|\mathcal{C}_{n,\mathfrak{D}}(x) - \mathcal{C}_{n,\mathfrak{D}}(y)\|_2 \leq \|x - y\|_2.$$

Proof of Corollary 2.5. The proof follows from a straightforward calculation. □

The next lemma states that the clipping function $\mathcal{C}_{n,\mathfrak{D}}$ can be represented as a small neural network.

Lemma 2.6 (Clipping Function as Neural Network)**.** *Assume Setting 2.3, let $n \in \mathbb{N}$, $\mathfrak{D} \in (0,\infty)$ and let*

$$\mathbf{a} = (n, 2n, 2n, n).$$

Then there exists $\boldsymbol{\theta} \in \mathbb{R}^{\mathcal{P}(\mathbf{a})}$ such that for all $x \in \mathbb{R}^n$ it holds that

$$\mathcal{C}_{n,\mathfrak{D}}(x) = \mathcal{F}_{\mathbf{a}}(\boldsymbol{\theta}, x).$$

Proof of Lemma 2.6. Denote by $\mathbf{I}$ the n-dimensional identity matrix and let

$$\boldsymbol{\theta} \in \left(\mathbb{R}^{2n\times n} \times \mathbb{R}^{2n}\right) \times \left(\mathbb{R}^{2n\times 2n} \times \mathbb{R}^{2n}\right) \times \left(\mathbb{R}^{n\times 2n} \times \mathbb{R}^{n}\right) \simeq \mathbb{R}^{\mathcal{P}(\mathbf{a})}$$

be given by

$$\boldsymbol{\theta} := \left(\left(\begin{bmatrix}\mathbf{I}\\ -\mathbf{I}\end{bmatrix}, \begin{bmatrix}0\\ \vdots\\ 0\end{bmatrix}\right), \left(\begin{bmatrix}-\mathbf{I} & \mathbf{0}\\ \mathbf{0} & -\mathbf{I}\end{bmatrix}, \begin{bmatrix}\mathfrak{D}\\ \vdots\\ \mathfrak{D}\end{bmatrix}\right), \left(\begin{bmatrix}-\mathbf{I} & \mathbf{I}\end{bmatrix}, \begin{bmatrix}0\\ \vdots\\ 0\end{bmatrix}\right)\right).$$

Lemma 2.4 and the Definition of $\mathcal{F}_{\mathbf{a}}(\boldsymbol{\theta}, \cdot)$ now prove the claim. □

The next result shows that the "clipped" neural network classes $\mathcal{N}^{u,v}_{\mathbf{a},R,\mathfrak{D}}$ are in fact subsets of "non-clipped" neural network classes.

Corollary 2.7. *Assume Setting 2.3, let $l \in \mathbb{N}_0$, $u \in \mathbb{R}$, $v \in (u,\infty)$, $\mathfrak{D}, R \in (0,\infty)$, let $\mathbf{a} = (a_0, a_1, \ldots, a_l, a_{l+1}) \in \mathbb{N}^{l+2}$ and*

$$\mathbf{b} := (a_0, a_1, \ldots, a_l, a_{l+1}, 2a_{l+1}, 2a_{l+1}, a_{l+1}) \in \mathbb{N}^{l+5}.$$

Then it holds that

$$\mathcal{N}^{u,v}_{\mathbf{a},R,\mathfrak{D}} \subseteq \mathcal{N}^{u,v}_{\mathbf{b},R}.$$

Proof of Corollary 2.7. The proof follows directly from Lemma 2.6. □

2.3 A Generalization Result

The following result is often referred to as the "Bias-Variance Decomposition".

Lemma 2.8. *Assume Setting 2.1, let $d \in \mathbb{N}$ and let $\mathcal{H} \subseteq C([u,v]^d, \mathbb{R})$ be compact. Then for every $f \in C\left([u,v]^d, \mathbb{R}\right)$ it holds that*

$$\left\|f - \widehat{f}_d\right\|^2_{L^2(\mathbb{P}_{X_d})} = \underbrace{\left\|\widehat{f}_{d,\mathcal{H}} - \widehat{f}_d\right\|^2_{L^2(\mathbb{P}_{X_d})}}_{\text{approximation error (bias)}} + \underbrace{\mathcal{E}_d(\widehat{f}_{d,m,\mathcal{H}}) - \mathcal{E}_d(\widehat{f}_{d,\mathcal{H}})}_{\text{generalization error (variance)}} + \mathcal{E}_d(f) - \mathcal{E}_d(\widehat{f}_{d,m,\mathcal{H}}).$$

Proof of Lemma 2.8. For $f \in C\left([u,v]^d, \mathbb{R}\right)$ it holds that

$$\begin{aligned}\mathcal{E}_d(f) &= \mathbb{E}\left[\left(f(X_d) - \widehat{f}_d(X_d) + \widehat{f}_d(X_d) - Y_d\right)^2\right]\\ &= \mathbb{E}\left[\left(f(X_d) - \widehat{f}_d(X_d)\right)^2\right] + \mathbb{E}\left[\left(\widehat{f}_d(X_d) - Y_d\right)^2\right]\\ &\quad + 2\mathbb{E}\left[\left(f(X_d) - \widehat{f}_d(X_d)\right)\left(\widehat{f}_d(X_d) - Y_d\right)\right]\end{aligned} \tag{14}$$

Observe that, due to the fact that $\widehat{f}_d(X_d) = \mathbb{E}\left[Y_d \middle| X_d\right]$, it holds by the tower property of the conditional expectation that

$$\begin{aligned}
&\mathbb{E}\left[\left(f(X_d) - \widehat{f}_d(X_d)\right)\left(\widehat{f}_d(X_d) - Y_d\right)\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[\left(f(X_d) - \widehat{f}_d(X_d)\right)\right)\left(\widehat{f}_d(X_d) - Y_d\right)\middle| X_d\right]\right] \\
&= \mathbb{E}\left[\left(f(X_d) - \widehat{f}_d(X_d)\right)\left(\widehat{f}_d(X_d) - \mathbb{E}\left[Y_d \middle| X_d\right]\right)\right] = 0
\end{aligned}$$

which, together with (14), implies that

$$\mathcal{E}_d(f) - \mathcal{E}_d(\widehat{f}_d) = \mathbb{E}\left[\left(f(X_d) - \widehat{f}_d(X_d)\right)^2\right] = \left\|f - \widehat{f}_d\right\|^2_{L^2(\mathbb{P}_{X_d})}. \tag{15}$$

Therefore, it follows that for every $f \in C\left([u,v]^d, \mathbb{R}\right)$ it holds that

$$\left\|f - \widehat{f}_d\right\|^2_{L^2(\mathbb{P}_{X_d})} = \mathcal{E}_d(f) - \mathcal{E}_d(\widehat{f}_{d,m,\mathcal{H}}) + \mathcal{E}_d(\widehat{f}_{d,m,\mathcal{H}}) - \mathcal{E}_d(\widehat{f}_{d,\mathcal{H}}) + \mathcal{E}_d(\widehat{f}_{d,\mathcal{H}}) - \mathcal{E}_d(\widehat{f}_d).$$

Finally, applying (15) (with $f \leftarrow \widehat{f}_{d,\mathcal{H}}$) implies that

$$\left\|f - \widehat{f}_d\right\|^2_{L^2(\mathbb{P}_{X_d})} = \mathcal{E}_d(f) - \mathcal{E}_d(\widehat{f}_{d,m,\mathcal{H}}) + \mathcal{E}_d(\widehat{f}_{d,m,\mathcal{H}}) - \mathcal{E}_d(\widehat{f}_{d,\mathcal{H}}) + \left\|\widehat{f}_{d,\mathcal{H}} - \widehat{f}_d\right\|^2_{L^2(\mathbb{P}_{X_d})}$$

which proves the lemma. □

The next lemma states that the function $\widehat{f}_{d,\mathcal{H}}$ is a best approximation of $\widehat{f}_d$ in $\mathcal{H}$ with respect to the $L^2(\mathbb{P}_{X_d})$ norm.

Lemma 2.9. *Assume Setting 2.1, let $d \in \mathbb{N}$ and let $\mathcal{H} \subseteq C([u,v]^d, \mathbb{R})$ be compact. Then for every $f \in \mathcal{H}$ it holds that*

$$\left\|\widehat{f}_{d,\mathcal{H}} - \widehat{f}_d\right\|^2_{L^2(\mathbb{P}_{X_d})} \leq \left\|f - \widehat{f}_d\right\|^2_{L^2(\mathbb{P}_{X_d})}.$$

Proof of Lemma 2.9. Observe that by assumption for all $f \in \mathcal{H}$ it holds that

$$\mathcal{E}_d(f) - \mathcal{E}_d(\widehat{f}_{d,\mathcal{H}}) \geq 0$$

which, by Lemma 2.8, implies that

$$\left\|\widehat{f}_{d,\mathcal{H}} - \widehat{f}_d\right\|^2_{L^2(\mathbb{P}_{X_d})} \leq \left\|f - \widehat{f}_d\right\|^2_{L^2(\mathbb{P}_{X_d})}.$$

□

The following theorem describes our main result related to the generalization of clipped ReLU networks.

Theorem 2.10 (Generalization Error Bound)**.** *Assume Settings 2.1 and 2.3, let $h \in C((0,\infty)^5, \mathbb{R})$ be given by*

$$h(x) = 128\mathfrak{D}^4 x_1^2\Big[1 + x_2 + x_4\Big(\ln\left(64\mathfrak{D}\max\{1,|u|,|v|\}x_1\right) + \left(x_5+1\right)\left(x_3+2\right)\Big)\Big],$$

let $\varepsilon, \varrho \in (0,1)$, $d \in \mathbb{N}$, $\mathbf{a} \in \mathbf{A}_d$, $R \in [1,\infty)$, let $\mathcal{H} := \mathcal{N}^{u,v}_{\mathbf{a},R,\mathfrak{D}}$ and

$$m \geq h\left(\varepsilon^{-1}, \ln(\varrho^{-1}), \ln(R\|\mathbf{a}\|_\infty), \mathcal{P}(\mathbf{a}), \mathcal{L}(\mathbf{a})\right).$$

Then it holds that

$$\mathbb{P}\left[\mathcal{E}_d\left(\widehat{f}_{d,m,\mathcal{H}}\right) - \mathcal{E}_d\left(\widehat{f}_{d,\mathcal{H}}\right) \leq \varepsilon\right] \geq 1 - \varrho.$$

A proof of Therem 2.10 will be given in Section 5. The next result shows how to use Theorem 2.10 to leverage bounds on the approximation error to obtain quantitative bounds on the generalization error.

Corollary 2.11 (Approximation implies Generalization)**.** *Assume Settings 2.1 and 2.3 and assume that for every $d \in \mathbb{N}$, $\varepsilon \in (0,1)$ there exist $\mathbf{a}_{d,\varepsilon} \in \mathbf{A}_d$, $R_{d,\varepsilon} \in [1,\infty)$ and a clipped neural network*

$$g_{d,\varepsilon} \in \mathcal{H}_{d,\varepsilon} := \mathcal{N}^{u,v}_{\mathbf{a}_{d,\varepsilon}, R_{d,\varepsilon}, \mathfrak{D}}$$

such that it holds that

$$\left\| \widehat{f}_d - g_{d,\varepsilon} \right\|^2_{L^2(\mathbb{P}_{X_d})} \leq \varepsilon/2.$$

Let $h \in C((0,\infty)^5, \mathbb{R})$ be given by

$$h(x) = 128\mathfrak{D}^4 x_1^2 \Big[1 + x_2 + x_4 \Big(\ln\big(64\mathfrak{D}\max\{1,|u|,|v|\}x_1\big) + \big(x_5+1\big)\big(x_3+2\big) \Big) \Big],$$

let $d \in \mathbb{N}$, $\varepsilon, \varrho \in (0,1)$ and

$$m \geq h\left(2\varepsilon^{-1}, \ln(\varrho^{-1}), \ln(R_{d,\varepsilon}\|\mathbf{a}_{d,\varepsilon}\|_\infty), \mathcal{P}(\mathbf{a}_{d,\varepsilon}), \mathcal{L}(\mathbf{a}_{d,\varepsilon})\right).$$

Then it holds that

$$\mathbb{P}\left[\left\| \widehat{f}_{d,m,\mathcal{H}_{d,\varepsilon}} - \widehat{f}_d \right\|^2_{L^2(\mathbb{P}_{X_d})} \leq \varepsilon \right] \geq 1 - \varrho.$$

Proof of Corollary 2.11. Since by assumption it holds that $g_{d,\varepsilon} \in \mathcal{H}_{d,\varepsilon}$ and

$$\left\| \widehat{f}_d - g_{d,\varepsilon} \right\|^2_{L^2(\mathbb{P}_{X_d})} \leq \varepsilon/2,$$

Lemma 2.9 implies that

$$\left\| \widehat{f}_d - \widehat{f}_{d,\mathcal{H}_{d,\varepsilon}} \right\|^2_{L^2(\mathbb{P}_{X_d})} \leq \varepsilon/2. \tag{16}$$

The Bias-Variance decomposition in Lemma 2.8 together with (16) hence assures that

$$\left\| \widehat{f}_d - \widehat{f}_{d,m,\mathcal{H}_{d,\varepsilon}} \right\|^2_{L^2(\mathbb{P}_{X_d})} \leq \varepsilon/2 + \mathcal{E}_d(\widehat{f}_{d,m,\mathcal{H}_{d,\varepsilon}}) - \mathcal{E}_d(\widehat{f}_{d,\mathcal{H}_{d,\varepsilon}}). \tag{17}$$

Theorem 2.10 (with $\varepsilon \leftarrow \varepsilon/2$) implies that

$$\mathbb{P}\left[\mathcal{E}_d(\widehat{f}_{d,m,\mathcal{H}_{d,\varepsilon}}) - \mathcal{E}_d(\widehat{f}_{d,\mathcal{H}_{d,\varepsilon}}) \leq \varepsilon/2 \right] \geq 1 - \varrho. \tag{18}$$

Finally, (17) and (18) directly imply the desired claim. □

The previous result in particular implies that, whenever the family $(\widehat{f}_d)_{d\in\mathbb{N}}$ from the mathematical learning problem of Setting 2.1 can be approximated by neural networks without curse of dimensionality, then also the number m of required training samples to achieve a desired accuracy with high probability does not suffer from the curse of dimensionality, either. A version of this statement is given in the next result.

Corollary 2.12 (Approximation without Curse implies Generalization without Curse)**.** *Assume Settings 2.1 and 2.3 and assume that there exists a polynomial $q : \mathbb{R}^2 \to \mathbb{R}$ such that for all $d \in \mathbb{N}$ and $\varepsilon \in (0,1)$ there is $\mathbf{a}_{d,\varepsilon} \in \mathbf{A}_d$, $R_{d,\varepsilon} \in [1,\infty)$ and*

$$g_{d,\varepsilon} \in \mathcal{N}^{u,v}_{\mathbf{a}_{d,\varepsilon}, R_{d,\varepsilon}, \mathfrak{D}} =: \mathcal{H}_{d,\varepsilon}$$

with

$$\max\{\ln(R_{d,\varepsilon}), \mathcal{P}(\mathbf{a}_{d,\varepsilon})\} \leq q(d, \varepsilon^{-1})$$

and

$$\left\| \widehat{f}_d - g_{d,\varepsilon} \right\|^2_{L^2(\mathbb{P}_{X_d})} \leq \varepsilon/2.$$

Then there exists a polynomial $s : \mathbb{R}^2 \to \mathbb{R}$ such that for all $d \in \mathbb{N}$, $\varepsilon, \varrho \in (0,1)$ and all

$$m \geq s(d, \varepsilon^{-1})(1 + \ln(\varrho^{-1}))$$

it holds that

$$\mathbb{P}\left[\left\| \widehat{f}_{d,m,\mathcal{H}_{d,\varepsilon}} - \widehat{f}_d \right\|^2_{L^2(\mathbb{P}_{X_d})} \leq \varepsilon \right] \geq 1 - \varrho.$$

Proof of Corollary 2.12. Observe that for every $d \in \mathbb{N}$, $\varepsilon \in (0,1)$ it holds that

$$\max\{\ln(\|\mathbf{a}_{d,\varepsilon}\|_\infty), \mathcal{L}(\mathbf{a}_{d,\varepsilon})\} \leq \mathcal{P}(\mathbf{a}_{d,\varepsilon}) \leq q(d, \varepsilon^{-1})$$

and that the function $h \in C((0,\infty)^5, \mathbb{R})$ from Corollary 2.11 satisfies

$$h(x) \leq 128\mathfrak{D}^4 x_1^2 (1 + x_2)\Big[1 + x_4\Big(\ln\big(64\mathfrak{D} \max\{1, |u|, |v|\} x_1\big) + \big(x_5 + 1\big)\big(x_3 + 2\big)\Big)\Big]$$

for every $x \in (0,\infty)^5$. This and the basic inequality $\ln(z) \leq z - 1$ for $z \in (0,\infty)$ establishes that for every $d \in \mathbb{N}$, $\varepsilon, \varrho \in (0,1)$ it holds that

$$h\left(2\varepsilon^{-1}, \ln(\varrho^{-1}), 2q(d, \varepsilon^{-1}), q(d, \varepsilon^{-1}), q(d, \varepsilon^{-1}))\right) \leq s(d, \varepsilon^{-1})(1 + \ln(\varrho^{-1}))$$

where the polynomial $s\colon \mathbb{R}^2 \to \mathbb{R}$ is defined by

$$\begin{aligned} s(x) = 512\mathfrak{D}^4 x_2^2 \Big[1 + q(x_1, x_2)\Big(&\ln\big(128\mathfrak{D} \max\{1, |u|, |v|\}\big) \\ &+ x_2 - 1 + \big(q(x_1, x_2) + 1\big)\big(2q(x_1, x_2) + 2\big)\Big)\Big] \end{aligned}$$

and thus Corollary 2.12 is a direct consequence of Corollary 2.11. □

3 Applications for the Numerical Approximation of High Dimensional PDEs

In the present section we apply the general results of Section 2 to the solution of high dimensional Kolmogorov PDEs. To this end we first reformulate the solution of a Kolmogorov PDE as a mathematical learning problem in Subsection 3.1. The next Subsection 3.2 establishes suitable approximation results for solutions of Kolmogorov PDEs. The following Subsection 3.3 contains the main result of this paper, Theorem 3.7, which states that ERM with deep neural networks is capable of numerically solving Kolmogorov PDEs with affine coefficients without curse of dimensionality. As a specific application we show in Subsection 3.4 that ERM with deep neural networks is capable of solving the Black-Scholes pricing problem for European Put options without curse of dimensionality.

3.1 Kolmogorov PDEs as Learning Problem

The following setting will be frequently used.

Setting 3.1. *Assume Setting 2.1, let $T, L \in (0,\infty)$, for all $d \in N$ let $\varphi_d \in C(\mathbb{R}^d, [-\mathfrak{D}, \mathfrak{D}])$ and let $\sigma_d \in C(\mathbb{R}^d, \mathbb{R}^{d\times d})$, $\mu_d \in C(\mathbb{R}^d, \mathbb{R}^d)$ be affine linear functions with*

$$\|\sigma_d(x)\|_2 + \|\mu_d(x)\|_2 \leq L(1 + \|x\|_2)$$

for all $x \in \mathbb{R}^d$. Let the probability space $(\Omega, \mathcal{G}, \mathbb{P})$ be equipped with a filtration $(\mathcal{G}_t)_{t\in[0,T]}$ which fulfills the usual conditions, for every $d \in \mathbb{N}$ let

$$(B_t^d)_{t\in[0,T]} \colon [0,T] \times \Omega \to \mathbb{R}^d$$

be a d-dimensional $(\mathcal{G}_t)$-Brownian motion, let the input data $X_d \colon \Omega \to [u,v]^d$ be $\mathcal{G}_0$-measurable and uniformly distributed on $[u,v]^d$, let

$$(S_t^{X_d})_{t\in[0,T]} \colon [0,T] \times \Omega \to \mathbb{R}^d$$

be the, up to indistinguishability, unique adapted stochastic processes with continuous sample paths satisfying the stochastic differential equation

$$dS_t^{X_d} = \sigma_d(S_t^{X_d})dB_t^d + \mu_d(S_t^{X_d})dt \quad \text{and} \quad S_0^{X_d} = X_d$$

$\mathbb{P}$-a.s. for every $t \in [0,T]$ (see, for instance, [2, Theorem 9.2]), define the label by

$$Y_d := \varphi_d(S_T^{X_d})$$

and for every $x \in \mathbb{R}^d$ let

$$(S_t^x)_{t\in[0,T]} \colon [0,T] \times \Omega \to \mathbb{R}^d$$

be the, up to indistinguishability, unique adapted stochastic processes with continuous sample paths satisfying the stochastic differential equation

$$dS_t^x = \sigma_d(S_t^x)dB_t^d + \mu_d(S_t^x)dt \quad \text{and} \quad S_0^x = x$$

$\mathbb{P}$-a.s. for every $t \in [0,T]$. We assume Setting 2.3 and we suppose that the initial values $(\varphi_d)_{d\in\mathbb{N}}$ can be approximated by neural networks in the following sense. Let $\mathbf{c} \in [1,\infty)$, $\nu \in [1/2,\infty)$, $\alpha, \beta, \gamma, \kappa, \lambda \in [0,\infty)$ and for every $d \in \mathbb{N}$, $\varepsilon \in (0,1)$ let

$$\mathbf{b}_{d,\varepsilon} \in \mathbf{A}_d, \quad \boldsymbol{\eta}_{d,\varepsilon} \in \mathbb{R}^{\mathcal{P}(\mathbf{b}_{d,\varepsilon})}$$

such that for all $d \in \mathbb{N}$, $\varepsilon \in (0,1)$ and $x \in \mathbb{R}^d$ it holds

(i) that

$$|\varphi_d(x) - \mathcal{F}_{\mathbf{b}_{d,\varepsilon}}(\boldsymbol{\eta}_{d,\varepsilon}, x)| \leq \mathbf{c}d^\alpha\varepsilon(1 + \|x\|_2^\nu),$$

(ii) that

$$\|\boldsymbol{\eta}_{d,\varepsilon}\|_\infty \leq \mathbf{c}d^\beta\varepsilon^{-\kappa},$$

(iii) and that

$$\mathcal{P}(\mathbf{b}_{d,\varepsilon}) \leq \mathbf{c}d^\gamma\varepsilon^{-\lambda}.$$

Finally, for all $d \in \mathbb{N}$ let $F_d \in C([0,T] \times \mathbb{R}^d, \mathbb{R})$ be the unique function satisfying

(i) that $F_d(0,x) = \varphi(x)$ for every $x \in \mathbb{R}^d$,

(ii) that F_d is at most polynomially growing, i.e. there exists $\vartheta \in (0,\infty)$ such that for every $x \in \mathbb{R}^d$ it holds that $\max_{t\in[0,T]} F_d(t,x) \leq \vartheta\left(1 + \|x\|_2^\vartheta\right)$,

(iii) and that F_d is a viscosity solution of the d-dimensional Kolmogorov PDE

$$\tfrac{\partial F_d}{\partial t}(t,x) = \tfrac{1}{2}\mathrm{Trace}\big(\sigma_d(x)[\sigma_d(x)]^*(\mathrm{Hess}_x F_d)(t,x)\big) + \big\langle \mu_d(x), (\nabla_x F_d)(t,x)\big\rangle_{\mathbb{R}^d}$$

for all $(t,x) \in (0,T)\times\mathbb{R}^d$,

see [23, Lemma 2.6 with $\varphi \leftarrow \sigma_d, \mu_d$ and Proposition 3.4(i) with $\varepsilon \leftarrow \mathbf{c}d^\alpha\varepsilon$, $c \leftarrow (\mathbf{c}+\mathfrak{D})d^\alpha$, $\mathbf{v},\mathbf{w} \leftarrow \nu$, $u \leftarrow F_d$].

The next result shows that computing the end value $[u,v]^d \ni x \mapsto F_d(T,x)$ of the solution to the Kolmogorov PDE can be restated as a learning problem.

Lemma 3.2 (Kolmogorov PDEs as Learning Problem)**.** *Assume Setting 3.1 and let $d \in \mathbb{N}$. Then for a.e. $x \in [u,v]^d$ it holds that*

$$F_d(T,x) = \widehat{f}_d(x).$$

Proof of Lemma 3.2. The proof is based on the Feynman-Kac formula for viscosity solutions of Kolmogorov equations in [23, Corollary 2.23(ii) with $u \leftarrow F_d$, $X_T \leftarrow S_T^x$] which assures that for every $x \in \mathbb{R}^d$ it holds that

$$F_d(T,x) = \mathbb{E}\big[\varphi_d\left(S_T^x\right)\big]. \tag{19}$$

We claim that for every $A \in \mathcal{B}([u,v]^d)$ it holds that

$$\mathbb{E}\left[\mathbf{1}_A(X_d)\varphi_d(S_T^{X_d})\right] = \int_A \mathbb{E}\left[\varphi_d(S_T^x)\right]\, d\mathbb{P}_{X_d}(x).$$

This would prove the lemma as it implies that for $\mathbb{P}_{X_d}$-a.s. $x \in [u,v]^d$ it holds that

$$\mathbb{E}\left[\varphi_d(S_T^{X_d})\Big| X_d = x\right] = \mathbb{E}\big[\varphi_d\left(S_T^x\right)\big]$$

and by (19) and the definition of $\widehat{f}_d$, Y_d, X_d this assures that for a.e. $x \in [u,v]^d$ it holds that

$$\widehat{f}_d(x) = \mathbb{E}\left[Y_d\middle| X_d = x\right] = \mathbb{E}\left[\varphi_d(S_T^{X_d})\Big| X_d = x\right] = \mathbb{E}\big[\varphi_d\left(S_T^x\right)\big] = F_d(T,x).$$

For the proof of the claim let us fix $A \in \mathcal{B}([u,v]^d)$, let $g_\varepsilon \in C^\infty(\mathbb{R}^d,\mathbb{R})$, $\varepsilon \in (0,1)$, be a family of mollifiers and for every $\varepsilon \in (0,1)$ define the convolution with the indicator function $\mathbf{1}_A$ by $\mathbf{1}_{A,\varepsilon} := \mathbf{1}_A * g_\varepsilon$. Then by the properties of a mollifier (see, for instance, [20, Appendix C.5 with $U \leftarrow \mathbb{R}^d$, $f^\epsilon \leftarrow \mathbf{1}_{A,\varepsilon}$]) it holds that $\mathbf{1}_{A,\varepsilon} \in C^\infty(\mathbb{R}^d,\mathbb{R})$ and

$$\lim_{\varepsilon\to 0} \mathbf{1}_{A,\varepsilon}(x) = \mathbf{1}_A(x) \tag{20}$$

for a.e. $x \in \mathbb{R}^d$. By defining the continuous and bounded mapping

$$\Phi_\varepsilon : \left\{ \begin{array}{ccc} C([0,T],\mathbb{R}^d) & \to & \mathbb{R} \\ f & \mapsto & \mathbf{1}_{A,\varepsilon}(f(0))\varphi(f(T)) \end{array} \right. ,$$

we obtain from [6, Lemma 2.6(v) with $\mathbb{X}_t \leftarrow S_t^{X_d}$, $X_t^x \leftarrow S_t^x$] that for every $\varepsilon \in (0,1)$ it holds that

$$\mathbb{E}\left[\mathbf{1}_{A,\varepsilon}(X_d)\varphi(S_T^{X_d})\right] = \frac{1}{(v-u)^d}\int_{[u,v]^d} \mathbf{1}_{A,\varepsilon}(x)\mathbb{E}\big[\varphi\left(S_T^x\right)\big]\, dx.$$

By (20), the dominated convergence theorem and the definition of $\mathbb{P}_{X_d}$ the claim follows when letting ε tend to zero. ☐

3.2 Neural Network Approximation Results for Solutions of Kolmogorov PDEs

In this subsection we prove the following approximation result.

Theorem 3.3 (Neural Network Regularity Result for Kolmogorov PDEs)**.** *Assume Setting 3.1 and let $\tau = \nu + 2\alpha$. Then there exist $C, c \in (0, \infty)$ such that for all $d \in \mathbb{N}$ and $\varepsilon \in (0,1)$ there is $\mathbf{a} \in \mathbf{A}_d$ and $\boldsymbol{\theta} \in \mathbb{R}^{\mathcal{P}(\mathbf{a})}$*

(i) with

$$\frac{1}{(v-u)^d} \|F_d(T,\cdot) - \mathcal{F}_{\mathbf{a},\mathfrak{D}}(\boldsymbol{\theta},\cdot)\|^2_{L^2[u,v]^d} \leq \varepsilon$$

(ii) with

$$\mathcal{P}(\mathbf{a}) \leq C d^{\tau(\lambda/2+2)+\gamma} \varepsilon^{-\lambda/2-2}$$

(iii) with

$$\|\boldsymbol{\theta}\|_\infty \leq C d^{\tau(\kappa/2+1)+\beta+3/2} \varepsilon^{-\kappa-1}$$

(iv) with

$$\mathcal{L}(\mathbf{a}) = \mathcal{L}(\mathbf{b}_{d,cd^{-\tau}\varepsilon^{1/2}})$$

(v) and with

$$\|\mathbf{a}\|_\infty \leq C d^\tau \varepsilon^{-1} \|\mathbf{b}_{d,cd^{-\tau}\varepsilon^{1/2}}\|_\infty.$$

Let us briefly sketch the idea of the proof. First we observe that in our case of affine linear drift μ_d and diffusion coefficient σ_d there exist random variables $\mathfrak{M}_d$ and $\mathfrak{N}_d$ such that for all $x \in \mathbb{R}^d$ it holds $\mathbb{P}$-a.s. that

$$S_T^x = \mathfrak{M}_d x + \mathfrak{N}_d,$$

see Lemma 3.4 below. Let

$$((\mathfrak{M}_d^{(j)}, \mathfrak{N}_d^{(j)}))_{j\in\mathbb{N}}$$

be i.i.d. samples with $(\mathfrak{M}_d^{(1)}, \mathfrak{N}_d^{(1)}) \sim (\mathfrak{M}_d, \mathfrak{N}_d)$. Then for fixed $x \in \mathbb{R}^d$ the Feynman-Kac formula, our assumptions, properties of Monte-Carlo approximation and a standard decomposition of the mean squared error into the sum of the squared bias and the variance yield that

$$\mathbb{E}\left[\left(F_d(T,x) - \frac{1}{n}\sum_{j=1}^n \mathcal{F}_{\mathbf{b}_{d,\delta}}\big(\boldsymbol{\eta}_{d,\delta}, \mathfrak{M}_d^{(j)} x + \mathfrak{N}_d^{(j)}\big)\right)^2\right] = \underbrace{\mathbb{E}\Big[\varphi_d(S_T^x) - \mathcal{F}_{\mathbf{b}_{d,\delta}}\left(\boldsymbol{\eta}_{d,\delta}, S_T^x\right)\Big]^2}_{\mathcal{O}(\delta^2)}$$
$$+ \underbrace{\mathbb{E}\left[\left(\mathbb{E}\big[\mathcal{F}_{\mathbf{b}_{d,\delta}}\left(\boldsymbol{\eta}_{d,\delta}, \mathfrak{M}_d x + \mathfrak{N}_d\right)\big] - \frac{1}{n}\sum_{j=1}^n \mathcal{F}_{\mathbf{b}_{d,\delta}}\big(\boldsymbol{\eta}_{d,\delta}, \mathfrak{M}_d^{(j)} x + \mathfrak{N}_d^{(j)}\big)\right)^2\right]}_{\mathcal{O}(n^{-1})}.$$

With more effort one can prove analogous estimates in the $L^2[u,v]^d$-norm and this suggests that, given ε, for sufficient large n and small δ there exists an outcome ω such that with

$$M_d^{(j)} := \mathfrak{M}_d^{(j)}(\omega),\ N_d^{(j)} := \mathfrak{N}_d^{(j)}(\omega)$$

it holds that

$$\frac{1}{(v-u)^d}\int_{[u,v]^d}\left(F_d(T,x) - \frac{1}{n}\sum_{j=1}^n \mathcal{F}_{\mathbf{b}_{d,\delta}}\big(\boldsymbol{\eta}_{d,\delta}, M_d^{(j)} x + N_d^{(j)}\big)\right)^2 dx \leq \varepsilon.$$

Finally we will prove in Lemma 3.6 that there exists a network architecture $\mathbf{a}$ and parameters $\boldsymbol{\theta}$ such that for every $x \in \mathbb{R}^d$ it holds that

$$\mathcal{F}_{\mathbf{a}}(\boldsymbol{\theta}, x) = \frac{1}{n} \sum_{j=1}^{n} \mathcal{F}_{\mathbf{b}_{d,\delta}}\big(\boldsymbol{\eta}_{d,\delta}, M_d^{(j)} x + N_d^{(j)}\big)$$

and we bound the parameters $\boldsymbol{\theta}$ with the help of Lemma 3.5 below. For a detailed presentation we refer the interested reader to [23]. The precise proof of Theorem 3.3 is based on the latter reference and is given after the following three auxiliary lemmas.

Lemma 3.4. *Assume Setting 3.1, let $d \in \mathbb{N}$, for every $i \in \{0, 1, \ldots, d\}$ let $e_i \in \mathbb{R}^d$ be the i-th standard basis vector in $\mathbb{R}^d$ and define the random variables $\mathfrak{M}_d : \Omega \to \mathbb{R}^{d \times d}$ and $\mathfrak{N}_d : \Omega \to \mathbb{R}^d$ by*

$$\mathfrak{N}_d := S_T^0, \quad \mathfrak{M}_d := \begin{bmatrix} S_T^{e_1} - S_T^0 & S_T^{e_2} - S_T^0 & \ldots & S_T^{e_d} - S_T^0 \end{bmatrix}.$$

Then for all $x \in \mathbb{R}^d$ it holds $\mathbb{P}$-a.s. that

$$S_T^x = \mathfrak{M}_d x + \mathfrak{N}_d.$$

Proof of Lemma 3.4. The proof is a simple consequence of [23, Lemma 2.15 with $X_T^x \leftarrow S_T^x$]. □

Lemma 3.5. *Assume Setting 3.1, let $d \in \mathbb{N}$ and let $\mathfrak{M}_d : \Omega \to \mathbb{R}^{d \times d}$, $\mathfrak{N}_d : \Omega \to \mathbb{R}^d$ as in Lemma 3.4. Then it holds that*

$$\mathbb{E}\big[\, \|\mathfrak{M}_d\|_2 + \|\mathfrak{N}_d\|_2 \,\big] \leq 3\sqrt{2} d \left(1 + LT + 2L\sqrt{T}\right) \exp\left(\big[L\sqrt{T} + 2L\big]^2 T\right).$$

Proof of Lemma 3.5. In [23, Proposition 2.14 with $p = 2$, $\xi \leftarrow x$, $\mathfrak{m}_1, \mathfrak{m}_2, \mathfrak{s}_1, \mathfrak{s}_2 \leftarrow L$, $X_T \leftarrow S_T^x$] it is shown that for all $x \in \mathbb{R}^d$ it holds that

$$\left(\mathbb{E}\left[\|S_T^x\|_2^2\right]\right)^{1/2} \leq \sqrt{2} \left(\|x\|_2 + LT + 2L\sqrt{T}\right) \exp\left(\big[L\sqrt{T} + 2L\big]^2 T\right). \tag{21}$$

The fact that

$$\mathbb{E}\big[\|S_T^x\|_2\big] \leq \left(\mathbb{E}\left[\|S_T^x\|_2^2\right]\right)^{1/2},$$

the triangle inequality, the subadditivity of the square root and (21) imply that

$$\begin{aligned}
&\mathbb{E}\big[\, \|\mathfrak{M}_d\|_2 + \|\mathfrak{N}_d\|_2 \,\big] = \mathbb{E}\Big[\, \big\| \begin{bmatrix} S_T^{e_1} - S_T^0 & S_T^{e_2} - S_T^0 & \ldots & S_T^{e_d} - S_T^0 \end{bmatrix} \big\|_2 + \big\|S_T^0\big\|_2 \Big] \\
&\leq \mathbb{E}\Big[\, \big\|S_T^0\big\|_2 + \sum_{i=1}^{d} \big\|S_T^{e_i} - S_T^0\big\|_2 \Big] \leq (d+1)\mathbb{E}\left[\big\|S_T^0\big\|_2\right] + \sum_{i=1}^{d} \mathbb{E}\left[\big\|S_T^{e_i}\big\|_2\right] \\
&\leq \sqrt{2}\Big[(d+1)\left(LT + 2L\sqrt{T}\right) + d\left(1 + LT + 2L\sqrt{T}\right)\Big] \exp\left(\big[L\sqrt{T} + 2L\big]^2 T\right) \\
&\leq 3\sqrt{2} d \left(1 + LT + 2L\sqrt{T}\right) \exp\left(\big[L\sqrt{T} + 2L\big]^2 T\right)
\end{aligned}$$

which is the desired estimate. □

In the next lemma we show that the average of the composition of a neural network with different affine functions can be represented by a single neural network and we bound the number and size of its parameters.

Lemma 3.6. *Assume Setting 2.3. Let $n \in \mathbb{N}$, $l \in \mathbb{N}_0$, $\mathbf{b} = (b_0, b_1, \ldots, b_l, b_{l+1}) \in \mathbb{N}^{l+2}$, $\boldsymbol{\eta} \in \mathbb{R}^{\mathcal{P}(\mathbf{b})}$, let*

$$((M^{(j)}, N^{(j)}))_{j=1}^{n} \in \left(\mathbb{R}^{b_0 \times b_0} \times \mathbb{R}^{b_0}\right)^n$$

and let

$$\mathbf{a} = (b_0, nb_1, \ldots, nb_l, b_{l+1}) \in \mathbb{N}^{l+2}.$$

Then it holds that

$$\mathcal{P}(\mathbf{a}) \leq n^2 \mathcal{P}(\mathbf{b})$$

and there exists $\boldsymbol{\theta} \in \mathbb{R}^{\mathcal{P}(\mathbf{a})}$

(i) with

$$\|\boldsymbol{\theta}\|_\infty \leq \sqrt{b_0}\|\boldsymbol{\eta}\|_\infty \max_{j=1}^{n} \left(\|M^{(j)}\|_2 + \|N^{(j)}\|_2 + 1\right)$$

(ii) and with

$$\mathcal{F}_{\mathbf{a}}(\boldsymbol{\theta}, x) = \frac{1}{n} \sum_{j=1}^{n} \mathcal{F}_{\mathbf{b}}(\boldsymbol{\eta}, M^{(j)} x + N^{(j)})$$

for all $x \in \mathbb{R}^{b_0}$.

Proof of Lemma 3.6. With the exception of Item (i) this result is proven in [23, Lemma 3.8 with $A_j \leftarrow M_j$, $b_j \leftarrow N_j$, $\mathcal{R}(\phi) \leftarrow \mathcal{F}_{\mathbf{b}}(\boldsymbol{\eta}, \cdot)$, $\mathcal{R}(\psi) \leftarrow \mathcal{F}_{\mathbf{a}}(\boldsymbol{\theta}, \cdot)$]. The latter reference shows that for

$$\boldsymbol{\eta} = ((V_i, A_i))_{i=0}^{l} \in \bigtimes_{i=0}^{l} \left(\mathbb{R}^{b_{i+1} \times b_i} \times \mathbb{R}^{b_{i+1}}\right) \simeq \mathbb{R}^{\mathcal{P}(\mathbf{b})}$$

and

$$\mathbf{a} = (a_0, a_1, \ldots, a_l, a_{l+1}) := (b_0, nb_1, \ldots, nb_l, b_{l+1}) \in \mathbb{N}^{l+2}$$

a suitable

$$\boldsymbol{\theta} = ((W_i, B_i))_{i=0}^{l} \in \bigtimes_{i=0}^{l} \left(\mathbb{R}^{a_{i+1} \times a_i} \times \mathbb{R}^{a_{i+1}}\right) \simeq \mathbb{R}^{\mathcal{P}(\mathbf{a})}$$

is given by

$$W_0 := \begin{bmatrix} V_0 M^{(1)} \\ V_0 M^{(2)} \\ \vdots \\ V_0 M^{(n)} \end{bmatrix}, \quad B_0 := \begin{bmatrix} V_0 N^{(1)} + A_0 \\ V_0 N^{(2)} + A_0 \\ \vdots \\ V_0 N^{(n)} + A_0 \end{bmatrix},$$

for $i \in \{1, 2, \ldots, l-1\}$ by

$$W_i := \begin{bmatrix} V_i & \mathbf{0} & \ldots & \mathbf{0} \\ \mathbf{0} & V_i & \ldots & \vdots \\ \vdots & \vdots & \ddots & \mathbf{0} \\ \mathbf{0} & \ldots & \mathbf{0} & V_i \end{bmatrix}, \quad B_i := \begin{bmatrix} A_i \\ A_i \\ \vdots \\ A_i \end{bmatrix},$$

and by

$$W_l := \begin{bmatrix} \frac{1}{n} V_l & \frac{1}{n} V_l & \ldots & \frac{1}{n} V_l \end{bmatrix}, \quad B_l := A_l.$$

Observe that using Hölder's inequality it holds that

$$\|W_0\|_\infty \leq \sqrt{b_0}\|V_0\|_\infty \max_{j=1}^{n} \|M^{(j)}\|_2 \leq \sqrt{b_0}\|\boldsymbol{\eta}\|_\infty \max_{j=1}^{n} \|M^{(j)}\|_2$$

and

$$\|B_0\|_\infty \leq \sqrt{b_0}\|V_0\|_\infty \max_{j=1}^n \|N^{(j)}\|_2 + \|A_0\|_\infty \leq \sqrt{b_0}\|\boldsymbol{\eta}\|_\infty \max_{j=1}^n \left(\|N^{(j)}\|_2 + 1\right).$$

Together with the facts that for every $i \in \{1, 2, \ldots, l\}$ it holds that

$$\|W_i\|_\infty \leq \|V_i\|_\infty \leq \|\boldsymbol{\eta}\|_\infty$$

and

$$\|B_i\|_\infty = \|A_i\|_\infty \leq \|\boldsymbol{\eta}\|_\infty$$

this proves the lemma. □

Now we are ready to prove Theorem 3.3.

Proof of Theorem 3.3. Except for Property (iii) a similar result was shown in [23, Corollary 3.13]. For this reason we will be brief in presenting the proof and refer to the results of [23] for further details. Recall that by Lemma 3.4 for every $d \in \mathbb{N}$ there exist random variables $\mathfrak{M}_d : \Omega \to \mathbb{R}^{d\times d}$, $\mathfrak{N}_d : \Omega \to \mathbb{R}^d$ such that for all $x \in \mathbb{R}^d$ it holds $\mathbb{P}$-a.s. that

$$S_T^x = \mathfrak{M}_d x + \mathfrak{N}_d = \mathcal{A}_{\mathfrak{M}_d,\mathfrak{N}_d}(x).$$

For all $d \in \mathbb{N}$, $\delta \in (0, 1)$ we define the function $f_{d,\delta} \in C(\mathbb{R}^d, \mathbb{R})$ by

$$f_{d,\delta}(x) = \mathcal{F}_{\mathbf{b}_{d,\delta}}(\boldsymbol{\eta}_{d,\delta}, x)$$

which satisfies for all $x \in \mathbb{R}^d$ that

$$|f_{d,\delta}(x)| \leq |f_{d,\delta}(x) - \varphi_d(x)| + |\varphi_d(x)| \leq (\mathbf{c} + \mathfrak{D})d^\alpha(1 + \|x\|_2^\nu).$$

We will now use [23, Proof of Proposition 3.4, display (217), (218), (224) and (229) with $\varepsilon \leftarrow \mathbf{c}d^\alpha\delta$, $c \leftarrow (\mathbf{c} + \mathfrak{D})d^\alpha$, $\phi \leftarrow f_{d,\delta}$, $\mathbf{v}, \mathbf{w} \leftarrow \nu$, $p \leftarrow 2$, $\nu \leftarrow \mathbb{P}_{X_d}$, $u \leftarrow F_d$, $\mathscr{A}_j \leftarrow \mathfrak{M}_d^{(j)}$, $\mathscr{B}_j \leftarrow \mathfrak{N}_d^{(j)}$] together with our assumptions in Setting 3.1 and the fact that, by an elementary calculation, for every $d \in \mathbb{N}$ it holds that

$$\int_{\mathbb{R}^d} \|x\|_2^{2\nu}\, d\mathbb{P}_{X_d}(x) \leq d^\nu \frac{v^{2\nu+1} - u^{2\nu+1}}{2\nu(v-u)}.$$

In particular, [23, Proof of Proposition 3.4] shows that there exists $\mathbf{C} \in [1, \infty)$ and for every $d \in \mathbb{N}$ there exist i.i.d. random variables

$$((\mathfrak{M}_d^{(j)}, \mathfrak{N}_d^{(j)}))_{j\in\mathbb{N}}$$

with $(\mathfrak{M}_d^{(1)}, \mathfrak{N}_d^{(1)}) \sim (\mathfrak{M}_d, \mathfrak{N}_d)$ such that for every $d, n \in \mathbb{N}$ and $\delta \in (0, 1)$ by defining the random variable

$$E_{d,\delta,n} := \frac{1}{(v-u)^{d/2}} \left\| F_d(T, \cdot) - \frac{1}{n}\sum_{j=1}^n f_{d,\delta} \circ \mathcal{A}_{\mathfrak{M}_d^{(j)},\mathfrak{N}_d^{(j)}} \right\|_{L^2[u,v]^d}$$

it holds that

$$\begin{aligned}
\mathbb{E}\left[E_{d,\delta,n}\right] &\leq \frac{1}{(v-u)^{d/2}} \left\| \mathbb{E}\left[\varphi_d \circ \mathcal{A}_{\mathfrak{M}_d^{(1)},\mathfrak{N}_d^{(1)}}\right] - \mathbb{E}\left[f_{d,\delta} \circ \mathcal{A}_{\mathfrak{M}_d^{(1)},\mathfrak{N}_d^{(1)}}\right] \right\|_{L^2[u,v]^d} \\
&\quad + \mathbb{E}\left[\frac{1}{(v-u)^{d/2}} \left\| \mathbb{E}\left[f_{d,\delta} \circ \mathcal{A}_{\mathfrak{M}_d^{(1)},\mathfrak{N}_d^{(1)}}\right] - \frac{1}{n}\sum_{j=1}^n f_{d,\delta} \circ \mathcal{A}_{\mathfrak{M}_d^{(j)},\mathfrak{N}_d^{(j)}} \right\|_{L^2[u,v]^d} \right] \\
&\leq \mathbf{C}d^{\nu/2+\alpha}\left(\delta + n^{-1/2}\right) = \mathbf{C}d^{\tau/2}\left(\delta + n^{-1/2}\right).
\end{aligned}$$

Let us fix $d \in \mathbb{N}$, $\varepsilon \in (0,1)$,

$$n \in \left[16\mathbf{C}^2 d^\tau \varepsilon^{-1}, 32\mathbf{C}^2 d^\tau \varepsilon^{-1}\right) \cap \mathbb{N} \tag{22}$$

and

$$\delta = (4\mathbf{C})^{-1} d^{-\tau/2} \varepsilon^{1/2}. \tag{23}$$

This implies that

$$\mathbb{E}\left[E_{d,\delta,n}\right] \leq \frac{\varepsilon^{1/2}}{2}. \tag{24}$$

Next, let

$$\mathfrak{C} = 6\sqrt{2}\left(1 + LT + 2L\sqrt{T}\right) \exp\left(\left[L\sqrt{T} + 2L\right]^2 T\right)$$

and let the random variables H and G be defined by

$$H := (\mathfrak{C}d)^{-1} \max_{j=1}^{n} \left(\left\|\mathfrak{M}_d^{(j)}\right\|_2 + \left\|\mathfrak{N}_d^{(j)}\right\|_2\right)$$

and

$$G := \varepsilon^{-1/2} E_{d,\delta,n} + n^{-1} H.$$

Lemma 3.5 establishes that it holds that

$$\mathbb{E}[H] \leq (\mathfrak{C}d)^{-1} \sum_{j=1}^{n} \mathbb{E}\left[\left\|\mathfrak{M}_d^{(j)}\right\|_2 + \left\|\mathfrak{N}_d^{(j)}\right\|_2\right] = n(\mathfrak{C}d)^{-1} \mathbb{E}\left[\, \|\mathfrak{M}_d\|_2 + \|\mathfrak{N}_d\|_2 \,\right] \leq \frac{n}{2}$$

and together with (24) this assures that

$$\mathbb{E}[|G|] \leq 1.$$

By [23, Proposition 3.3 with $X \leftarrow G$, $\varepsilon \leftarrow 1$] it follows that there exists $\omega \in \Omega$ such that by defining for all $j \in \{1, \ldots, n\}$

$$M_d^{(j)} := \mathfrak{M}_d^{(j)}(\omega), \quad N_d^{(j)} := \mathfrak{N}_d^{(j)}(\omega)$$

it holds that

$$E_{d,\delta,n}^2(\omega) = \frac{1}{(v-u)^d} \left\| F_d(T,\cdot) - \frac{1}{n} \sum_{j=1}^{n} f_{d,\delta} \circ \mathcal{A}_{M_d^{(j)}, N_d^{(j)}} \right\|_{L^2[u,v]^d}^2 \leq \varepsilon \tag{25}$$

and

$$H(\omega) = (\mathfrak{C}d)^{-1} \max_{j=1}^{n} \left(\|M_d^{(j)}\|_2 + \|N_d^{(j)}\|_2\right) \leq n.$$

Using (22) this implies that

$$\max_{j=1}^{n} \left(\|M_d^{(j)}\|_2 + \|N_d^{(j)}\|_2\right) \leq \mathfrak{C}dn \leq 32\mathfrak{C}\mathbf{C}^2 d^{\tau+1} \varepsilon^{-1}.$$

By Lemma 3.6, our assumptions and (22), (23) there is $\mathbf{a} \in \mathbf{A}_d$ and $\boldsymbol{\theta} \in \mathbb{R}^{\mathcal{P}(\mathbf{a})}$ with

$$\mathcal{F}_{\mathbf{a}}\left(\boldsymbol{\theta}, x\right) = \frac{1}{n} \sum_{j=1}^{n} \mathcal{F}_{\mathbf{b}_{d,\delta}}\left(\boldsymbol{\eta}_{d,\delta}, M_d^{(j)} x + N_d^{(j)}\right) = \frac{1}{n} \sum_{j=1}^{n} f_{d,\delta} \circ \mathcal{A}_{M_d^{(j)}, N_d^{(j)}}(x) \tag{26}$$

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

$$\mathcal{P}(\mathbf{a}) \leq n^2 \mathcal{P}(\mathbf{b}_{d,\delta}) \leq 32^2 \mathbf{C}^4 \mathbf{c} d^{2\tau+\gamma} \varepsilon^{-2} \delta^{-\lambda} \leq C d^{\tau(\lambda/2+2)+\gamma} \varepsilon^{-\lambda/2-2},$$

$$\|\boldsymbol{\theta}\|_\infty \leq \sqrt{d}\|\boldsymbol{\eta}_{d,\delta}\|_\infty \left(32\mathfrak{C}\mathbf{C}^2 d^{\tau+1}\varepsilon^{-1}+1\right) \leq C d^{\tau(\kappa/2+1)+\beta+3/2}\epsilon^{-\kappa/2-1},$$
$$\mathcal{L}(\mathbf{a}) = \mathcal{L}(\mathbf{b}_{d,\delta}) = \mathcal{L}(\mathbf{b}_{d,cd^{-\tau/2}\varepsilon^{1/2}})$$

and with

$$\|\mathbf{a}\|_\infty = n\|\mathbf{b}_{d,\delta}\|_\infty \leq 32\mathbf{C}^2 d^\tau \varepsilon^{-1}\|\mathbf{b}_{d,\delta}\|_\infty \leq C d^\tau \varepsilon^{-1}\|\mathbf{b}_{d,cd^{-\tau/2}\varepsilon^{1/2}}\|_\infty$$

where $C, c \in (0,\infty)$ are defined by

$$C = \max\left\{32^2 4^\lambda \mathbf{C}^{4+\lambda}\mathbf{c}, (4\mathbf{C})^\kappa \mathbf{c}(32\mathfrak{C}\mathbf{C}^2+1)\right\}, \quad c = (4\mathbf{C})^{-1}.$$

Further, observe that, due to the Feynman-Kac formula in [23, Corollary 2.23(ii) with $u \leftarrow F_d$, $X_T \leftarrow S_T^x$] and the fact that $\varphi_d : \mathbb{R}^d \to [-\mathfrak{D}, \mathfrak{D}]$ for every $x \in \mathbb{R}^d$ it holds that

$$|F_d(T,x)| = \left|\mathbb{E}\left[\varphi_d(S_T^x)\right]\right| \leq \mathfrak{D},$$

which implies that

$$\mathcal{C}_{\mathfrak{D},1} \circ F_d(T,x) = F_d(T,x)$$

for all $x \in \mathbb{R}^d$. Corollary 2.5, (25) and (26) hence imply that

$$\frac{1}{(v-u)^d}\left\|F_d(T,\cdot) - \mathcal{F}_{\mathbf{a},\mathfrak{D}}(\boldsymbol{\theta},\cdot)\right\|^2_{L^2[u,v]^d} \leq \varepsilon$$

and this proves the theorem. □

3.3 Neural Network Generalization Results for Solutions of Kolmogorov PDEs

The next theorem represents the main result of this paper.

Theorem 3.7 (Neural Network Generalization Result for Kolmogorov PDEs)**.** *Assume Setting 3.1, let $\tau = \nu + 2\alpha$ and let $h \in C((0,\infty)^5, \mathbb{R})$ be given by*

$$h(x) = 128\mathfrak{D}^4 x_1^2\Big[1 + x_2 + x_4\Big(\ln\left(64\mathfrak{D}\max\{1,|u|,|v|\}x_1\right) + \left(x_5+1\right)\left(x_3+2\right)\Big)\Big].$$

Then there exist $C, c \in (0,\infty)$ such that for all $d \in \mathbb{N}$ and $\varepsilon, \varrho \in (0,1)$ there is $\mathbf{a} \in \mathbf{A}_d$ and $R \in [1,\infty)$

(i) with

$$\mathcal{P}(\mathbf{a}) \leq C d^{\tau(\lambda/2+2)+\gamma}\varepsilon^{-\lambda/2-2}$$

(ii) with

$$R \leq C d^{\tau(\kappa/2+1)+\beta+3/2}\varepsilon^{-\kappa/2-1}$$

(iii) with

$$\mathcal{L}(\mathbf{a}) = \mathcal{L}(\mathbf{b}_{d,cd^{-\tau}\varepsilon^{1/2}})$$

(iv) and with

$$\|\mathbf{a}\|_\infty \leq C d^\tau \varepsilon^{-1}\|\mathbf{b}_{d,cd^{-\tau}\varepsilon^{1/2}}\|_\infty$$

such that with

$$\mathcal{H} = \mathcal{N}^{u,v}_{\mathbf{a},R,\mathfrak{D}},$$

and

$$m \geq h\left(2\varepsilon^{-1}, \ln(\varrho^{-1}), \ln\left(R\|\mathbf{a}\|_\infty\right), \mathcal{P}(\mathbf{a}), \mathcal{L}(\mathbf{a})\right)$$

it holds that

$$\mathbb{P}\left[\frac{1}{(v-u)^d}\left\|\widehat{f}_{d,m,\mathcal{H}} - F_d(T,\cdot)\right\|^2_{L^2[u,v]^d} \leq \varepsilon\right] \geq 1-\varrho.$$

Proof of Theorem 3.7. This is a direct consequence of Theorem 3.3 (with $\varepsilon \leftarrow \varepsilon/2$), Corollary 2.11 and the fact that for every $d \in \mathbb{N}$ it holds that $\mathbb{P}_{X_d} = \frac{1}{(v-u)^d}\lambda_{[u,v]^d}$. □

We can also reformulate this in a more compact form.

Corollary 3.8. *Assume Setting 3.1 Then there exists a polynomial $p\colon \mathbb{R}^2 \to \mathbb{R}$ such that for all $d \in \mathbb{N}$ and $\varepsilon, \varrho \in (0,1)$ there is $\mathbf{a} \in \mathbf{A}_d$ and $R \in [1,\infty)$ with*

$$\max\{R, \mathcal{P}(\mathbf{a})\} \leq p(d, \varepsilon^{-1})$$

such that with

$$\mathcal{H} = \mathcal{N}^{u,v}_{\mathbf{a},R,\mathfrak{D}}$$

and

$$m \geq p(d,\varepsilon^{-1})(1 + \ln(\varrho^{-1}))$$

it holds that

$$\mathbb{P}\left[\frac{1}{(v-u)^d}\left\|\widehat{f}_{d,m,\mathcal{H}} - F_d(T,\cdot)\right\|^2_{L^2[u,v]^d} \leq \varepsilon\right] \geq 1 - \varrho.$$

Proof of Corollary 3.8. This is a direct consequence of Theorem 3.3, Corollary 2.12 and the fact that given arbitrary polynomials $q\colon \mathbb{R}^2 \to \mathbb{R}$ and $s\colon \mathbb{R}^2 \to \mathbb{R}$ there exists a new polynomial $p\colon \mathbb{R}^2 \to \mathbb{R}$ satisfying $\max\{q(x), s(x)\} \leq p(x)$ for every $x \in \mathbb{R}^2$. □

3.4 Pricing of High-Dimensional Options

The proof of Theorem 1.1 from the introductory section dealing with the pricing of high-dimensional European Put Options is now an easy consequence of the above theory.

Proof of Theorem 1.1. We first show that the approximation of $(\varphi_d)_{d\in\mathbb{N}}$ by clipped neural networks according to Setting 3.1 is possible. Note that for every $z \in [0,\infty)$ it holds that

$$\min\{z, \mathfrak{D}\} = \tfrac{1}{2}\left(\max\{\mathfrak{D}+z, 0\} - \max\{\mathfrak{D}-z,0\} - \max\{z-\mathfrak{D},0\}\right).$$

That implies that for every $d \in \mathbb{N}$ it holds that

$$\begin{aligned}\varphi_d(x) &= \min\left\{\max\left\{\mathfrak{D} - \sum_{i=1}^d c_{d,i}x_i, 0\right\}, \mathfrak{D}\right\} \\ &= \mathcal{A}_{V_2,A_2} \circ \mathrm{ReLU}_3 \circ \mathcal{A}_{V_1,A_1} \circ \mathrm{ReLU}_1 \circ A_{V_0,A_0}(x) = \mathcal{F}_{\mathbf{b}_d}(\boldsymbol{\eta}_d, x)\end{aligned}$$

where

$$V_0 = \begin{bmatrix}-c_{d,1} & -c_{d,2} & \dots & -c_{d,d}\end{bmatrix},\ A_0 = \mathfrak{D},\ V_1 = \begin{bmatrix}1\\-1\\1\end{bmatrix},\ A_1 = \begin{bmatrix}\mathfrak{D}\\ \mathfrak{D}\\ -\mathfrak{D}\end{bmatrix},$$

$$V_2 = \begin{bmatrix}\frac{1}{2} & -\frac{1}{2} & -\frac{1}{2}\end{bmatrix},\ A_1 = 0$$

$\mathbf{b}_d = (d,1,3,1)$ and

$$\boldsymbol{\eta}_d = (V_i, A_i)_{i=0}^2 \in \left(\mathbb{R}^{1\times d}\times\mathbb{R}^1\right)\times\left(\mathbb{R}^{3\times 1}\times\mathbb{R}^3\right)\times\left(\mathbb{R}^{1\times 3}\times\mathbb{R}^1\right).$$

Accordingly Setting 3.1 is satisfied with

$$\mathbf{c} = \max\{\mathfrak{D}, 11\},\ \nu = 1/2,\ \gamma = 1,\ \alpha = \beta = \kappa = \lambda = 0,\ \mathbf{b}_{d,\varepsilon} = \mathbf{b}_d,\ \boldsymbol{\eta}_{d,\varepsilon} = \boldsymbol{\eta}_d$$

for every $d \in \mathbb{N}$, $\varepsilon \in (0,1)$. Now Theorem 3.7 shows that there exists $C \in [1,\infty)$ such that for all $d \in \mathbb{N}$ and $\varepsilon, \varrho \in (0,1)$ there is $\mathbf{a} \in \mathbf{A}_d$ and $R \in [1,\infty)$

(i) with
$$\mathcal{P}(\mathbf{a}) \leq Cd^2\varepsilon^{-2}$$
(ii) with
$$R \leq Cd^2\varepsilon^{-1}$$
(iii) with
$$\mathcal{L}(\mathbf{a}) = \mathcal{L}(\mathbf{b}_d) = 2$$
(iv) and with
$$\|\mathbf{a}\|_\infty \leq Cd^{1/2}\varepsilon^{-1}\|\mathbf{b}_d\|_\infty = Cd^{3/2}\varepsilon^{-1}$$
such that with
$$\mathcal{H} = \mathcal{N}^{u,v}_{\mathbf{a},R,\mathfrak{D}} \tag{27}$$
and
$$m = \lceil h\left(2\varepsilon^{-1}, \ln(\varrho^{-1}), \ln\left(C^2d^{7/2}\varepsilon^{-2}\right), Cd^2\varepsilon^{-2}, 2\right)\rceil$$
it holds that
$$\mathbb{P}\left[\frac{1}{(v-u)^d}\left\|\widehat{f}_{d,m,\mathcal{H}} - F_d(T,\cdot)\right\|^2_{L^2[u,v]^d} \leq \varepsilon\right] \geq 1-\varrho.$$
A simple calculation shows that by defining
$$\mathbf{C} = 512\mathfrak{D}^4C\left[7 + \ln\left(128\mathfrak{D}C^6\max\{1,|u|,|v|\}\right)\right] + 1$$
for every $d \in \mathbb{N}$, $\varepsilon, \varrho \in (0,1)$ it holds that
$$\lceil h\left(2\varepsilon^{-1}, \ln(\varrho^{-1}), \ln\left(C^2d^4\varepsilon^{-2}\right), Cd^2\varepsilon^{-2}, 2\right)\rceil \leq \mathbf{C}d^2\varepsilon^{-4}\big(1 + \ln(d\varepsilon^{-1}\varrho^{-1})\big)$$
and this concludes the proof. □

4 Covering Number Estimates

In this section we prove some estimates on the covering numbers of neural network hypothesis classes. We will use the following setting.

Setting 4.1. *Let $(\mathcal{H}, d)$ be a compact metric space, for $r \in (0,\infty)$ and $f \in \mathcal{H}$ let*
$$B_r(f) := \{g \in \mathcal{H} : \ d(f,g) \leq r\}$$
be the ball of radius r around f*, for $r \in (0,\infty)$ let*
$$N(\mathcal{H}, r) := \inf\left\{n \in \mathbb{N} : \ \text{There exists } (f_i)_{i=1}^n \subseteq \mathcal{H} \ \text{ with } \ \mathcal{H} \subseteq \bigcup_{i=1}^n B_r(f_i)\right\} < \infty$$
*be the r-*covering number of $\mathcal{H}$*.*

Let us first prove that neural networks with fixed architecture and bounded parameters are Lipschitz continuous with respect to their weights and biases.

Theorem 4.2. *Assume Setting 2.3, let $u \in \mathbb{R}$, $v \in (u,\infty)$, $R \in [1,\infty)$, $l \in \mathbb{N}_0$, let*
$$\mathfrak{m} = \max\{1, |u|, |v|\}$$
and let $\mathbf{a} = (a_0, a_1, a_2, \ldots, a_l, a_{l+1}) \in \mathbb{N}^{l+2}$. Then for every $\boldsymbol{\theta}, \boldsymbol{\eta} \in \mathbb{R}^{\mathcal{P}(\mathbf{a})}$ with
$$\max\{\|\boldsymbol{\theta}\|_\infty, \|\boldsymbol{\eta}\|_\infty\} \leq R$$
it holds that
$$\sup_{x \in [u,v]^{a_0}} \left\|\mathcal{F}_{\mathbf{a}}(\boldsymbol{\theta}, x) - \mathcal{F}_{\mathbf{a}}(\boldsymbol{\eta}, x)\right\|_\infty \leq \|\boldsymbol{\theta} - \boldsymbol{\eta}\|_\infty \frac{\mathfrak{m}R^l(3\|\mathbf{a}\|_\infty + 3)^{l+1}}{2}.$$

Proof of Theorem 4.2. Let us fix $\boldsymbol{\theta}, \boldsymbol{\eta} \in \mathbb{R}^{\mathcal{P}(\mathbf{a})}$ with $\max\{\|\boldsymbol{\theta}\|_\infty, \|\boldsymbol{\eta}\|_\infty\} \leq R$ given by

$$\boldsymbol{\theta} = \big((W_i, B_i)\big)_{i=0}^{l} \in \bigtimes_{i=0}^{l} \big(\mathbb{R}^{a_{i+1}\times a_i} \times \mathbb{R}^{a_{i+1}}\big) \simeq \mathbb{R}^{\mathcal{P}(\mathbf{a})}$$

and

$$\boldsymbol{\eta} = \big((V_i, A_i)\big)_{i=0}^{l} \in \bigtimes_{i=0}^{l} \big(\mathbb{R}^{a_{i+1}\times a_i} \times \mathbb{R}^{a_{i+1}}\big) \simeq \mathbb{R}^{\mathcal{P}(\mathbf{a})}.$$

To simplify the notation we define $D := [u, v]^{a_0}$ and for every $s \in \{1, \dots, l+1\}$ we define the partial architecture $\mathbf{a}(s)$ by

$$\mathbf{a}(s) = (a_0, a_1, \dots, a_s) \in \mathbb{R}^{s+1},$$

the partial parameters $\boldsymbol{\theta}(s)$, $\boldsymbol{\eta}(s)$ by

$$\boldsymbol{\theta}(s) = \big((W_i, B_i)\big)_{i=0}^{s-1} \in \bigtimes_{i=0}^{s-1} \big(\mathbb{R}^{a_{i+1}\times a_i} \times \mathbb{R}^{a_{i+1}}\big) \simeq \mathbb{R}^{\mathcal{P}(\mathbf{a}(s))}$$

and

$$\boldsymbol{\eta}(s) = \big((V_i, A_i)\big)_{i=0}^{s-1} \in \bigtimes_{i=0}^{s-1} \big(\mathbb{R}^{a_{i+1}\times a_i} \times \mathbb{R}^{a_{i+1}}\big) \simeq \mathbb{R}^{\mathcal{P}(\mathbf{a}(s))},$$

the partial networks f_s, g_s by

$$f_s : \left\{ \begin{array}{ccc} D & \to & \mathbb{R}^{a_s} \\ x & \mapsto & \mathcal{F}_{\mathbf{a}(s)}(\boldsymbol{\theta}(s), x) \end{array} \right. ,$$

and

$$g_s : \left\{ \begin{array}{ccc} D & \to & \mathbb{R}^{a_s} \\ x & \mapsto & \mathcal{F}_{\mathbf{a}(s)}(\boldsymbol{\eta}(s), x) \end{array} \right. ,$$

the partial errors $\mathfrak{e}_s$ by

$$\mathfrak{e}_s = \sup_{x\in D} \big\|f_s(x) - g_s(x)\big\|_\infty$$

and the partial maxima $\mathfrak{m}_s$ by

$$\mathfrak{m}_s = \max\left\{1, \sup_{x\in D} \big\|f_s(x)\big\|_\infty, \sup_{x\in D} \big\|g_s(x)\big\|_\infty\right\}.$$

Note that it holds that

$$f_{l+1}(x) = \mathcal{F}_{\mathbf{a}}(\boldsymbol{\theta}, x)$$

and

$$g_{l+1}(x) = \mathcal{F}_{\mathbf{a}}(\boldsymbol{\eta}, x)$$

for every $x \in D$ and we are therefore interested in estimating the error e_{l+1}. For the convenience of the reader we further define

$$\mathfrak{e}_0 = 0,\ \mathfrak{m}_0 = \mathfrak{m},\ r = \|\boldsymbol{\theta} - \boldsymbol{\eta}\|_\infty.$$

We will try to bound $\mathfrak{m}_{s+1}$ relative to $\mathfrak{m}_s$ and $\mathfrak{e}_{s+1}$ relative to $\mathfrak{e}_s$ by observing how the error and the maximum is amplified in each layer. Using the triangle and Hölder's inequality it holds that

$$\begin{aligned}
&\sup_{x\in D} \big\|f_{s+1}(x)\big\|_\infty = \sup_{x\in D} \Big\|W_s \operatorname{ReLU}_{a_s}\big(f_s(x)\big) + B_s\Big\|_\infty \\
&\leq \sup_{x\in D} a_s \big\|W_s\big\|_\infty \big\|\operatorname{ReLU}_{a_s}\big(f_s(x)\big)\big\|_\infty + \big\|B_s\big\|_\infty \\
&\leq R\|\mathbf{a}\|_\infty \sup_{x\in D} \big\|f_s(x)\big\|_\infty + R \leq R\mathfrak{m}_s\big(\|\mathbf{a}\|_\infty + 1\big).
\end{aligned}$$

An analogous computation for the parameters $\boldsymbol{\eta}$ and the case $s = 0$ shows that for every $s \in \{0, 1, \dots, l\}$ it holds that

$$\mathfrak{m}_{s+1} \leq R\mathfrak{m}_s\big(\|\mathbf{a}\|_\infty + 1\big)$$

and thus also

$$\mathfrak{m}_{s+1} \leq R^{s+1}\mathfrak{m}\big(\|\mathbf{a}\|_\infty + 1\big)^{s+1}. \tag{28}$$

For estimating the partial error we will make use of the following basic inequality. For every $k, n \in \mathbb{N}$, $M_1, M_2 \in \mathbb{R}^{k\times n}$, $N_1, N_2 \in \mathbb{R}^{n\times 1}$ it holds that

$$\begin{aligned}\|M_1N_1 - M_2N_2\|_\infty &= \|(M_1 - M_2)(N_1 - N_2) + M_2(N_1 - N_2) + (M_1 - M_2)N_2\|_\infty \\ &\leq n\big(\|M_1 - M_2\|_\infty\|N_1 - N_2\|_\infty + \|M_2\|_\infty\|N_1 - N_2\|_\infty + \|M_1 - M_2\|_\infty\|N_2\|_\infty\big).\end{aligned}$$

This assures that for every $s \in \{1, \dots, l\}$ and $x \in D$ it holds that

$$\begin{aligned}&\Big\|\Big(W_s\,\mathrm{ReLU}_{a_s}\big(f_s(x)\big) + B_s\Big) - \Big(V_s\,\mathrm{ReLU}_{a_s}\big(g_s(x)\big) + A_s\Big)\Big\|_\infty \\ &\leq a_s\Big(\big\|W_s - V_s\big\|_\infty\big\|\,\mathrm{ReLU}_{a_s}\big(f_s(x)\big) - \mathrm{ReLU}_{a_s}\big(g_s(x)\big)\big\|_\infty \\ &+ \big\|V_s\big\|_\infty\big\|\,\mathrm{ReLU}_{a_s}\big(f_s(x)\big) - \mathrm{ReLU}_{a_s}\big(g_s(x)\big)\big\|_\infty \\ &+ \big\|W_s - V_s\big\|_\infty\big\|\,\mathrm{ReLU}_{a_s}\big(g_s(x)\big)\big\|_\infty\Big) + \big\|B_s - A_s\big\|_\infty \\ &\leq \|\mathbf{a}\|_\infty\big(r\mathfrak{e}_s + R\mathfrak{e}_s + \mathfrak{m}_s r\big) + r \leq \big(\|\mathbf{a}\|_\infty + 1\big)\big(r\mathfrak{e}_s + R\mathfrak{e}_s + \mathfrak{m}_s r\big).\end{aligned} \tag{29}$$

Together with the fact that $r \leq 2R$ this establishes that for every $s \in \{1, 2, \dots, l\}$ it holds that

$$\mathfrak{e}_{s+1} \leq \big(\|\mathbf{a}\|_\infty + 1\big)\big(3R\mathfrak{e}_s + \mathfrak{m}_s r\big). \tag{30}$$

We now claim that for every $s \in \{0, 1, \dots, l\}$ it holds that

$$\mathfrak{e}_{s+1} \leq \tfrac{1}{2}\big(3^{s+1} - \tfrac{1}{2}\big)R^s\mathfrak{m}\big(\|\mathbf{a}\|_\infty + 1\big)^{s+1}r, \tag{31}$$

which we will prove by induction. The base case $s = 0$ is proved similar to (29), namely

$$\begin{aligned}\mathfrak{e}_1 &= \sup_{x\in D}\big\|\big(W_0x + B_0\big) - \big(V_0x + A_0\big)\big\|_\infty \\ &\leq \|\mathbf{a}\|_\infty\mathfrak{m}r + r \leq \mathfrak{m}\big(\|\mathbf{a}\|_\infty + 1\big)r.\end{aligned}$$

For the induction step let us assume that (31) holds for a given $s \in \{0, 1, \dots, l-1\}$, which implies by (28) and (30) that

$$\begin{aligned}\mathfrak{e}_{s+2} &\leq \big(\|\mathbf{a}\|_\infty + 1\big)\big(3R\mathfrak{e}_{s+1} + \mathfrak{m}_{s+1}r\big) \\ &= \tfrac{1}{2}\big(\|\mathbf{a}\|_\infty + 1\big)\Big(\big(3^{s+2} - \tfrac{3}{2}\big)R^{s+1}\mathfrak{m}\big(\|\mathbf{a}\|_\infty + 1\big)^{s+1}r + R^{s+1}\mathfrak{m}\big(\|\mathbf{a}\|_\infty + 1\big)^{s+1}r\Big) \\ &\leq \tfrac{1}{2}\big(3^{s+2} - \tfrac{1}{2}\big)R^{s+1}\mathfrak{m}\big(\|\mathbf{a}\|_\infty + 1\big)^{s+2}r.\end{aligned}$$

Consequently, our claim (31) holds for every $s \in \{0, 1, \dots, l\}$ and in particular assures that

$$\begin{aligned}\sup_{x\in D}\big\|\mathcal{F}_{\mathbf{a}}(\boldsymbol{\theta}, x) - \mathcal{F}_{\mathbf{a}}(\boldsymbol{\eta}, x)\big\|_\infty = e_{l+1} &\leq \tfrac{1}{2}\big(3^{l+1} - \tfrac{1}{2}\big)R^l\mathfrak{m}\big(\|\mathbf{a}\|_\infty + 1\big)^{l+1}r \\ &\leq r\frac{\mathfrak{m}R^l\big(3\|\mathbf{a}\|_\infty + 3\big)^{l+1}}{2}\end{aligned}$$

This proves the theorem. □

Next we state a proposition on the covering number of balls in an euclidean space (or general any finite-dimensional Banach space) and with Theorem 4.2 this allows us to bound the covering number of our hypothesis class.

Proposition 4.3. *Assume Setting 4.1. Let $n \in \mathbb{N}$, $R \in [1, \infty)$, $r \in (0, 1)$ and define the metric space*

$$B_R = \big\{\boldsymbol{\theta} \in \mathbb{R}^n : \|\boldsymbol{\theta}\|_\infty \leq R\big\}$$

with its metric induced by the norm $\|\cdot\|_\infty$. Then it holds that

$$\ln N(B_R, r) \leq n \ln\Big(\frac{4R}{r}\Big).$$

Proof of Proposition 4.3. A proof is given, for instance, in [13, Proposition 5]. □

This implies the following lemma for the covering number of the hypothesis class of (clipped) neural networks.

Lemma 4.4. *Assume Setting 2.3 and 4.1, let $u \in \mathbb{R}$, $v \in (u, \infty)$, $d \in \mathbb{N}$, $R \in [1, \infty)$, $r \in (0, 1)$, let $\mathfrak{m} = \max\big\{1, |u|, |v|\big\}$ and let $\mathbf{a} \in \mathbf{A}_d$. Then it holds that*

$$\ln N\big(\mathcal{N}^{u,v}_{\mathbf{a},R,\mathfrak{D}}, r\big) \leq \ln N\big(\mathcal{N}^{u,v}_{\mathbf{a},R}, r\big) \leq \mathcal{P}(\mathbf{a})\Big[\ln\Big(\frac{2\mathfrak{m}}{r}\Big) + (\mathcal{L}(\mathbf{a}) + 1)\ln\big(3R\|\mathbf{a}\|_\infty + 3R\big)\Big].$$

Proof of Lemma 4.4. To simplify the notation we define

$$B_R = \big\{\boldsymbol{\theta} \in \mathbb{R}^{\mathcal{P}(\mathbf{a})} : \|\boldsymbol{\theta}\|_\infty \leq R\big\},\ \Delta = r\frac{2}{\mathfrak{m}R^{\mathcal{L}(\mathbf{a})}\big(3\|\mathbf{a}\|_\infty + 3\big)^{\mathcal{L}(\mathbf{a})+1}},\ N = N\Big(B_R, \Delta\Big).$$

Choose

$$\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \ldots, \boldsymbol{\theta}_N \in B_R$$

such that the balls with center $\boldsymbol{\theta}_i$, $i \in \{1, 2, \ldots, N\}$, and radius Δ cover B_R. For arbitrary $\boldsymbol{\theta} \in B_R$ there exists $i \in \{1, 2, \ldots, N\}$ such that it holds that

$$\|\boldsymbol{\theta} - \boldsymbol{\theta}_i\|_\infty \leq \Delta$$

and by Lemma 2.6 and Theorem 4.2 this implies that

$$\begin{aligned}\sup_{x \in [u,v]^{a_0}} \big|\mathcal{F}_{\mathbf{a},\mathfrak{D}}(\boldsymbol{\theta}, x) - \mathcal{F}_{\mathbf{a},\mathfrak{D}}(\boldsymbol{\theta}_i, x)\big| &\leq \sup_{x \in [u,v]^{a_0}} \big|\mathcal{F}_{\mathbf{a}}(\boldsymbol{\theta}, x) - \mathcal{F}_{\mathbf{a}}(\boldsymbol{\theta}_i, x)\big| \\ &\leq \|\boldsymbol{\theta} - \boldsymbol{\theta}_i\|_\infty \frac{\mathfrak{m}R^{\mathcal{L}(\mathbf{a})}\big(3\|\mathbf{a}\|_\infty + 3\big)^{\mathcal{L}(\mathbf{a})+1}}{2} \leq r\end{aligned}$$

This shows that

$$N\big(\mathcal{N}^{u,v}_{\mathbf{a},R,\mathfrak{D}}, r\big) \leq N\big(\mathcal{N}^{u,v}_{\mathbf{a},R}, r\big) \leq N = N\Big(B_R, \Delta\Big)$$

and Proposition 4.3 implies that it holds that

$$\begin{aligned}\ln N\big(\mathcal{N}^{u,v}_{\mathbf{a},R,\mathfrak{D}}, r\big) &\leq \ln N\big(\mathcal{N}^{u,v}_{\mathbf{a},R}, r\big) \leq \mathcal{P}(\mathbf{a})\ln\Big(\frac{4R}{\Delta}\Big) \\ &= \mathcal{P}(\mathbf{a})\Big[\ln\Big(\frac{2\mathfrak{m}}{r}\Big) + (\mathcal{L}(\mathbf{a}) + 1)\ln\big(3R\|\mathbf{a}\|_\infty + 3R\big)\Big].\end{aligned}$$

This proves the lemma. □

5 Proof of Theorem 2.10

In this section we prove our main generalization result. As a main tool we will use the following result.

Theorem 5.1 (Bound on the Sample Error)**.** *Assume Settings 2.1, 2.3 and 4.1. Let* $\varepsilon > 0$*,* $d \in \mathbb{N}$*,* $R \in [1, \infty)$*,* $\mathbf{a} \in \mathbf{A}_d$*,* $\mathcal{H} := \mathcal{N}^{u,v}_{\mathbf{a},R,\mathfrak{D}}$ *and* $m \in \mathbb{N}$*. Then it holds that*

$$\mathbb{P}\left[\mathcal{E}_d\left(\widehat{f}_{d,m,\mathcal{H}}\right) - \mathcal{E}_d\left(\widehat{f}_{d,\mathcal{H}}\right) \leq \varepsilon\right] \geq 1 - 2N\Big(\mathcal{H}, \frac{\varepsilon}{32\mathfrak{D}}\Big) \exp\left(-\frac{m\varepsilon^2}{128\mathfrak{D}^4}\right).$$

Proof of Theorem 5.1. The proof is adapted for our purposes from [13] and [41, End of Chapter 3]. First note that by assumptions for every $f \in \mathcal{H}$ it holds that

$$|f(X_d) - Y_d| \leq \sup_{x \in [u,v]^d} |f(x)| + |Y_d| \leq 2\mathfrak{D}$$

and analogously for the samples $((X_d^{(i)}, Y_d^{(i)}))_{i=1}^m$. The elementary identity

$$(y_1 - z)^2 - (y_2 - z)^2 = (y_1 - y_2)(y_1 + y_2 - 2z)$$

for real numbers $y_1, y_2, z \in \mathbb{R}$ and Jensen's inequality imply that

$$\begin{aligned} |\mathcal{E}_d(f) - \mathcal{E}_d(g)| &\leq \mathbb{E}\left[\left|\big(f(X_d) - g(X_d)\big)\big(f(X_d) + g(X_d) - 2Y_d\big)\right|\right] \\ &\leq 4\mathfrak{D} \sup_{x \in [u,v]^d} |f(x) - g(x)| \end{aligned}$$

and

$$\begin{aligned} |\mathcal{E}_{d,m}(f) - \mathcal{E}_{d,m}(g)| &\leq \frac{1}{m}\sum_{i=1}^m \left|\big(f(X_d^{(i)}) - g(X_d^{(i)})\big)\big(f(X_d^{(i)}) + g(X_d^{(i)}) - 2Y_d^{(i)}\big)\right| \\ &\leq 4\mathfrak{D} \sup_{x \in [u,v]^d} |f(x) - g(x)| \end{aligned}$$

for every $f, g \in \mathcal{H}$. Now define $N = N\big(\mathcal{H}, \frac{\varepsilon}{32\mathfrak{D}}\big)$ and choose

$$f_1, f_2, \ldots, f_N \in \mathcal{H}$$

such that the balls

$$B_i = \left\{ f \in \mathcal{H} \colon \sup_{x \in [u,v]^d} |f(x) - f_i(x)| \leq \frac{\varepsilon}{32\mathfrak{D}} \right\}, \ i \in \{1, 2, \ldots, N\},$$

cover $\mathcal{H}$. This establishes that for every $i \in \{1, 2, \ldots, N\}$ and $f \in B_i$ it holds that

$$\begin{aligned} |\mathcal{E}_d(f) - \mathcal{E}_{d,m}(f)| &\leq |\mathcal{E}_d(f) - \mathcal{E}_d(f_i)| + |\mathcal{E}_d(f_i) - \mathcal{E}_{d,m}(f_i)| + |\mathcal{E}_{d,m}(f_i) - \mathcal{E}_{d,m}(f)| \\ &\leq 8\mathfrak{D} \sup_{x \in [u,v]^d} |f(x) - f_i(x)| + |\mathcal{E}_d(f_i) - \mathcal{E}_{d,m}(f_i)| \\ &\leq \varepsilon/4 + |\mathcal{E}_d(f_i) - \mathcal{E}_{d,m}(f_i)|. \end{aligned} \tag{32}$$

Our assumptions yield that for every $\omega \in \Omega$ it holds that

$$\begin{aligned} &\mathcal{E}_d\left(\widehat{f}_{d,m,\mathcal{H}(\omega)}\right) - \mathcal{E}_d\left(\widehat{f}_{d,\mathcal{H}}\right) \\ &\leq \mathcal{E}_d\left(\widehat{f}_{d,m,\mathcal{H}(\omega)}\right) - \mathcal{E}_{d,m}\left(\widehat{f}_{d,m,\mathcal{H}(\omega)}\right)(\omega) + \mathcal{E}_{d,m}\left(\widehat{f}_{d,\mathcal{H}}\right)(\omega) - \mathcal{E}_d\left(\widehat{f}_{d,\mathcal{H}}\right) \\ &\leq 2 \sup_{f \in \mathcal{H}} |\mathcal{E}_d(f) - \mathcal{E}_{d,m}(f)(\omega)| \end{aligned} \tag{33}$$

In summary (32) and (33) give

$$\begin{aligned}
&\left\{\omega \in \Omega\colon \mathcal{E}_d\left(\widehat{f}_{d,m,\mathcal{H}}(\omega)\right) - \mathcal{E}_d\left(\widehat{f}_{d,\mathcal{H}}\right) \geq \varepsilon\right\} \\
&\subseteq \bigcup_{i=1}^{N} \left\{\omega \in \Omega\colon \sup_{f\in B_i} |\mathcal{E}_d(f) - \mathcal{E}_{d,m}(f)(\omega)| \geq \varepsilon/2\right\} \\
&\subseteq \bigcup_{i=1}^{N} \left\{\omega \in \Omega\colon |\mathcal{E}_d(f_i) - \mathcal{E}_{d,m}(f_i)(\omega)| \geq \varepsilon/4\right\}.
\end{aligned}$$

Observe that for fixed $f \in \mathcal{H}$ it holds that the random variables $E_i := \left(f(X_d^{(i)}) - Y_d^{(i)}\right)^2$, $i \in \{1, 2, \ldots, m\}$, are independent and satisfy

$$\mathbb{E}\left[E_i\right] = \mathcal{E}_d(f),\ \frac{1}{m}\sum_{i=1}^{m} E_i = \mathcal{E}_{d,m}(f),\ 0 \leq E_i \leq 4\mathfrak{D}^2$$

which by Hoeffding's inequality (see, for instance, [29, Theorem 2 with $X_i \leftarrow \pm E_i$]) assures that

$$\mathbb{P}\left[\left|\mathcal{E}_d(f) - \mathcal{E}_{d,m}(f)\right| \geq \varepsilon/4\right] \leq 2\exp\left(-\frac{m\varepsilon^2}{128\mathfrak{D}^4}\right). \tag{34}$$

Together with (5), the monotonicity and subadditivity of the probability measure and the measurability assumptions according to Lemma 2.2 this proves that

$$\begin{aligned}
\mathbb{P}\left[\mathcal{E}_d\left(\widehat{f}_{d,m,\mathcal{H}}\right) - \mathcal{E}_d\left(\widehat{f}_{d,\mathcal{H}}\right) \geq \varepsilon\right] &\leq \sum_{i=1}^{N} \mathbb{P}\left[\left|\mathcal{E}_d(f_i) - \mathcal{E}_{d,m}(f_i)\right| \geq \varepsilon/4\right] \\
&\leq 2N\exp\left(-\frac{m\varepsilon^2}{128\mathfrak{D}^4}\right).
\end{aligned}$$

Using the complement rule and plugging in the definition of N proves the theorem. □

We now have everything in place to proceed to the proof of Theorem 2.10.

Proof of Theorem 2.10. This is a direct consequence of Theorem 5.1 and Lemma 4.4. We assume Setting 4.1 and and observe that

$$1 - 2N\Big(\mathcal{H}, \frac{\varepsilon}{32\mathfrak{D}}\Big)\exp\left(-\frac{m\varepsilon^2}{128\mathfrak{D}^4}\right) \geq 1 - \varrho \tag{35}$$

holds for every

$$m \geq 128\mathfrak{D}^4\varepsilon^{-2}\left[\ln N\Big(\mathcal{H}, \frac{\varepsilon}{32\mathfrak{D}}\Big) + \ln(2/\varrho)\right]. \tag{36}$$

Lemma 4.4 and some basic inequalities assure that with $\mathfrak{m} = \max\{1, |u|, |v|\}$ it holds that

$$\begin{aligned}
&128\mathfrak{D}^4\varepsilon^{-2}\left[\ln N\Big(\mathcal{H}, \frac{\varepsilon}{32\mathfrak{D}}\Big) + \ln(2/\varrho)\right] \\
&\leq 128\mathfrak{D}^4\varepsilon^{-2}\left[\mathcal{P}(\mathbf{a})\Big(\ln\Big(\frac{64\mathfrak{D}\mathfrak{m}}{\varepsilon}\Big) + \big(\mathcal{L}(\mathbf{a})+1\big)\ln\big(3R\|\mathbf{a}\|_\infty + 3R\big)\Big) + \ln(2/\varrho)\right] \\
&\leq 128\mathfrak{D}^4\varepsilon^{-2}\Big[\mathcal{P}(\mathbf{a})\Big(\ln\Big(\frac{64\mathfrak{D}\mathfrak{m}}{\varepsilon}\Big) + \big(\mathcal{L}(\mathbf{a})+1\big)\big(\ln(R\|\mathbf{a}\|_\infty) + 2\big)\Big) + \ln(\varrho^{-1}) + 1\Big] \\
&= h\left(\varepsilon^{-1}, \ln(\varrho^{-1}), \ln(R\|\mathbf{a}\|_\infty), \mathcal{P}(\mathbf{a}), \mathcal{L}(\mathbf{a})\right).
\end{aligned} \tag{37}$$

Combining (35), (36), (37) and Theorem 5.1 shows that for every

$$m \geq h\left(\varepsilon^{-1}, \ln(\varrho^{-1}), \ln(R\|\mathbf{a}\|_\infty), \mathcal{P}(\mathbf{a}), \mathcal{L}(\mathbf{a})\right)$$

it holds that

$$\mathbb{P}\left[\mathcal{E}_d\left(\widehat{f}_{d,m,\mathcal{H}}\right) - \mathcal{E}_d\left(\widehat{f}_{d,\mathcal{H}}\right) \leq \varepsilon\right] \geq 1 - 2N\Big(\mathcal{H}, \frac{\varepsilon}{32\mathfrak{D}}\Big)\exp\left(-\frac{m\varepsilon^2}{128\mathfrak{D}^4}\right) \geq 1 - \varrho$$

and this proves the theorem. □

Acknowledgements

The authors are grateful to Shahar Mendelson and Stefan Steinerberger for their useful comments. The research of JB was supported by the Austrian Science Fund (FWF) under grant I3403-N32.

References

[1] Aliprantis, C., and Border, K. *Infinite Dimensional Analysis: A Hitchhiker's Guide (third edition)*. Springer, 2007.

[2] Baldi, P. *Stochastic Calculus: An Introduction Through Theory and Exercises*. Universitext. Springer International Publishing, 2017.

[3] Barron, A. R. Universal approximation bounds for superpositions of a sigmoidal function. *IEEE Trans. Inf. Theory 39*, 3 (1993), 930–945.

[4] Barron, A. R. Approximation and estimation bounds for artificial neural networks. *Mach. Learn. 14*, 1 (1994), 115–133.

[5] Bartlett, P. L., Bousquet, O., Mendelson, S., et al. Local rademacher complexities. *The Annals of Statistics 33*, 4 (2005), 1497–1537.

[6] Beck, C., Becker, S., Grohs, P., Jaafari, N., and Jentzen, A. Solving stochastic differential equations and Kolmogorov equations by means of deep learning. *arXiv:1806.00421* (2018).

[7] Beck, C., E, W., and Jentzen, A. Machine learning approximation algorithms for high-dimensional fully nonlinear partial differential equations and second-order backward stochastic differential equations. *arXiv:1709.05963* (2017).

[8] Bellman, R. *Dynamic Programming*. Dover Books on Computer Science Series. Dover Publications, 2003.

[9] Bölcskei, H., Grohs, P., Kutyniok, G., and Petersen, P. Optimal approximation with sparsely connected deep neural networks. *arXiv:1705.01714* (2017).

[10] Burger, M., and Neubauer, A. Error Bounds for Approximation with Neural Networks . *Journal of Approximation Theory 112*, 2 (2001), 235–250.

[11] Candès, E. J. Ridgelets: Theory and Applications, 1998. Ph.D. thesis, Stanford University.

[12] Chui, C. K., Li, X., and Mhaskar, H. N. Neural networks for localized approximation. *Math. Comp. 63*, 208 (1994), 607–623.

[13] Cucker, F., and Smale, S. On the mathematical foundations of learning. *Bulletin of the American mathematical society 39*, 1 (2002), 1–49.

[14] Cybenko, G. Approximation by superpositions of a sigmoidal function. *Mathematics of Control, Signals and Systems 2*, 4 (1989), 303–314.

[15] DeVore, R., Oskolkov, K., and Petrushev, P. Approximation by feed-forward neural networks. *Ann. Numer. Math. 4* (1996), 261–287.

[16] E, W., Han, J., and Jentzen, A. Deep learning-based numerical methods for high-dimensional parabolic partial differential equations and backward stochastic differential equations. *arXiv:1706.04702* (2017).

[17] E, W., and Yu, B. The Deep Ritz method: A deep learning-based numerical algorithm for solving variational problems. *arXiv:1710.00211* (2017).

[18] Eldan, R., and Shamir, O. The power of depth for feedforward neural networks. *Proceedings of the 29th Conference on Learning Theory, COLT 2016, New York, USA* (2016), 907–940.

[19] Ellacott, S. Aspects of the numerical analysis of neural networks. *Acta Numer. 3* (1994), 145–202.

[20] Evans, L. *Partial Differential Equations (second edition).* Graduate studies in mathematics. American Mathematical Society, 2010.

[21] Fujii, M., Takahashi, A., and Takahashi, M. Asymptotic Expansion as Prior Knowledge in Deep Learning Method for high dimensional BSDEs. *arXiv:1710.07030* (2017).

[22] Funahashi, K.-I. On the approximate realization of continuous mappings by neural networks. *Neural Networks 2*, 3 (1989), 183–192.

[23] Grohs, P., Hornung, F., Jentzen, A., and von Wurstemberger, P. A proof that artificial neural networks overcome the curse of dimensionality in the numerical approximation of black-scholes partial differential equations. *arXiv-submit:2386269* (2018).

[24] Györfi, L., Kohler, M., Krzyzak, A., and Walk, H. *A distribution-free theory of nonparametric regression.* Springer Science & Business Media, 2006.

[25] Hairer, M., Hutzenthaler, M., and Jentzen, A. Loss of regularity for Kolmogorov equations. *Ann. Probab. 43*, 2 (2015), 468–527.

[26] Han, J., Jentzen, A., and E, W. Overcoming the curse of dimensionality: Solving high-dimensional partial differential equations using deep learning. *arXiv:1707.02568* (2017).

[27] Henry-Labordere, P. Deep Primal-Dual Algorithm for BSDEs: Applications of Machine Learning to CVA and IM. Available at SSRN: https://ssrn.com/abstract=3071506.

[28] Hinton, G., Deng, L., Yu, D., Dahl, G. E., r. Mohamed, A., Jaitly, N., Senior, A., Vanhoucke, V., Nguyen, P., Sainath, T. N., and Kingsbury, B. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *IEEE Signal Processing Magazine 29*, 6 (2012), 82–97.

[29] Hoeffding, W. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association 58*, 301 (1963), 13–30.

[30] Hornik, K. Approximation capabilities of multilayer feedforward networks. *Neural Networks 4*, 2 (1991), 251 – 257.

[31] Hornik, K., Stinchcombe, M., and White, H. Multilayer feedforward networks are universal approximators. *Neural Networks 2*, 5 (1989), 359–366.

[32] Khoo, Y., Lu, J., and Ying, L. Solving parametric PDE problems with artificial neural networks. *arXiv:1707.03351* (2017).

[33] Kingma, D. P., and Ba, J. Adam: A method for stochastic optimization. *arXiv:1412.6980* (2014).

[34] Koltchinskii, V. Introduction. In *Oracle Inequalities in Empirical Risk Minimization and Sparse Recovery Problems.* Springer, 2011, pp. 1–16.

[35] LeCun, Y., Bengio, Y., and Hinton, G. Deep learning. *nature 521*, 7553 (2015), 436.

[36] LeCun, Y., Cortes, C., and Burges, C. J. C. The MNIST database of handwritten digits, 1998. http://yann.lecun.com/exdb/mnist/ [online; accessed August 22, 2018].

[37] Massart, P. *Concentration inequalities and model selection*. Springer, 2007.

[38] Mhaskar, H., and Micchelli, C. Degree of approximation by neural and translation networks with a single hidden layer. *Adv. Appl. Math. 16*, 2 (1995), 151–183.

[39] Mhaskar, H. N. Neural networks for optimal approximation of smooth and analytic functions. *Neural Comput. 8*, 1 (1996), 164–177.

[40] Mhaskar, H. N., and Poggio, T. Deep vs. shallow networks: An approximation theory perspective. *Analysis and Applications 14*, 06 (2016), 829–848.

[41] Mohri, M., Rostamizadeh, A., Talwalkar, A., and Bach, F. *Foundations of Machine Learning*. Adaptive computation and machine learning series. MIT Press, 2012.

[42] Nguyen-Thien, T., and Tran-Cong, T. Approximation of functions and their derivatives: A neural network implementation with applications. *Appl. Math. Model. 23*, 9 (1999), 687–704.

[43] Nielsen, M. Neural networks and deep learning, 2015. http://neuralnetworksanddeeplearning.com/chap1.html [online; accessed March 05, 2018].

[44] Perekrestenko, D., Grohs, P., Elbrächter, D., and Bölcskei, H. The universal approximation power of finite-width deep relu networks. *arXiv:1806.01528* (2018).

[45] Petersen, P., and Voigtlaender, F. Optimal approximation of piecewise smooth functions using deep ReLU neural networks. *arXiv:1709.05289* (2017).

[46] Pinkus, A. Approximation theory of the MLP model in neural networks. *Acta Numer. 8* (1999), 143–195.

[47] Schmitt, M. Lower bounds on the complexity of approximating continuous functions by sigmoidal neural networks. In *Proceedings of the 12th International Conference on Neural Information Processing Systems* (Cambridge, MA, USA, 1999), NIPS'99, MIT Press, pp. 328–334.

[48] Shaham, U., Cloninger, A., and Coifman, R. R. Provable approximation properties for deep neural networks. *Applied and Computational Harmonic Analysis 44*, 3 (2018), 537 – 557.

[49] Sirignano, J., and Spiliopoulos, K. DGM: A deep learning algorithm for solving partial differential equations. *arXiv:1708.07469* (2017).

[50] Van de Geer, S. A. Applications of empirical process theory, volume 6 of cambridge series in statistical and probabilistic mathematics, 2000.

[51] Yarotsky, D. Error bounds for approximations with deep relu networks. *Neural Networks 94* (2017), 103–114.

[52] Yarotsky, D. Universal approximations of invariant maps by neural networks. *arXiv:1804.10306* (2018).

[53] Zhang, C., Bengio, S., Hardt, M., Recht, B., and Vinyals, O. Understanding deep learning requires rethinking generalization. *arXiv:1611.03530* (2016).