ON ACCURACY BARRIERS AND ERROR-CONTROLLING PARAMETERS IN NEURAL NETWORKS

Anonymous authors
Paper under double-blind review

ABSTRACT

Many advances have improved our understanding of the depth, width, or number of parameters of a neural network (NN) that are needed to achieve a certain accuracy, and also our understanding of related topics such as overparameterization. Nevertheless, our theoretical understanding of these topics is incomplete, and perspectives with elementary principles can be valuable. A main aim here is to show, using simple concepts, that a relatively small number of parameters in the NN can be identified as the error-controlling parameters. Other parameters, which are more numerous, are identified as playing the role of defining the basis functions or fundamental features. Taken together, the error decays at multiple rates, where a rapid decay in error is governed by the error-controlling parameters, and subsequently a barrier or a slower decay in error can be seen. Theoretical results are proved for many common function classes (e.g., continuously differentiable functions, analytic functions, piecewise smooth functions), when approximated by a continuous piecewise linear function. Empirical results demonstrate these aspects in ReLU neural networks with standard training, including both synthetic examples and the MNIST dataset. Owing to a finite amount of training data, a second error barrier can arise, in addition to the representation error barrier mentioned above. We relate this data error barrier to a Monte Carlo numerical integration error, arising from the optimization problem for training.

1 INTRODUCTION

In understanding the approximation error of artificial neural networks (NNs), many interesting results have advanced our understanding. An early type of result was universal approximation theorems (e.g., Cybenko, 1989; Pinkus, 1999; Kidger & Lyons, 2020) that give limited information about the width and/or depth, which are considered to be arbitrarily large. In recent years, in other types of results, many studies have proven relationships between the size of an NN (e.g., the depth, width, and/or number of parameters) and its accuracy (e.g., Yarotsky, 2017; Yarotsky & Zhevnerchuk, 2020; DeVore et al., 2021).

In addition, for our understanding of training, many insightful studies have added to our understanding of optimization methods, such as stochastic gradient descent, as well as related phenomena such as overparameterization (e.g., Jacot et al., 2018; Belkin et al., 2019; Du et al., 2019; Rahaman et al., 2019; Bartlett et al., 2020; Hastie et al., 2022; Loog et al., 2020; Schaeffer et al., 2024).

While these insights and theoretical advances are significant, our understanding of these topics is incomplete, and perspectives with elementary principles can be valuable.

1.1 OUR CONTRIBUTION

Here we show, using simple concepts, that a relatively small number of parameters in an NN can be identified as the error-controlling parameters. From this perspective, multiple rates of error decay are identified. A rapid decay in error is associated with the error-controlling parameters, and a slower decay in error is associated with the other parameters. Transitioning from the rapid decay to the slower decay is like hitting an accuracy barrier.

These results are shown both theoretically and empirically, in section 2. Theoretical results are shown for approximating functions from many common function classes (such as continuously differentiable

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

functions, analytic functions, or discontinuous piecewise smooth functions), when approximated by continuous piecewise linear functions. As is also common in theoretical studies, we consider the domain of the function (i.e., the input space) to be the d-dimensional hypercube, such as $[0, 2\pi]^d$ or $[-1, 1]^d$. Empirical results are shown for neural networks with standard training, including synthetic examples and also the MNIST dataset (LeCun et al., 1998).

Furthermore, a second accuracy barrier also arises, due to a finite amount of training data. We relate this data error barrier to a Monte Carlo numerical integration error, arising from the optimization problem for training. This data error barrier is described in section 3, and a concluding discussion is in section 4.

Taken together, a scaling law is presented for the error, in equation 8, with different terms associated with different error barriers. In this simplified perspective on NN scaling, the different contributions to the error are expressed in terms of the finiteness in representation, parameters, and training data.

1.2 Related work

For understanding training, many perspectives have added to our understanding. For example, phenomena such as low-frequency spectral bias have been identified (e.g., Rahaman et al., 2019), and theorems on the convergence of randomly initialized gradient descent have been proved (e.g., Du et al., 2019). Examining low frequencies and high frequencies has been valuable in the past (e.g., Rahaman et al., 2019), and our work herein will also utilize a frequency-by-frequency viewpoint. Our perspective here highlights further aspects and is different in many ways, including a theoretical perspective on the training error, by relating it to Monte Carlo numerical integration error, arising from the optimization problem for training. Also, here we consider not only training error but also representation error.

In the setting of theoretical work on NN accuracy and error, many interesting results have been proved on the relationship between the size of an NN and its accuracy (e.g., Yarotsky (2017); Yarotsky & Zhevnerchuk (2020); DeVore et al. (2021)). Here our aim is not to get a better bound or sharper bound. Instead, we aim for a more refined view of accuracy and its dependence on NN size. It is more refined in the sense that we identify the presence of two error decay rates. A consequence of the multiple decay rates is that there are also multiple sets of parameters: each set of parameters is in control of a different rate of error decay.

Some past theoretical proofs on NN accuracy are related in the sense of considering piecewise-linear approximations of basis functions such as polynomials or trigonometric polynomials (sines and cosines) (e.g., Yarotsky (2017); Elbrächter et al. (2021) using concepts from Telgarsky (2015)). To an extent, some of the concepts herein are implicitly immersed in such proofs. However, we are not aware of any clear or prominent statements or clear or prominent illustrations of the main concepts herein on error-controlling parameters, accuracy barriers, and multiple rates of error decay. To the authors' knowledge, the results presented herein have not been presented in past literature.

In other past work, results on multiple convergence rates of NN approximations have been presented as a phase diagram of convergence rates (Yarotsky & Zhevnerchuk, 2020). We note that, in that phase diagram, it is a relationship between a convergence rate and the smoothness of a function. The convergence rate is related to a function's smoothness, and different convergence rates were identified upon considering different amounts of function smoothness. In contrast, here, the multiple rates of error decay are identified in the approximation of a single function.

Lastly, note that a single scaling law is typically sought for the relationship between NN size and accuracy (e.g., Yarotsky (2017); Yarotsky & Zhevnerchuk (2020); DeVore et al. (2021)). Here, we highlight that a more refined viewpoint can be taken, and multiple rates of error decay can be identified.

2 A first accuracy barrier and error-controlling parameters

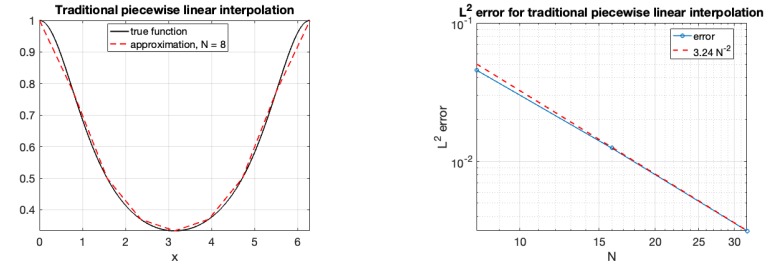
In this section, we illustrate the concepts of the first accuracy barrier and the error-controlling parameters. The role of a neural network (NN) activation function enters through the particular types of functions that are considered in section 2. For example, continuous piecewise-linear functions will be considered for their connection with NNs using ReLU activation functions (e.g., Arora et al.

(2018)). In what follows, section 2.1 presents the theoretical results without referring to the details of NN training, beginning with a 1D example and its error analysis, followed by extensions to other activation functions, basis functions, function classes, and general dimension d. Section 2.2 presents empirical results for trained NNs that illustrate the theoretical findings.

2.1 THEORY

We study approximation by ReLU neural networks through piecewise linear approximation, motivated by the fact that ReLU networks produce piecewise linear functions. We begin with a simple 1D setting by considering the approximation of an analytic function $f : \Omega \to \mathbb{R}$ on $\Omega = [0, 2\pi]^d$, where $d = 1$ and Ω is uniformly partitioned into a grid with N subintervals.

Classical approximation theory suggests that local piecewise linear approximation on this grid achieves algebraic accuracy, typically of order $\mathcal{O}(N^{-2})$. Figure 1 illustrates this traditional intuition.



(a) Piecewise linear approximation (b) Error vs N, with $N = 8, 16, 32$

Figure 1: Classical viewpoint of local piecewise linear approximation, illustrated using the target function $f(x) = \frac{1}{2-\cos x}$. Panel (a) shows the approximation for a fixed number of segments $N = 8$. Panel (b) shows the approximation error for different values of $N = 8, 16, 32$; the red dashed curve shows the theoretical upper bounds. This compares with the accuracy barrier that appears in Figure 3 upon considering piecewise linear approximations that are built from a different procedure.

We now show that a different behavior emerges when one considers piecewise linear approximations that are built from a different procedure. As an alternative procedure, rather than defining a local approximation, consider a global approximation, as follows. Let

$$\phi_1(x) = 1, \qquad \phi_{2m}(x) = \cos(mx), \qquad \phi_{2m+1}(x) = \sin(mx), \qquad m \geq 1,$$

denote the Fourier basis functions on Ω. For a given $K \leq N$, we take the first K modes and interpolate them onto the uniform grid with N subintervals using piecewise linear functions. More precisely, let

$$\Phi_j := I_N \phi_j, \qquad j = 1, \ldots, K, \tag{1}$$

where I_N denotes the piecewise linear interpolation operator on the grid. Figure 2 illustrates examples of these piecewise linear mode functions.

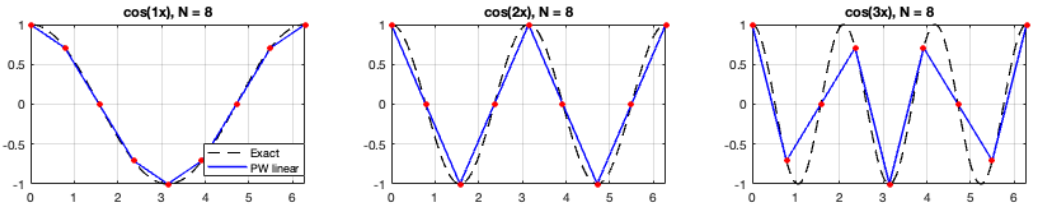


Figure 2: Piecewise linear approximation of Fourier basis.

To create a global approximation, we then take linear combinations of the piecewise versions of sines and cosines from equation 1 and Figure 2. The corresponding approximation space is then defined as

$$W_{K,N} := \text{span}\{\Phi_1, \Phi_2, \ldots, \Phi_K\}.$$

We then define a global approximation to f by its best $L^2(\Omega)$ approximation in $W_{K,N}$, namely

$$f_{K,N} := \operatorname*{argmin}_{v \in W_{K,N}} \|f - v\|_{L^2(\Omega)}. \tag{2}$$

Figure 3 shows the accuracy barrier that becomes visible in the global version of piecewise linear approximation, using the basis functions from equation 1 and Figure 2. The classical local approximation and the global construction both have an $\mathcal{O}(N^{-2})$ dependence on the grid resolution. However, for fixed N (number of piecewise segments), the global construction reveals two distinct error regimes as K increases: an initial regime of rapid decay, followed by a saturation regime. The rapid decay is consistent with the exponential decay of the Fourier coefficients of an analytic function, and this small number of parameters is in control of the error. On the other hand, the saturation regime is an accuracy barrier, beyond which the error no longer improves significantly as K increases. Consequently, although the approximant remains piecewise linear on the same mesh, the error does not follow a single algebraic law when global approximations are considered. Figure 8(a) in the Appendix also shows a representative global piecewise linear approximation at a fixed grid size.

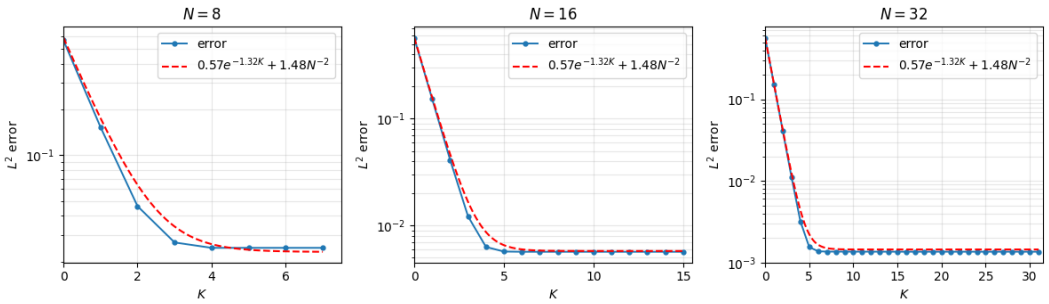


Figure 3: Accuracy barrier and error-controlling parameters in a global version of piecewise linear approximation, using the basis functions from equation 1 and Figure 2. The plots show the approximation error $\|f - f_{K,N}\|_2$ for the target function $f(x) = \frac{1}{2-\cos(x)}$, as the number of modes K increases for a fixed grid size N. Each panel shows a different value of $N = 8, 16, 32$, and the red dashed curves show the theoretical upper bound in equation 3. The error exhibits two regimes: an initial rapid decay as K increases, until an N-dependent accuracy barrier is reached.

The following theorem quantifies the approximation error and separates it into two contributions: the term e^{-cK}, which is the spectral truncation error arising from the analytic regularity of f, and the term N^{-2}, which is the floor or barrier imposed by piecewise linear approximation on a grid of fixed resolution.

Theorem 1. *Let $f : \Omega \to \mathbb{R}$ be periodic and analytic on $\Omega = [0, 2\pi]$. Then the best approximation $f_{K,N} \in W_{K,N}$ defined in equation 2 satisfies*

$$\|f - f_{K,N}\|_{L^2(\Omega)} \lesssim e^{-cK} + N^{-2}, \tag{3}$$

where $c > 0$ depends on the analyticity of f, and the implicit constant is independent of K and N.

A detailed proof and extensions of this theorem are given in the Appendix. Once the approximation reaches the grid-dependent floor, additional parameters may still be present in the representation, but they no longer lead to meaningful accuracy gains. In this sense, only a subset of the available degrees of freedom is actively contributing to error reduction. This suggests that a substantial fraction of the weights in a ReLU network may not be actively involved in improving accuracy within a finite resolution.

We are not aware of this phenomenon appearing before in the literature. It is a vast literature that has considered approximations of functions using Fourier series and related approximations, so perhaps it has appeared in some field before. In any case, if it has appeared in the literature before, it does not seem to be widely known. Perhaps there is a reason it has not drawn attention in the past: in many applications, one would simply use the sine and cosine functions themselves, rather than considering piecewise-linear approximations of sine and cosine. There are some theoretical studies of NNs in the past, which have considered piecewise-linear approximations of polynomials or sines and cosines

(e.g., Yarotsky (2017); Elbrächter et al. (2021) using concepts from Telgarsky (2015)). However, the concepts of error-controlling parameters and accuracy barriers are not emphasized in these past studies, and they are immersed inside the proofs. Here, we show that the concept of error-controlling parameters can provide a more refined viewpoint of accuracy, including the roles of multiple decay rates of error.

Remark 1. For each fixed grid size N, we denote by K_* the crossover point at which the two terms in the error bound are comparable. The point exists for all sufficiently large N and grows as $K_* \asymp \log N$ for analytic functions. For small N, the finite-resolution term may dominate, and the spectral-decay regime may be very short or not visible. When $K_* > 0$, the bound has two regimes. For $K < K_*$, the exponential truncation term dominates, and K is the main error-controlling parameter. For $K \gg K_*$, the finite-resolution term dominates, and N controls the resulting representation error floor. See Remark 4 in the Appendix for further details.

The approximation extends from the Fourier basis above to other types of bases, such as Legendre, Chebyshev, etc. On a grid of size N, each basis function can be approximated using a piecewise constant, piecewise linear, or higher-order piecewise polynomial approximation. Using the first K modes, we construct the best approximation to the target function as a linear combination of these piecewise basis functions. Figure 11 in the Appendix shows similar two-regime error behavior for piecewise linear Legendre modes. This construction does not impose additional assumptions on the target function beyond those required by the underlying exact basis. For example, Fourier bases naturally require periodic target functions to achieve spectral convergence, whereas Legendre or Chebyshev bases can be used for nonperiodic functions on finite domains.

Next, we consider other common types of target functions beyond the analytic setting, such as functions that are m times continuously differentiable, say $f \in C^m(\Omega)$, or functions with finite Sobolev regularity, say $f \in H^m(\Omega)$. The results here admit straightforward generalizations to these cases. More specifically, one still expects similar two-regime behavior, but the initial error decay is generally algebraic rather than exponential, with a rate depending on the regularity of f. See Remark 2 in the Appendix for details and Figures 14–15 for examples with C^1 target functions.

For another common class of functions, such as piecewise-smooth functions with a finite number of discontinuities or jumps in function values, a single global spectral approximation typically does not produce exponential initial convergence. Instead, the convergence is usually algebraic. In such cases, it is natural to use an adaptive approximation strategy, where the grid is chosen so that element boundaries align with the nonsmooth points or interfaces. Then the target can be smooth on each subdomain, and local basis functions, such as Legendre polynomial bases, can be used on those smooth subdomains. On each smooth domain, the local error can preserve the same two-regime behavior: a basis-controlled decay regime followed by a mesh-controlled error floor. See Remark 3 in the Appendix for more details and Figures 16 and 17 in the Appendix for examples.

We now extend the 1D estimate to a general dimension d. The main difference is that the total number of spectral modes K and the total number of mesh elements N scale with the dimension. More specifically, for an analytic target function and a piecewise linear approximation operator, the error estimate becomes

$$\|f - f_{K,N}\|_{L^2(\Omega)} \lesssim e^{-cK^{1/d}} + N^{-2/d}.$$

A proof of this estimate and illustrations for analytic target functions in dimensions $d = 5$ and $d = 10$ are provided in Appendix A.1; see Figure 13. The dimension dependence here is standard for the spectral truncation and approximation on a finite grid. Nevertheless, the two error-controlling parameters persist: an initial spectral-decay regime controlled by the number of modes K, followed by an error floor controlled by the mesh size N.

2.2 Neural network results

We present empirical results for trained NNs to illustrate the theoretical findings in section 2.1. A standard feedforward ReLU NN with three hidden layers of width 64 is randomly initialized and trained using the stochastic gradient-based Adam optimizer to approximate a target function f. After training, we regard its output f^{NN} as an analogue of the exact function in the theoretical analysis. This viewpoint is natural when the NN provides a sufficiently accurate approximation of f, and is also useful when the target f is not explicitly known. We then approximate the NN output using the global piecewise linear basis described above and examine the resulting approximation error.

In the first example, the ReLU NN is trained using mean-squared error loss to approximate the analytic target function considered in Figure 3. We construct a global approximation of the NN output using the piecewise linear Fourier basis, followed equation 2. As shown in Figure 4, the approximation error initially decays rapidly before reaching a grid-dependent accuracy barrier, similar to the two-regime behavior observed in Figure 3. Additional NN results for $d = 5$ and $d = 10$ show similar error behavior and are provided in the Appendix.

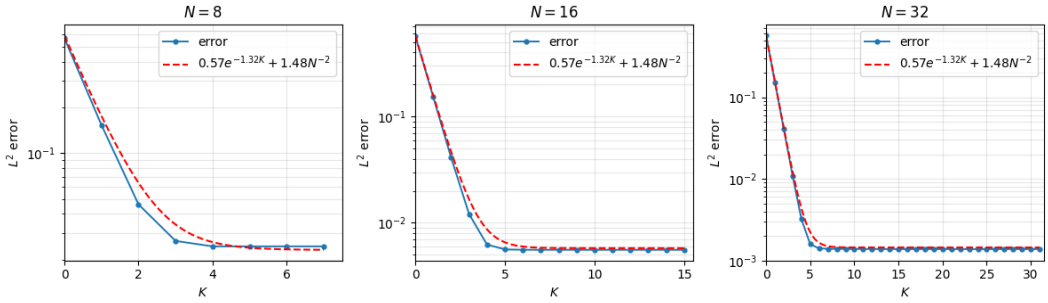


Figure 4: Accuracy barrier and error-controlling parameters, as in Figure 3, for a trained ReLU NN. The plots show the approximation error $\|f^{\mathrm{NN}} - f^{\mathrm{NN}}_{K,N}\|_2$ for the network output trained to approximate the target function $f(x) = \frac{1}{2-\cos(x)}$. The approximation uses the global piecewise linear Fourier basis from equation 1 and Figure 2. The red dashed curves show the theoretical bound for $\|f - f_{K,N}\|_2$ for reference. The results provide an empirical counterpart to the theoretical analysis and exhibit the same two-regime error behavior as in Figure 3.

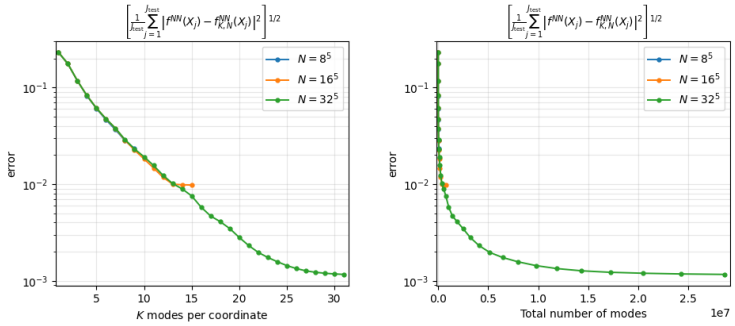


Figure 5: Accuracy barrier and error-controlling parameters for a trained ReLU NN on the MNIST dataset. The plots show the empirical approximation error on the test data; the underlying target function $f(x)$ given by the conditional class probabilities is unknown. For the approximation analysis, the data are represented using the leading five PCA components, and the global approximation is constructed using the piecewise linear Legendre tensor-product basis. The errors are shown versus the number of modes per coordinate and the total number of modes, respectively. A two-regime behavior is observed: the initial decay is relatively slow, consistent with the greater complexity and dimensionality of the classification problem, before the error reaches an accuracy floor.

In the second example, we consider the MNIST dataset, where the underlying target function is not explicitly known. For the approximation analysis, the images are projected onto the leading five principal components using PCA fitted on the training set, and the resulting coordinates are rescaled to $\Omega = [-1, 1]^5$. A standard ReLU NN is then trained on these non-uniformly distributed inputs for digit classification using softmax and cross-entropy loss. The network produces logits $g(x) \in \mathbb{R}^{10}$ with output $f^{NN}(x) = \mathrm{softmax}(g(x))$ approximating the unknown conditional class probabilities $f(x) = \big(\mathbb{P}(Y = 0 \mid X = x), \ldots, \mathbb{P}(Y = 9 \mid X = x)\big)$, or equivalently $f(x) = \mathbb{E}[Y \mid X = x]$ when Y is the one-hot digit label. After training, the network output is analyzed using a tensor-product basis constructed from piecewise-linear interpolants of Legendre polynomials. Figure 5 shows a two-regime behavior in the empirical approximation error: the initial decay is slower than in the analytic examples, consistent with the more complex, high-dimensional classification setting, before

the error reaches an accuracy floor. Additional experiments with other activations, such as Softplus, on the MNIST dataset and with the NASA Airfoil Self-Noise dataset (Brooks & Marcolini, 1989) show similar behavior and are provided in the Appendix, see Figures 21 - 23.

3 Training data and a second accuracy barrier

The error analysis in section 2 identifies an accuracy barrier associated with finite resolution. We now consider the additional effect of finite training data, which can introduce a second accuracy barrier. We first illustrate this effect using trained NNs, then relate this data error barrier to a Monte Carlo numerical integration error, arising from the optimization problem for training.

Figure 6 compares networks trained with different sample sizes. We use a standard feedforward ReLU NN with 3 hidden layers of width 64, trained on sample pairs $\{(x_j, f(x_j))\}$, $j = 1, \ldots, J_{\text{sample}}$, with i.i.d. samples $x_j \in [0, 2\pi]$. The approximation error exhibits two regimes: an initial decay in K followed by saturation. For small sample sizes, the approximation floor error lies above the theoretical resolution error floor. As the sample size increases, this approximation floor error decreases toward that resolution floor, suggesting an additional accuracy barrier associated with finite training data. For the global approximation, we fix $N = 32$, sufficiently large to make the effect of finite sample size visible before the finite resolution barrier is reached; results for other N are provided in the Appendix.

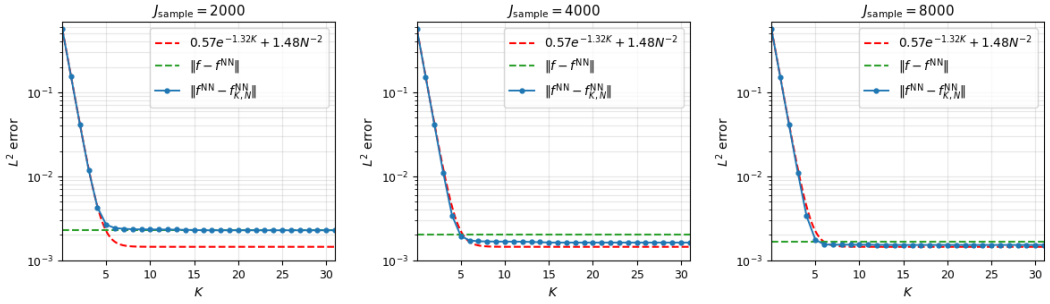


Figure 6: A second accuracy barrier arises from finite training data. The plots show the errors $\|f^{\text{NN}} - f^{\text{NN}}_{K,N}\|_2$ for ReLU NNs trained to approximate $f(x) = \frac{1}{2-\cos x}$ as in Figure 4, using different sample sizes $J_{\text{sample}} = 2000,\ 4000,\ 8000$. The learning error $\|f - f^{\text{NN}}\|_2$ and the theoretical bound for $\|f - f_{K,N}\|_2$ from Figure 3 are shown for reference. As the sample size decreases, the finiteness of the training data gives rise to a second accuracy barrier.

To examine the contribution of finite data, let samples $X_1, \ldots, X_J \overset{\text{i.i.d.}}{\sim} \mu$, where μ is a general probability distribution on Ω with density p so that $d\mu = p\, d\mathbf{x}$, and define $\|f\|^2_{L^2_\mu} = \int_\Omega |f(\mathbf{x})|^2\, d\mu(\mathbf{x})$. Consider the optimization problem

$$\min_{\{c_\mathbf{k}\}} \|f - \sum_{\mathbf{k}\in\Lambda_K} c_\mathbf{k}\Phi_\mathbf{k}\|^2_{L^2_\mu}, \tag{4}$$

where the $\{\Phi_\mathbf{k}\}$ are orthonormalized in L^2_μ and Λ_K denotes the set of retained spectral modes. With the data $\{(X_j, f(X_j))\}_{j=1}^J$ available, a data-driven version would replace the integral in the objective by a sample average:

$$\min_{\{c_\mathbf{k}\}} \frac{1}{J}\sum_{j=1}^{J} |f(X_j) - \sum_{\mathbf{k}\in\Lambda_K} c_\mathbf{k}\Phi_\mathbf{k}(X_j)|^2. \tag{5}$$

Alternatively, we can first solve equation 4 and then estimate the resulting coefficients using Monte Carlo numerical integration (MCNI) approximation:

$$\widehat{f_\mathbf{k}} = \int_\Omega f(\mathbf{x})\overline{\Phi_\mathbf{k}(\mathbf{x})}\, d\mu(\mathbf{x}) \approx \frac{1}{J}\sum_{j=1}^{J} f(X_j)\overline{\Phi_\mathbf{k}(X_j)} =: \widehat{f}^{\text{MC}}_\mathbf{k}. \tag{6}$$

We then define

$$f_{K,N}^{\mathrm{MC}} = \sum_{\mathbf{k}\in\Lambda_K} \widehat{f}_{\mathbf{k}}^{\mathrm{MC}} \Phi_{\mathbf{k}}, \tag{7}$$

and consider the error averaged over random datasets and derive the error estimate:

$$(\mathbb{E}[\|f - f_{K,N}^{\mathrm{MC}}\|_{L^2_\mu}^2])^{1/2} \lesssim e^{-cK^{1/d}} + N^{-2/d} + \sqrt{\frac{K}{J}}, \tag{8}$$

where the implicit constant is independent of K, N and J. See the Appendix for further details.

Figure 7 compares global approximation errors obtained by optimization and MCNI for a trained NN output in 2D. Both exhibit an initial decay as additional retained parameters are included. The optimization error then saturates, while the MCNI error is followed by a short plateau associated with the finite representation and an eventual increase as finite-sample errors accumulate, consistent with the multiple error contributions in equation 8.

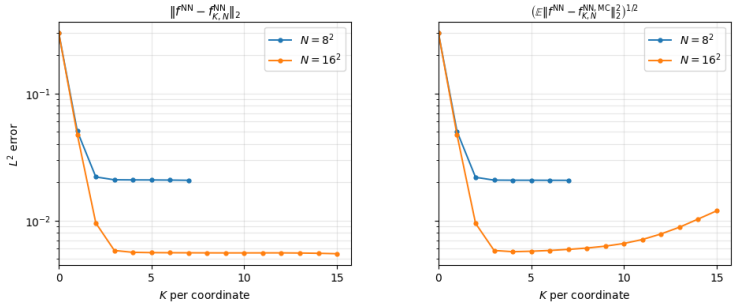


Figure 7: Relation of the finite-data error barrier to a Monte Carlo numerical integration error arising from the optimization problem for training. A ReLU NN is trained to approximate the target function $f(x) = \frac{1}{2d - \sum_{i=1}^{d} \cos(x_i)}$, with $d = 2$, and its output is approximated using a global tensor-product piecewise linear Fourier basis. The left panel shows the error $\|f^{NN} - f_{K,N}^{NN}\|_2$ obtained by optimization. The right panel shows the error $(\mathbb{E}\left[\|f - f_{K,N}^{\mathrm{MC}}\|_2^2\right])^{1/2}$ when the coefficients are estimated from finite samples using MCNI, which reveals multiple error contributions in terms of the number of retained parameters, the finite representation, and the amount of training data.

In some other settings, the approach of estimating expansion coefficients from sample averages has also been used, for instance, in uncertainty quantification (e.g., Maître & Knio, 2010) and nonparametric statistics (e.g., Tsybakov, 2009). In some detail, one can note that discrete least squares returns $G_J^{-1} b_J$ with empirical Gram matrix $G_J = \frac{1}{J}\sum_j \Phi(x_j)\Phi(x_j)^\top$ and $b_J = \frac{1}{J}\sum_j f(x_j)\Phi(x_j)$, whereas $\widehat{f}^{\mathrm{MC}} = b_J$; i.e., MCNI corresponds to least squares with G_J replaced by its expectation I.

Furthermore, the decomposition in equation 8 separates the approximation error into the effects of retaining finitely many modes and representing those modes at finite resolution. Together with the finite-sample error, this yields the three-term decomposition in equation 8, making explicit the different roles of the number of retained coefficients K, the representation resolution N, and the sample size J.

The data error term $\sqrt{K/J}$ in equation 8 is roughly in line with empirical neural scaling laws (e.g., Hoffmann et al., 2022) as well as results from statistical learning theory (e.g., Bartlett & Mendelson, 2002). Here the aim is not to derive a more precise estimate or a more precise power-law exponent for such scaling. Rather the result in equation 8 provides a simple perspective.

A simplified perspective on the training process is suggested by Figure 7 and equation 8. First, the error-controlling parameters (i.e., the coefficients of the first features, when $K = 1, 2, 3$ in Figure 7) can be determined from an explicit formula, by Monte Carlo numerical integration over the training data. Then the other basis functions or features should have their coefficients set to zero, since their values are too small to accurately estimate from the given data, and since the MCNI estimate would lead to an accumulation of error. In this way, the optimization provides a "smart" use of MCNI, where MCNI is used when it works and ignored otherwise.

4 CONCLUDING DISCUSSION

In this paper, we have shown that a relatively small number of parameters of an NN can be identified as the error-controlling parameters. A related aspect is the presence of multiple rates of error decay, which is a refinement of the typical perspective of many theoretical studies (e.g., Yarotsky, 2017; Yarotsky & Zhevnerchuk, 2020; DeVore et al., 2021), where a single scaling law is sought for the relationship between NN size and accuracy. Here, we highlight that a more refined viewpoint can be taken, and multiple rates of error decay can be identified.

The source of the first error floor is the piecewise nature of ReLU neural networks, and the consequent errors in representing basis functions or fundamental features. Even when the ReLU activation function is replaced by a smoothed version, such as softmax activation, we showed in empirical results that the first error floor still arises.

In addition to this first error floor that is related to representation error, a second error floor was related to finite amounts of data. In a theoretical perspective on the data error floor, we described its relation to Monte Carlo numerical integration (MCNI) error. Moreover, in trained NNs, one could see that the results of the training process could be obtained through a more transparent procedure, where MCNI is used in a "smart" way, i.e., MCNI is used when it works and ignored otherwise.

Future work could investigate these aspects in further detail, beyond the results presented here. For example, here the MNIST dataset was investigated as an example, and future work could investigate settings such as text and language, to explore the error floors in other applications.

AI USE STATEMENT

Generative AI tools were used to assist with polishing the writing. All scientific content, analyses, and results were produced and verified by the authors.

ETHICS STATEMENT

This work does not raise any specific ethical concerns.

REPRODUCIBILITY STATEMENT

Details supporting reproducibility, including theoretical assumptions, proofs, and experimental settings, are provided in the main text and appendix.

REFERENCES

Raman Arora, Amitabh Basu, Poorya Mianjy, and Anirbit Mukherjee. Understanding deep neural networks with rectified linear units. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=B1J_rgWRW.

Peter L Bartlett and Shahar Mendelson. Rademacher and Gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3(Nov):463–482, 2002.

Peter L Bartlett, Philip M Long, Gábor Lugosi, and Alexander Tsigler. Benign overfitting in linear regression. *Proceedings of the National Academy of Sciences*, 117(48):30063–30070, 2020.

Mikhail Belkin, Daniel Hsu, Siyuan Ma, and Soumik Mandal. Reconciling modern machine-learning practice and the classical bias–variance trade-off. *Proceedings of the National Academy of Sciences*, 116(32):15849–15854, 2019. doi: 10.1073/pnas.1903070116.

Pope D. Brooks, Thomas and Michael Marcolini. Airfoil Self-Noise. UCI Machine Learning Repository, 1989. DOI: https://doi.org/10.24432/C5VW2C.

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

Ronald DeVore, Boris Hanin, and Guergana Petrova. Neural network approximation. *Acta Numerica*, 30:327–444, 2021.

Simon S. Du, Xiyu Zhai, Barnabas Poczos, and Aarti Singh. Gradient descent provably optimizes over-parameterized neural networks. In *International Conference on Learning Representations*, 2019. URL https://arxiv.org/abs/1810.02054.

Dennis Elbrächter, Dmytro Perekrestenko, Philipp Grohs, and Helmut Bölcskei. Deep neural network approximation theory. *IEEE Transactions on Information Theory*, 67(5):2581–2623, 2021.

Trevor Hastie, Andrea Montanari, Saharon Rosset, and Ryan J Tibshirani. Surprises in high-dimensional ridgeless least squares interpolation. *Annals of Statistics*, 50(2):949–986, 2022.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, et al. Training compute-optimal large language models. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, pp. 30016–30030, 2022.

Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: Convergence and generalization in neural networks. In *Advances in Neural Information Processing Systems*, volume 31, 2018. URL https://proceedings.neurips.cc/paper/2018/hash/5a4be1fa34e62bb8a6ec6b91d2462f5a-Abstract.html.

Patrick Kidger and Terry Lyons. Universal Approximation with Deep Narrow Networks. In Jacob Abernethy and Shivani Agarwal (eds.), *Proceedings of Thirty Third Conference on Learning Theory*, volume 125 of *Proceedings of Machine Learning Research*, pp. 2306–2327. PMLR, 09–12 Jul 2020. URL https://proceedings.mlr.press/v125/kidger20a.html.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998. doi: 10.1109/5.726791.

Marco Loog, Tom Viering, Alexander Mey, Jesse H Krijthe, and David MJ Tax. A brief prehistory of double descent. *Proceedings of the National Academy of Sciences*, 117(20):10625–10626, 2020.

Olivier P. Le Maître and Omar M. Knio. *Spectral Methods for Uncertainty Quantification: With Applications to Computational Fluid Dynamics*. Scientific Computation. Springer, Dordrecht, 2010.

Allan Pinkus. Approximation theory of the MLP model in neural networks. *Acta Numerica*, 8: 143–195, 1999.

Nasim Rahaman, Aristide Baratin, Devansh Arpit, Felix Draxler, Min Lin, Fred Hamprecht, Yoshua Bengio, and Aaron Courville. On the spectral bias of neural networks. In Kamalika Chaudhuri and Ruslan Salakhutdinov (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 5301–5310. PMLR, 09–15 Jun 2019. URL https://proceedings.mlr.press/v97/rahaman19a.html.

Rylan Schaeffer, Zachary Robertson, Akhilan Boopathy, Mikail Khona, Kateryna Pistunova, Jason William Rocks, Ila R Fiete, Andrey Gromov, and Sanmi Koyejo. Double descent demystified: Identifying, interpreting & ablating the sources of a deep learning puzzle. In *The Third Blogpost Track at ICLR 2024*, 2024. URL https://openreview.net/forum?id=muC7uLvGHr.

Matus Telgarsky. Representation benefits of deep feedforward networks. *arXiv preprint arXiv:1509.08101*, 2015. URL https://arxiv.org/abs/1509.08101.

Alexandre B. Tsybakov. *Introduction to Nonparametric Estimation*. Springer Series in Statistics. Springer, New York, 2009.

Dmitry Yarotsky. Error bounds for approximations with deep ReLU networks. *Neural Networks*, 94: 103–114, 2017.

Dmitry Yarotsky and Anton Zhevnerchuk. The phase diagram of approximation rates for deep neural networks. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 13005–13015. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/979a3f14bae523dc5101c52120c535e9-Paper.pdf.

A APPENDIX

This appendix contains additional details supporting the main text. Appendix A.1 gives proof of Theorem 1 and presents extensions of the error analysis to more general target functions, higher-dimensional settings, and finite-sample data. Appendix A.2 provides additional numerical test cases illustrating the main results.

A.1 ERROR ESTIMATES: PROOFS AND EXTENSIONS

A.1.1 PROOF OF THEOREM 1 IN DIMENSION d

We first prove Theorem 1 in a general setting: a piecewise polynomial approximation operator of degree p for an analytic target function in dimension d. The case $p = 1$ corresponds to piecewise linear approximation, which is closely related to ReLU NN.

Proof. Let

$$f_K = \sum_{\|\mathbf{k}\|_\infty \le \lfloor K^{1/d} \rfloor} \hat{f}_{\mathbf{k}} \phi_{\mathbf{k}}$$

be the exact Fourier truncation of f in the Fourier basis $\{\phi_{\mathbf{k}}\}_{\mathbf{k}\in\mathbb{Z}^d}$, where $\hat{f}_{\mathbf{k}}$ denotes the corresponding Fourier coefficient. Here K is a spectral budget parameter: the largest frequency included in the truncation is of order $K^{1/d}$, and the number of basis modes included is $\mathcal{O}(K)$.

Define

$$v_{K,N} := \sum_{\|\mathbf{k}\|_\infty \le \lfloor K^{1/d} \rfloor} \hat{f}_{\mathbf{k}} \Phi_{\mathbf{k}} \in W_{K,N},$$

where $\Phi_{\mathbf{k}}$ is the piecewise polynomial approximation of degree p to $\phi_{\mathbf{k}}$ on a mesh with N total elements.

Since $f_{K,N}$ is the best approximation to f in $W_{K,N}$, we have

$$\|f - f_{K,N}\|_{L^2(\Omega)} \le \|f - v_{K,N}\|_{L^2(\Omega)}.$$

Therefore,

$$\|f - f_{K,N}\|_{L^2(\Omega)} \le \|f - f_K\|_{L^2(\Omega)} + \|f_K - v_{K,N}\|_{L^2(\Omega)}.$$

We first estimate the truncation error. Since f is periodic and analytic, its Fourier coefficients satisfy

$$|\hat{f}_{\mathbf{k}}| \le c_1 e^{-c_0\|\mathbf{k}\|_\infty}, \qquad \mathbf{k} \in \mathbb{Z}^d,$$

for some constants $c_0, c_1 > 0$ independent of $\mathbf{k}$. Hence the truncation error of the exact Fourier approximation satisfies

$$\|f - f_K\|_{L^2(\Omega)} \le C_1 e^{-cK^{1/d}},$$

where $C_1, c > 0$ are independent of K and N.

Next, we estimate the error from replacing each Fourier mode by its piecewise polynomial approximation. Assume that the mesh has N total elements in d dimensions, with mesh size $h \sim N^{-1/d}$, as in uniformly partitioned or quasi-uniform meshes. The standard interpolation estimate gives

$$\|\phi_{\mathbf{k}} - \Phi_{\mathbf{k}}\|_{L^2(\Omega)} \le c_2(1 + \|\mathbf{k}\|_\infty)^{p+1} N^{-(p+1)/d},$$

for some constant $c_2 > 0$ independent of $\mathbf{k}$ and N.

Thus,

$$\begin{aligned}\|f_K - v_{K,N}\|_{L^2(\Omega)} &\le \sum_{\|\mathbf{k}\|_\infty \le \lfloor K^{1/d} \rfloor} |\hat{f}_{\mathbf{k}}| \|\phi_{\mathbf{k}} - \Phi_{\mathbf{k}}\|_{L^2(\Omega)} \\ &\le c_2 N^{-(p+1)/d} \sum_{\|\mathbf{k}\|_\infty \le \lfloor K^{1/d} \rfloor} |\hat{f}_{\mathbf{k}}| (1 + \|\mathbf{k}\|_\infty)^{p+1} \\ &\le c_2 c_1 N^{-(p+1)/d} \sum_{\|\mathbf{k}\|_\infty \le \lfloor K^{1/d} \rfloor} (1 + \|\mathbf{k}\|_\infty)^{p+1} e^{-c_0 \|\mathbf{k}\|_\infty} \\ &\le c_2 c_1 N^{-(p+1)/d} \sum_{\mathbf{k} \in \mathbb{Z}^d} (1 + \|\mathbf{k}\|_\infty)^{p+1} e^{-c_0 \|\mathbf{k}\|_\infty} \\ &= C_2 N^{-(p+1)/d},\end{aligned}$$

where

$$C_2 = c_1 c_2 C_{p,d}, \qquad C_{p,d} = \sum_{\mathbf{k} \in \mathbb{Z}^d} (1 + \|\mathbf{k}\|_\infty)^{p+1} e^{-c_0 \|\mathbf{k}\|_\infty}.$$

Since $c_0 > 0$, the constant $C_{p,d}$ is finite. Therefore, $C_2 > 0$ is independent of K and N.

Combining the two estimates gives

$$\|f - f_{K,N}\|_{L^2(\Omega)} \le C_1 e^{-cK^{1/d}} + C_2 N^{-(p+1)/d}.$$

This proves the result. □

A.1.2 Extensions to other classes of target functions

Remark 2. This framework has straightforward extensions to more general function classes, including functions in C^m, which are m times continuously differentiable, and functions in H^m, which have finite Sobolev regularity.

For example, suppose $f \in H^m(\Omega)$ for some $m > 0$. Then the Fourier coefficients decay algebraically:

$$|\hat{f}_{\mathbf{k}}| \le c_1 (1 + \|\mathbf{k}\|_\infty)^{-m},$$

where $c_1 > 0$ is independent of $\mathbf{k}$. Hence the exact spectral truncation error satisfies

$$\|f - f_K\|_{L^2(\Omega)} \le C_1 K^{-m/d},$$

where $C_1 > 0$ is independent of K and N.

We also have an approximation estimate for each Fourier mode:

$$\|\phi_{\mathbf{k}} - \Phi_{\mathbf{k}}\|_{L^2(\Omega)} \le c_2 (1 + \|\mathbf{k}\|_\infty)^{p+1} N^{-(p+1)/d}.$$

Similarly,

$$\begin{aligned}\|f_K - v_{K,N}\|_{L^2(\Omega)} &\le \sum_{\|\mathbf{k}\|_\infty \le \lfloor K^{1/d} \rfloor} |\hat{f}_{\mathbf{k}}| \, \|\phi_{\mathbf{k}} - \Phi_{\mathbf{k}}\|_{L^2(\Omega)} \\ &\le c_1 c_2 N^{-(p+1)/d} \sum_{\mathbf{k} \in \mathbb{Z}^d} (1 + \|\mathbf{k}\|_\infty)^{p+1-m} = C_2 N^{-(p+1)/d},\end{aligned}$$

the last sum is finite when $m > p + 1 + d$. This condition is sufficient but not sharp.

Combining the two estimates also gives

$$\|f - f_{K,N}\|_{L^2(\Omega)} \le C_1 K^{-m/d} + C_2 N^{-(p+1)/d},$$

here $C_1, C_2 > 0$ is independent of K and N.

We refer to Figure 14 for a 1D example with a C^1 continuously differentiable target function. In this case, the same two-regime behavior is observed: an initial algebraic decay in the number of modes, followed by an error floor controlled by the mesh size.

Remark 3. For another common class of functions, such as piecewise-smooth functions with a finite number of discontinuities or jumps in function values, it is natural to use an adaptive approximation strategy instead of a single global approximation. The adaptive grid can be chosen so that the element boundaries align with the nonsmooth points or interfaces. Then, on each subdomain where the function is smooth, one can recover the same type of two-regime error behavior as in the globally smooth case.

More specifically, assume that f is piecewise smooth on $\Omega \subset \mathbb{R}^d$, and $\Omega = \bigcup_{\ell=1}^{L} \Omega_\ell$, is partitioned into finitely many subdomains, where the subdomains are disjoint up to their boundaries, and that $f_\ell := f|_{\Omega_\ell} \in H^{m_\ell}(\Omega_\ell)$ (or $C^{m_\ell}(\Omega_\ell)$) for each $\ell = 1, \ldots, L$. We approximate f locally on each smooth subdomain:

Let

$$f_{K,\ell} = \sum_{\|\mathbf{k}\|_\infty \le \lfloor K_\ell^{1/d} \rfloor} \hat{f}_{\mathbf{k}}^{(\ell)} \phi_{\mathbf{k}}^{(\ell)}$$

be the local spectral truncation of f_ℓ on Ω_ℓ. Define

$$v_{K,N}|_{\Omega_\ell} := \sum_{\|\mathbf{k}\|_\infty \le \lfloor K_\ell^{1/d} \rfloor} \hat{f}_{\mathbf{k}}^{(\ell)} \Phi_{\mathbf{k}}^{(\ell)},$$

where $\Phi_{\mathbf{k}}^{(\ell)}$ is the piecewise polynomial approximation of degree p to $\phi_{\mathbf{k}}^{(\ell)}$ on Ω_ℓ. Let $K = \sum_{\ell=1}^{L} K_\ell$ and $N = \sum_{\ell=1}^{L} N_\ell$ denote the total numbers of spectral modes and mesh budgets, respectively.

On each subdomain Ω_ℓ, the local error satisfies

$$\|f_\ell - v_{K,N}|_{\Omega_\ell}\|_{L^2(\Omega_\ell)} \le C_{\ell,1} K_\ell^{-m_\ell/d} + C_{\ell,2} N_\ell^{-(p+1)/d},$$

where $C_{\ell,1}, C_{\ell,2} > 0$ are independent of K and N. Then, the global error satisfies

$$\|f - f_{K,N}\|_{L^2(\Omega)} \le \sum_{\ell=1}^{L} \left(C_{\ell,1} K_\ell^{-m_\ell/d} + C_{\ell,2} N_\ell^{-(p+1)/d} \right).$$

We refer to Figures 16 and 17 for 1D and 2D examples of approximations of piecewise-smooth target functions. These examples show that adaptive approximation on locally smooth subdomains exhibits the same two-regime error behavior.

Remark 4 (Crossover point K_*). Consider the error bound

$$\|f - f_{K,N}\| \le C_1 e^{-cK} + C_2 N^{-2}.$$

For each fixed grid size N, define the crossover point K_* (or "elbow") as the value of K modes where the two terms are comparable $C_1 e^{-cK_*} \approx C_2 N^{-2}$. Solving for K_* gives

$$K_* \approx \frac{1}{c} \log\left(\frac{C_1}{C_2}\right) + \frac{2}{c} \log N.$$

This formula shows that the elbow K_* shifts slowly to the right as the grid is refined, since K_* grows logarithmically with N. The elbow point exists when $K_* > 0$, or equivalently, $\frac{C_1}{C_2} N^2 > 1$. This condition is satisfied for all sufficiently large N grid sizes.

If this condition does not hold for small values of N, then the floor error is already comparable to, or larger than, the spectral truncation error for the first few modes. In that case, the spectral-decay regime may be very short or may not be visible.

If the condition holds, then the error exhibits multiple levels. In the first regime, where $K < K_*$, the exponential truncation error $C_1 e^{-cK}$ dominates, and the error decreases rapidly with spectral-type decay as the number of modes K increases. In this regime, K is the main error-controlling parameter.

In the second regime, where $K \gg K_*$, the finite-resolution floor $C_2 N^{-2}$ dominates. Increasing the number of spectral modes K no longer significantly reduces the error. Instead, the grid size N becomes the error-controlling parameter. This produces an accuracy barrier: the underlying piecewise linear approximation space has fixed resolution, which imposes an algebraic floor of order $O(N^{-2})$.

Together, this shows that a substantial refinement of the grid N produces only a moderate increase, of order $\mathcal{O}(\log N)$, in the number of spectral modes that can be used before the floor error is reached.

If $f \in H^s(\Omega)$ with $s \geq 2$, the truncation term becomes algebraic: $\|f - f_{K,N}\| \leq C_1 K^{-s} + C_2 N^{-2}$. The first regime then shows algebraic decay with a more gradual transition to the floor. Balancing the two terms gives $K_* \sim N^{2/s}$, so the elbow grows algebraically with N, approaching the analytic case as $s \to \infty$. For $1/2 < s < 2$, the floor degrades to $\mathcal{O}(N^{-s})$, and $K_* \sim N$.

A.1.3 Approximation from finite samples

To demonstrate the main ideas, we consider the tensor-product Fourier basis $\{\phi_{\mathbf{k}}\}_{\mathbf{k}}$, $\phi_{\mathbf{k}}(\mathbf{x}) = e^{i\mathbf{k}\cdot\mathbf{x}}$ on $\Omega = [0, 2\pi]^d$ and show that for an analytic function f,

$$\left(\mathbb{E}\left[\|f - f_{K,N}^{\mathrm{MC}}\|_{L^2_\mu}^2\right]\right)^{1/2} \lesssim e^{-cK^{1/d}} + N^{-2/d} + \sqrt{\frac{K}{J}},$$

The implicit constant is independent of K, N, and J. The overall argument can be extended to other tensor-product spectral bases, such as Legendre or Chebyshev bases.

Proof. Define the weighted Euclidean norm $\|\mathbf{k}\|_{w,2} = \left(\sum_{i=1}^d w_i k_i^2\right)^{1/2}$, for some $w_1, \ldots, w_d > 0$, and the spectral index set $\Lambda_K = \left\{\mathbf{k} \in \mathbb{Z}^d : \|\mathbf{k}\|_{w,2} \leq \lfloor K^{1/d} \rfloor\right\}$, so that $|\Lambda_K| \lesssim K$. The same analysis also applies to the full tensor-product cutoff based on $\|\mathbf{k}\|_\infty$. Here, we use the weighted Euclidean cutoff since it can retain fewer modes while maintaining the same order of spectral resolution. Again, assume that the mesh has N total elements in d dimensions, with mesh size $h \sim N^{-1/d}$, as in uniformly partitioned meshes, and that the retained Fourier modes lie below the Nyquist frequency of the mesh, corresponding to $K \lesssim N$.

Let $\{\Phi_{\mathbf{k}}\}$ be the tensor-product piecewise linear interpolant of $\{\phi_{\mathbf{k}}\}$ on the uniform mesh, we orthonormalize $\{\Phi_{\mathbf{k}}\}$ in L^2_μ and denote the resulting basis by $\{\Psi_{\mathbf{k}}\}$. In the main text, for simplicity, we continue to use the notation $\Phi_{\mathbf{k}}$ for these functions.

Denote

$$f_{K,N} = \sum_{\mathbf{k}\in\Lambda_K} \widehat{f}_{\mathbf{k}} \Psi_{\mathbf{k}}, \qquad \widehat{f}_{\mathbf{k}} = \langle f, \Psi_{\mathbf{k}} \rangle_{L^2_\mu} = \int_\Omega f(\mathbf{x}) \overline{\Psi_{\mathbf{k}}(\mathbf{x})}\, d\mu(\mathbf{x}),$$

We have $X_1, \ldots, X_J \overset{\text{i.i.d.}}{\sim} \mu$. The corresponding finite-sample approximation and Monte Carlo coefficients are

$$f_{K,N}^{\mathrm{MC}} = \sum_{\mathbf{k}\in\Lambda_K} \widehat{f}_{\mathbf{k}}^{\mathrm{MC}} \Psi_{\mathbf{k}}, \qquad \widehat{f}_{\mathbf{k}}^{\mathrm{MC}} = \frac{1}{J} \sum_{j=1}^J f(X_j) \overline{\Psi_{\mathbf{k}}(X_j)}.$$

We split the error as

$$f - f_{K,N}^{MC} = (f - f_{K,N}) + (f_{K,N} - f_{K,N}^{MC}).$$

By a derivation similar to that in Appendix A.1.1, we can have

$$\|f - f_{K,N}\|_{L^2_\mu} \lesssim e^{-cK^{1/d}} + N^{-2/d},$$

for some $c > 0$ that depends on the analyticity of f, and the implicit constant is independent of K and N. The first term represents the spectral truncation error, which decays exponentially for analytic functions and algebraically for functions with finite Sobolev regularity, while the second term represents the approximation-resolution floor.

It remains to estimate the sampling error. We have

$$\mathbb{E}\left[\left|\widehat{f}_{\mathbf{k}}^{\mathrm{MC}} - \widehat{f}_{\mathbf{k}}\right|^2\right] = \frac{1}{J} \operatorname{Var}\left(f(X)\overline{\Psi_{\mathbf{k}}(X)}\right) \leq \frac{1}{J} \mathbb{E}\left[|f(X)|^2 |\Psi_{\mathbf{k}}(X)|^2\right].$$

Then by the orthonormality of $\{\Psi_{\mathbf{k}}\}$, we derive

$$\mathbb{E}\left[\|f_{K,N} - f_{K,N}^{\mathrm{MC}}\|_{L^2_\mu}^2\right] = \sum_{\mathbf{k}\in\Lambda_K} \mathbb{E}\left[\left|\widehat{f}_{\mathbf{k}} - \widehat{f}_{\mathbf{k}}^{\mathrm{MC}}\right|^2\right] \le \frac{1}{J}\,\mathbb{E}\left[|f(X)|^2 \sum_{\mathbf{k}\in\Lambda_K} |\Psi_{\mathbf{k}}(X)|^2\right]$$
$$\le \frac{\|f\|_{L^\infty}^2}{J} \sum_{\mathbf{k}\in\Lambda_K} \|\Psi_{\mathbf{k}}\|_{L^2_\mu}^2 \le \frac{|\Lambda_K|}{J}\,\|f\|_{L^\infty}^2 \lesssim \frac{K}{J}.$$

Finally, since $f - f_{K,N}$ is deterministic and $\mathbb{E}\left[f_{K,N}^{\mathrm{MC}}\right] = f_{K,N}$, the cross term has expectation zero, and

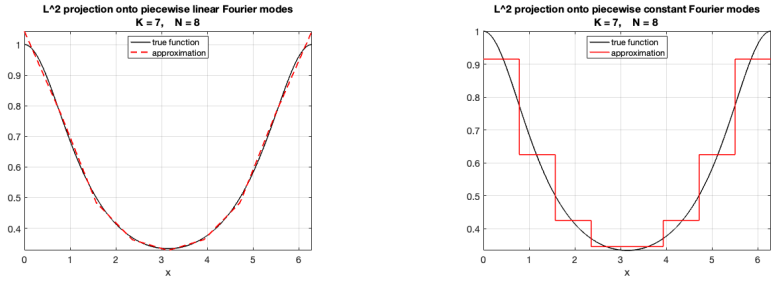
$$\mathbb{E}\left[\|f - f_{K,N}^{\mathrm{MC}}\|_{L^2_\mu}^2\right] = \|f - f_{K,N}\|_{L^2_\mu}^2 + \mathbb{E}\left[\|f_{K,N} - f_{K,N}^{\mathrm{MC}}\|_{L^2_\mu}^2\right].$$

Taking square roots, combining the preceding estimates, and using $\sqrt{a^2+b^2} \le a+b$ for all $a, b \ge 0$ concludes the proof.

□

A.2 Additional test cases

In this section, we present several test cases to further support results in the main text.



(a) Piecewise linear approximation (b) Piecewise constant approximation

Figure 8: Approximation of the analytic target function $f(x) = \frac{1}{2-\cos(x)}$ on $\Omega = [0, 2\pi]$, using $N = 8$ uniformly distributed piecewise segments and $K = 7$ modes. The approximations are constructed using the basis functions defined in equation 1 and illustrated in Figure 2. Panel (a) corresponds to the piecewise linear approximation, with the error plot shown in Figure 3 in Section 2.1. Panel (b) corresponds to the piecewise constant approximation, with the error plot shown in Figure 9 below.

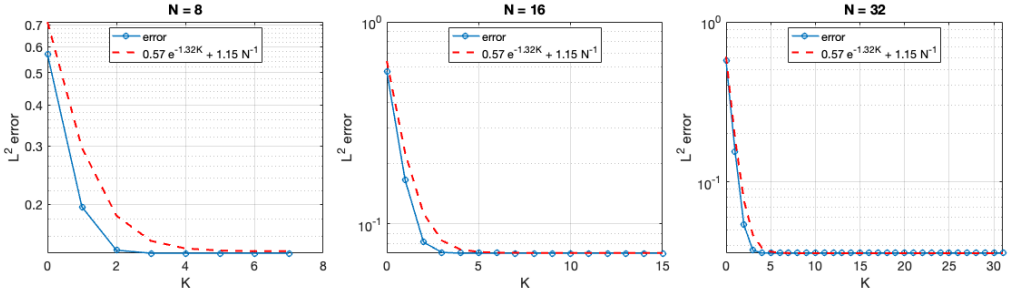
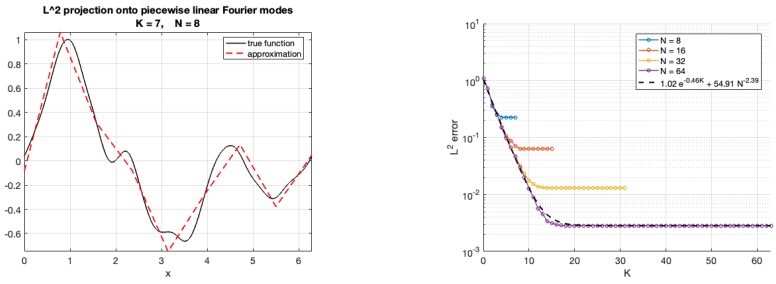
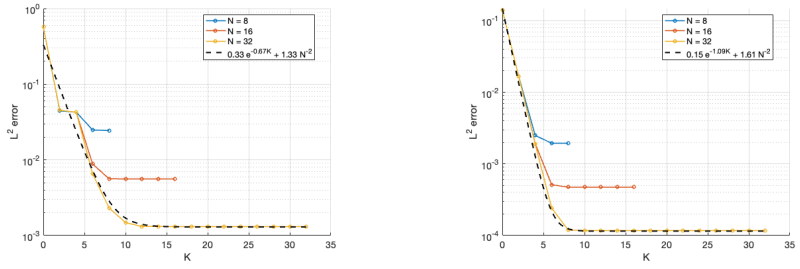


Figure 9: Piecewise constant approximation error for the target function $f(x) = \frac{1}{2-\cos(x)}$ on $\Omega = [0, 2\pi]$. Each panel shows the error versus the number of modes K for a fixed grid size N, with $N = 8, 16, 32$. This figure illustrates the piecewise constant setting mentioned in Section 2.1. Here I_N in equation 1 denotes the piecewise constant interpolation operator, and the global approximation is constructed using the same procedure. The same two-regime behavior appears: rapid spectral decay followed by an accuracy barrier, with the expected $O(N^{-1})$ grid-size decay in the barrier regime.



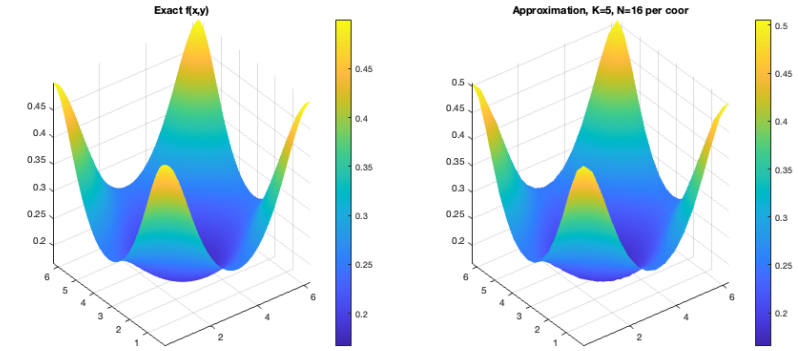
(a) Piecewise linear approximation (b) Error vs total K modes

Figure 10: Another 1D piecewise linear approximation using the basis functions from equation 1 and Figure 2. Here the target is a general analytic function $f(x) = \sum_{n=1}^{M} a_n \cos(nx + \theta_n)$, where $|a_n| \leq Ce^{-cn}$. We choose $a_n = Ce^{-cn}\,(\eta + (1-\eta)\xi_n)$, $\theta_n = 2\pi\zeta_n$, with $\xi_n, \zeta_n \sim \mathrm{Unif}(0,1)$. Here, $C = 1, c = 0.45, M = 80$, and $\eta = 0.55$.

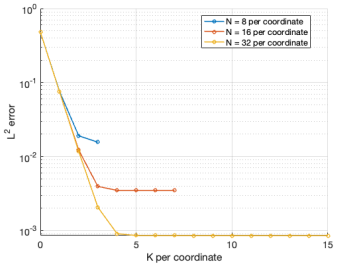


(a) Error vs K modes on $\Omega = [0, 2\pi]$ (b) Error vs K modes on $\Omega = [-1, 1]$

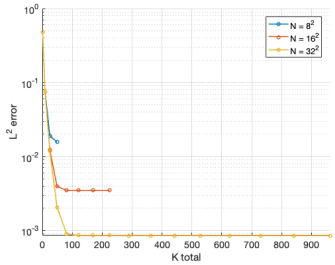
Figure 11: L^2 error for the global piecewise linear Legendre approximation of the analytic target function $f(x) = \frac{1}{2-\cos(x)}$ on different domains. This figure illustrates a setting discussed in Section 2.1, where Legendre modes are used instead of Fourier modes. The basis functions are obtained by piecewise linear interpolation of Legendre polynomials on a fixed grid. The same two-regime behavior appears: spectral-type decay followed by a mesh-dependent accuracy floor. The black dashed curves in each panel shows theoretical upper bound.



(a) Piecewise linear approximation

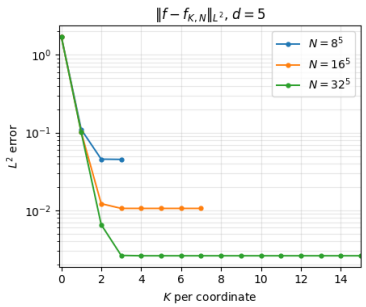


(b) Error vs K modes per coordinate

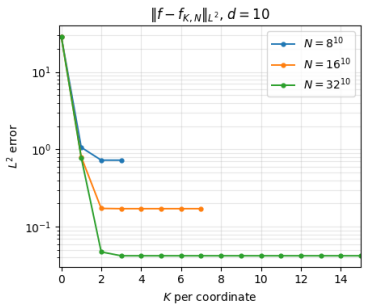


(c) Error vs total K modes

Figure 12: 2D approximation of the analytic target function $f(x) = \frac{1}{2d - \sum_{i=1}^{d} \cos(x_i)}$, $d = 2$, using the piecewise linear spectral basis functions as in equation 1. More precisely, let ϕ_m denote the Fourier mode of frequency m on $[0, 2\pi]$, and let I_N be the piecewise linear interpolation operator on a uniform grid with N segments in each coordinate direction. We use the piecewise linearized Fourier tensor-product basis functions $\Phi_{mn}(x) = (I_N\phi_m)(x_1)(I_N\phi_n)(x_2)$. Here, we plot the error for $K \leq N/2$, since the two-regime behavior is already visible for relatively small values of K. This example supports the discussion in Section 2.1 and illustrates the same two-regime error behavior in 2D. Cases for $d = 5$ and $d = 10$ are shown in Figure 13.

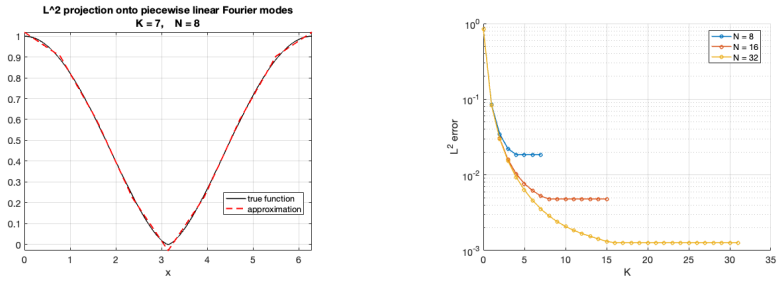


(a) $d = 5$



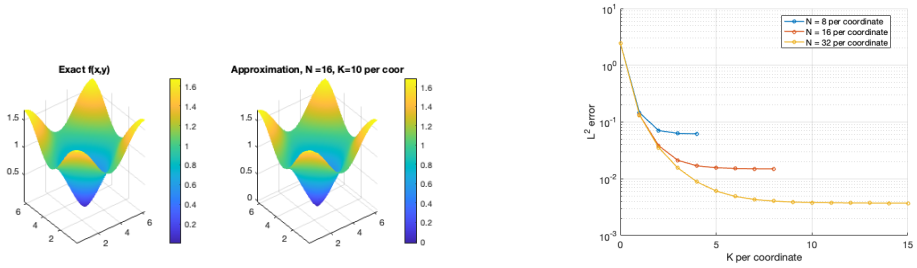
(b) $d = 10$

Figure 13: Approximation error versus the number of modes K per coordinate in dimensions $d = 5$ and $d = 10$ for the target $f(x) = \frac{1}{2d - \sum_{i=1}^{d} \cos(x_i)}$. The approximation uses a global tensor-product basis formed from piecewise linear Fourier modes. Results are shown for uniform grids with 8, 16, and 32 elements per coordinate, corresponding to total grid sizes $N = 8^d, 16^d, 32^d$.



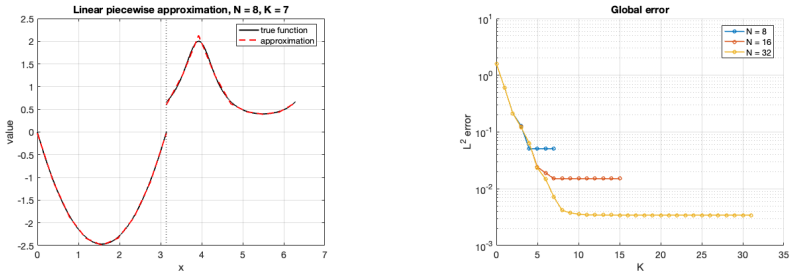
(a) Piecewise linear approximation (b) Error vs total K modes

Figure 14: 1D approximation of a non-analytic but continuously differentiable C^1 target function, $f(x) = |\sin((x-\pi)/2)|^{3/2}$ using the global piecewise spectral basis functions from equation 1. This example supports the discussion in Appendix A, Remark 2. The same two-regime behavior appears, but the initial decay is algebraic rather than exponential, followed by a mesh-dependent accuracy barrier.



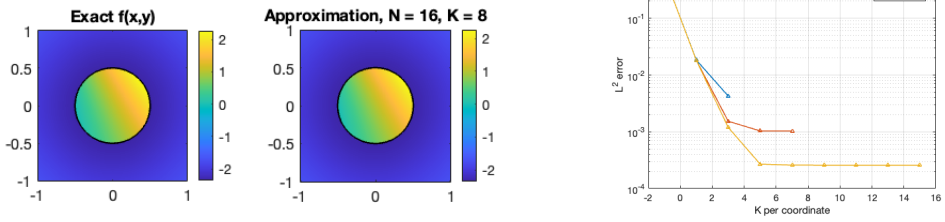
(a) Piecewise linear approximation (b) Error vs total K modes

Figure 15: 2D approximation of a C^1 target function $f(x) = \left[\sin^2\left(\frac{x_1-\pi}{2}\right) + \sin^2\left(\frac{x_2-\pi}{2}\right)\right]^{3/4}$ using 2D piecewise spectral (global) basis functions. This example supports the discussion in Appendix A, Remark 2, and Section **??**. The same two-regime behavior appears, but the initial decay is algebraic rather than exponential, followed by a mesh-dependent accuracy barrier.



(a) Piecewise linear approximation (b) Error vs total K modes

Figure 16: 1D approximation of a discontinuous, piecewise smooth target function $f(x) = x(x-\pi)$ for $x < \pi$ and $f(x) = 1/(1.5 - \cos(2x - 0.5\pi))$ for $x \geq \pi$ on $\Omega = [0, 2\pi]$, using linear piecewise Legendre basis functions as equation 1, applied on each subinterval of smoothness. On each smooth subinterval, we observe the same two-regime error behavior: an initial spectral-decay regime controlled by the number of modes K, followed by an error floor controlled by the mesh size N. The global error obtained by combining the local errors over the two subintervals follows the same behavior. This example supports the discussion in Appendix A, Remark 3.



(a) Approximation ($N = 16$, $K = 8$ per coordinate) (b) Error vs K modes per coordinate

Figure 17: 2D approximation of a discontinuous, piecewise smooth target function $f(x) = 2x_1 + exp(x_2) + x_1x_2$ for $|x| < 0.55$ and $f(x) = \frac{-5}{4-\cos(x_1)-\cos(x_2)}$ for $|x| \geq 0.55$ on $\Omega = [-1, 1]^2$, using linear piecewise Legendre basis functions as applied on each subdomain of smoothness. More precisely, let P_m denote the Legendre polynomial of degree m on $[-1, 1]$, and let I_N be the piecewise linear interpolation operator on a uniform grid with N segments in each coordinate direction. We use the piecewise linearized Legendre basis functions $\Phi_{mn}(\mathbf{x}) = (I_N P_m)(x_1)(I_N P_n)(x_2)$. The best approximation is computed locally on the two subregions of smoothness $\Omega_1 = \{x \in \Omega : |x| < 0.55\}$ and $\Omega_2 = \Omega \backslash \Omega_1$, separated by the circular interface $|x| = 0.55$. This example supports the discussion in Appendix A, Remark 3.

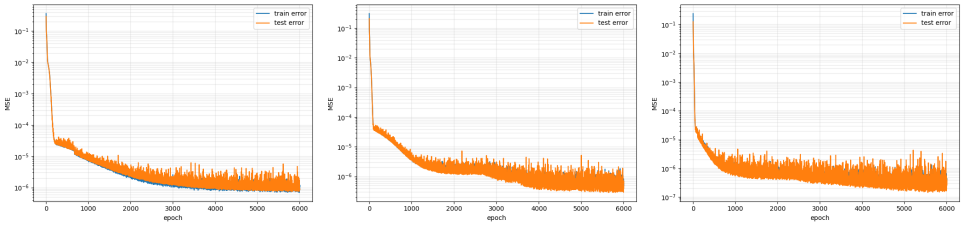


Figure 18: Training histories versus epoch for approximating $f(x) = \frac{1}{2d-\sum_{i=1}^{d}\cos(x_i)}, d = 1$, as in Figure 6. For simplicity, a constant learning rate of 10^{-4} is used for all 6000 epochs.

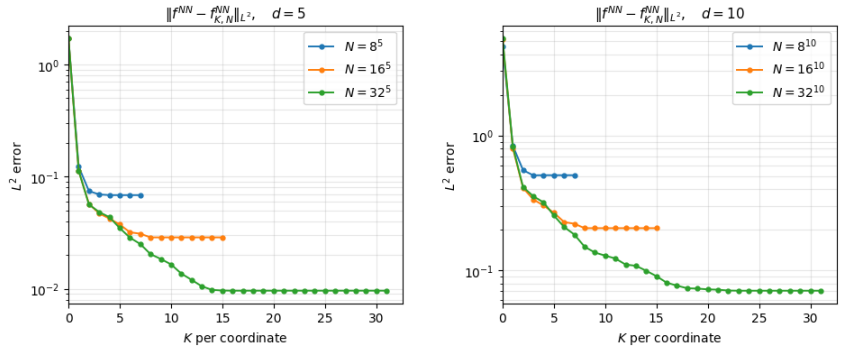


Figure 19: Approximation error for the output of a ReLU neural network trained to approximate the target function $f(x) = \frac{1}{2d-\sum_{i=1}^{d}\cos(x_i)}$ with $d = 5$ and $d = 10$, respectively, as in Figure 13. The corresponding training and validation losses are shown in Figure 20.

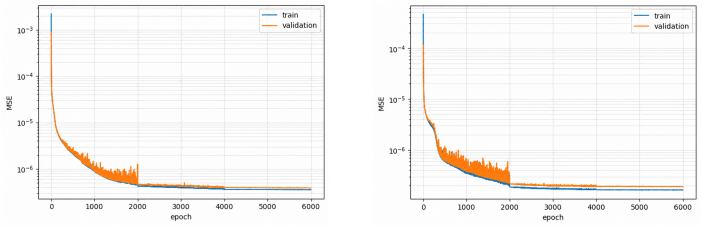


Figure 20: Training histories versus epoch to approximate $f(x) = \frac{1}{2d - \sum_{i=1}^{d} \cos(x_i)}$, with $d = 5$ and $d = 10$, respectively. A learning rate schedule is used, with learning rates 10^{-3}, 10^{-4}, and 10^{-5}, each applied for 2000 epochs.

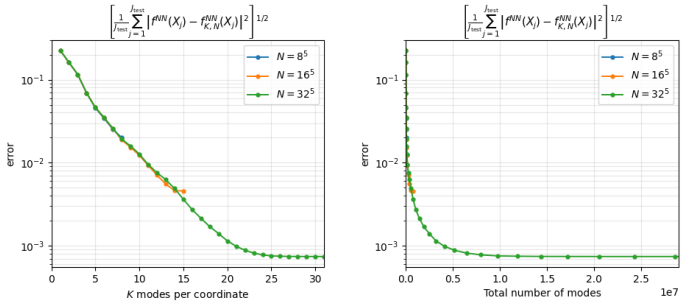


Figure 21: Accuracy barrier and error-controlling parameters for a trained neural network on the MNIST dataset, as in Figure 5, using the Softplus activation function. See Figure 22 for its training histories.

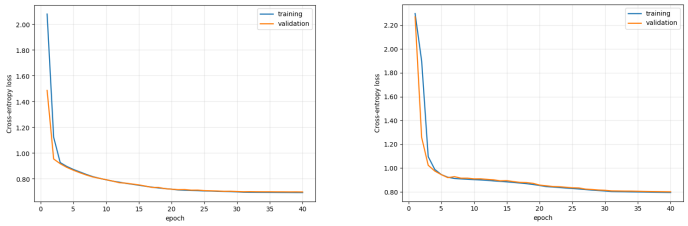
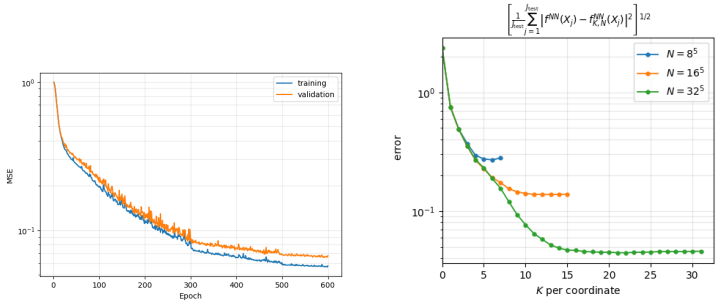


Figure 22: Training histories versus epoch for the MNIST classification problem using ReLU activation (left; see also Figure 5) and Softplus activation (right; see also Figure 21).



(a) Training histories versus epoch. (b) Approximation error of the trained NN output.

Figure 23: Training and approximation results for the NASA Airfoil Self-Noise dataset. The dataset consists of five input features and a continuous scalar output. The five input variables are rescaled to $\Omega = [-1, 1]^5$. A standard feedforward ReLU neural network is trained for regression using the mean-squared error loss. The network output $f^{NN}(x)$ approximates the unknown regression function $f(x) = \mathbb{E}[Y \mid X = x]$, where Y denotes the measured sound-pressure level. After training, f^{NN} is analyzed using a tensor-product basis constructed from piecewise-linear interpolants of Legendre polynomials.