Universal Consistency of Wide and Deep ReLU Neural Networks and Minimax Optimal Convergence Rates for Kolmogorov-Donoho Optimal Function Classes

Hyunouk Ko [1] **Xiaoming Huo** [1]

Abstract

In this paper, we prove the universal consistency of wide and deep ReLU neural network classifiers. We also give sufficient conditions for a class of probability measures for which classifiers based on neural networks achieve minimax optimal rates of convergence. The result applies to a wide range of known function classes. In particular, while most previous works impose explicit smoothness assumptions on the regression function, our framework encompasses more general settings. The proposed neural networks are either the minimizers of the 0-1 loss that exhibit a benign overfitting behavior.

While the development of statistical theory for binary classification dates back to the 1970s and is well-summarized in (Devroye et al., 2013) and (Boucheron et al., 2005), a general theory explaining the generalizability of classifiers based on neural networks is far from complete. The problem can be roughly formulated as follows. The random vector (X, Y) takes values in $\mathbb{R}^d \times \{0, 1\}$, and we have n independent, identically distributed samples $\{(X_1, Y_1), \ldots, (X_n, Y_n)\}$. The goal is to build a function $g : \mathbb{R}^d \to \{0, 1\}$ based on n samples such that the classification risk of g, $E[g(X) \neq Y]$, is minimal. The function $\eta(x) = E[Y|X = x]$ is called the regression function. It is well-known that the Bayes classifier defined by

$$g^*(\boldsymbol{x}) := \begin{cases} 1 & \text{if } \eta(x) \geq \frac{1}{2}; \\ 0 & \text{otherwise} \end{cases}$$

achieves the minimal classification risk, $L^* := E[g^*(X) \neq Y]$. Thus, it is natural to study the non-negative excess risk $E[g(X) \neq Y] - L^*$ of a classifier g as a measure of its performance.

The first classical result on classifiers based on neural networks is the paper (Faragó & Lugosi, 1993), which establishes two results. First, it shows that there exists a sequence of 1-hidden layer sigmoidal neural network classifiers whose widths grow in the order $o\left(\frac{n}{\log n}\right)$ such that their excess risks converge to 0 uniformly over all possible distributions, i.e., they are universally consistent. Second, for distributions whose regression function belongs to the Barron space (Barron, 1993), a wide class of functions for which shallow neural networks enjoy dimension-free approximation rate, it is shown that there exist neural network classifiers whose excess risks converge at a uniform rate $O(n^{-\frac{1}{4}})$.

However, the first result has room for improvement because it does not apply to deep or wide neural networks, and the proposed classifier is computationally infeasible. The second result on rates of convergence may be tightened in that there is no indication of whether the rate is tight in such a regime.

In practice, state-of-the-art neural networks have become increasingly more complex with number of parameters employed scaling to the order of hundreds of trillions. A very recent work (Radhakrishnan et al., 2023) studied the weak consistency of infinitely wide and deep neural networks using polynomial and sinusoidal activation functions, interpreting them as the neural tangent kernel (NTK) machines. However, they leave the question of weak consistency, let alone strong consistency, of finitely wide and deep neural networks as an open problem (see Section 3 in their paper). **Our result answers this question and provides a theoretical guarantee that for an arbitrary distribution, a computationally feasible sequence of classifiers based on deep and wide neural networks is strongly consistent**.

A number of recent results study classification problems with overparametrized deep neural networks. (Kim et al., 2021) shows that in the classical regimes characterized by Hölder-smoothness, neural networks that minimize the empirical risk of the hinge loss or logistic loss achieve competitive rates of convergence. (Bos & Schmidt-Hieber, 2022)

[1]H. Milton Stewart School of Industrial and Systems Engineering, Georgia Institute of Technology, Atlanta, United States. Correspondence to: Hyunouk Ko <hko39@gatech.edu>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024. Copyright 2024 by the author(s).

considers multi-class classification under the smooth compositional structural assumption on the regression function. Others derive convergence rates of convolutional neural networks optimizing the square loss (Kohler et al., 2022) and the logistic loss (Kohler & Langer, 2020).

A common assumption employed above and in the classical statistics literature including (Mammen & Tsybakov, 1999), (Tsybakov, 2004), (Audibert & Tsybakov, 2007), (Kerkyacharian et al., 2014), is to impose a smoothness assumption on the regression function (see Section 2.3 of (Suh & Cheng, 2024) for details). Then, they derive upper bounds on the convergence rate and in some regimes, prove minimax optimality by deriving a matching minimax lower bound.

We take a somewhat different view and **ask in what distributional regimes neural network classifiers are capable of achieving minimax optimal rates**. In doing so, we relax the smoothness assumption on the regression function and allow for a study of much more general classes of L^2 functions.

Specifically, we consider a family of L^2 functions with a finite Kolmogorov-Donoho optimal exponent, which is an information-theoretic number that quantifies the number of bits needed to construct an encoder-decoder pair that can approximate a given function class to a target accuracy (details in Section 1.3). **The significance of this characterization is that it applies to a much wider class of functions without explicit smoothness constraints, allowing for more realistic distributional settings**. A series of works from the past two decades ((Donoho et al., 1998),(Grohs et al., 2023),(Hinrichs et al., 2008),(Petersen & Voigtlaender, 2018)) have provided optimal exponents for many general classes of functions including L^p-Sobolev spaces, Besov spaces, bounded variation spaces, modulation spaces, and cartoon functions. Moreover, (Elbrächter et al., 2021) shows that most of these spaces are well-approximated by neural networks from the perspective of distortion theory (details in Section 1.3).

To put our work into context, we discuss some related works on the performance of neural network classifiers. A series of papers (Kohler et al., 2020), (Kohler & Langer, 2020), (Kohler et al., 2022), (Kohler & Walter, 2023), (Walter, 2023) study performance guarantee of empirical-risk-minimizing convolutional neural network classifiers under structural and smoothness assumptions on the regression function. Note that in these works, the optimization aspect of how the classifier is obtained is not considered which is also true of this paper. Some works do consider optimization procedures, albeit under stronger distributional assumptions. For example, (Frei et al., 2022) showed that two-layer neural networks with smoothed leaky ReLU activations trained with gradient descent exhibit exponentially fast convergence rates for distributions (roughly) with strongly log-concave covariate distribution and regression function whose norm is bounded by 1. (Cao et al., 2022) also derived exponential rates for convolutional neural networks under assumptions that imply the regression function is binary-valued. (Kou et al., 2023) obtained similar results for a slightly more general setting with ReLU convolutional neural networks. We emphasize that the distributional assumptions in these works are quite restrictive compared to our flexible distributional setting.

To summarize, we first show the universal consistency of wide and deep ReLU neural networks and second, give a characterization of some general classes of distributions for which neural network classifiers achieve minimax optimal rates of convergence.

0.1. Organization

In Section 1, we give a rigorous formulation of binary classification problems, provide definitions involving neural networks, and introduce basic concepts from Kolmogorov-Donoho approximation theory. In Section 2, we establish our first main result on the universal consistency of wide and deep ReLU neural networks. In Section 3, we give our second main result on rates of convergence for neural network classifiers for functions with Kolmogorov-Donoho optimal exponents and demonstrate with examples how the theorems may be applied to specific function spaces.

1. Preliminaries

We first give a rigorous formulation of the classification problem. Suppose we have $Z = (X, Y)$ and $Z_i = (X_i, Y_i), i = 1, 2, \ldots$ countably infinite, independent, identically distributed random vectors that map from a common probability space (Ω, Σ, P) to $[0,1]^d \times \{0,1\}$.

Fix a positive integer n. By a **classifier**, we mean a measurable function $g_n : [0,1]^d \times \{[0,1]^d \times \{0,1\}\}^n \to \{0,1\}$ where $[0,1]^d$ is endowed with the usual Borel σ-algebra it inherits from $\mathbb{R}^d$. Then, we can define

$$L(g_n) := P(g_n(X, Z_1, \ldots, Z_n) \neq Y | Z_1, \ldots, Z_n)$$

which is the conditional probability with respect to the σ-algebra generated by $Z_1, \ldots, Z_n$. Note that $L(g_n)$ is well-defined up to P-null set and is $\sigma(Z_1, \ldots, Z_n)$-measurable by the Radon-Nikodym theorem. For $n = 0$, we let $L_0 = L(g) = P(g(X) \neq Y)$ in the obvious way. We will be interested in $E[L(g_n)]$, the classification risk, as a measure of the performance of a classifier g_n.

Given a real-valued function $f : [0,1]^d \times \{[0,1]^d \times \{0,1\}\}^n \to \mathbb{R}$, the **plug-in classifier** corresponding to f

will be defined as:

$$p_f(\boldsymbol{x}) := \mathbb{1}_{\{x: f(x) \geq 1/2\}}(\boldsymbol{x}), \tag{1}$$

where for any subset $A \subset \mathbb{R}^d$, $\mathbb{1}_A$ is the indicator function defined by

$$\mathbb{1}_A(x) := \begin{cases} 1, & \text{if } x \in A; \\ 0, & \text{otherwise.} \end{cases} \tag{2}$$

When clear from context, we will also write $L(f) := L(p_f)$.

Denote by $\mathbb{N}$ the set of natural numbers $\{1, 2, \dots\}$. Any sequence of classifiers $\{g_n\}_{n\in\mathbb{N}}$ will be called a **classification rule**. A classification rule will be called weakly consistent if $L(g_n) \to L^*$ in probability (equivalently, $E[L(g_n)] \to L^*$) and strongly consistent if $L(g_n) \to L^*$ almost surely. Note these notions depend on the underlying probability measure P. We will call a classification rule universally weakly (strongly) consistent if for all probability measures P, the rule is weakly (strongly) consistent.

1.1. Notations

The symbols $\mathbb{Z}, \mathbb{R}$ denote the set of integers and real numbers respectively, and $\mathbb{R}_{>0}$ denotes the positive real numbers. For any $x \in \mathbb{R}$, we define $\lfloor x \rfloor := \max\{m \in \mathbb{Z} : m \leq x\}$. We write $L^p([0,1]^d, \mu)$ or $L^p(\mu)$ to denote the L^p space with respect to a positive Borel measure μ. This metric space has the usual $L^p(\mu)$-norm and has the corresponding metric topology. We write $C([0,1]^d)$ to denote the space of all continuous functions on $[0,1]^d$ equipped with the uniform norm, $\|f\|_u := \sup_{x\in[0,1]} |f(x)|$, and the usual norm topology. For an integer $k \geq 0$ and $0 < \beta \leq 1$, we define the Hölder space $C^{k,\beta} = C^{k,\beta}([0,1]^d)$ as the space of all k-times continuously differentiable functions on $[0,1]^d$ equipped with the norm:

$$\|f\|_{C^{k,\beta}} = \max\Bigg\{ \max_{\mathbf{k}:|\mathbf{k}|\leq k} \max_{\mathbf{x}\in[0,1]^d} |D^{\mathbf{k}} f(\mathbf{x})|, \max_{\mathbf{k}:|\mathbf{k}|=k} \sup_{\substack{\mathbf{x},\mathbf{y}\in[0,1]^d \\ \mathbf{x}\neq\mathbf{y}}} \frac{\left\|D^{\mathbf{k}} f(\mathbf{x}) - D^{\mathbf{k}} f(\mathbf{y})\right\|_2}{\|\mathbf{x}-\mathbf{y}\|_2^{\beta}} \Bigg\}.$$

For either a matrix $A \in \mathbb{R}^{m\times n}$ or a vector $v \in \mathbb{R}^n$, $\|A\|_{\max} := \max_{i=1,\dots,m} \max_{j=1,\dots,n} |A_{ij}|$ and $\|v\|_{\max} := \max_{i=1,\dots,n} |v_i|$ where the subscript notation refers to the indexed component of the matrix and vector. For a real-valued measurable function f whose domain is a measurable space (Ω, Σ, P), we write $P(f)$ to denote the integral of f with respect to P. For a probability measure P, we will write P_n to mean the empirical measure corresponding to n i.i.d. random variables with distribution P, $\frac{1}{n}\sum_{i=1}^n \delta_{X_i}(B)$ where $\delta_X(B) = \mathbb{1}_B(X)$ for any measurable set B. For a metric space S, $\mathcal{B}(S)$ denotes the Borel σ-algebra associated with S.

1.2. Neural networks

In this section, we rigorously define neural networks and their realization functions and equip the space with the right topology to obtain an adequate compactification of the space of neural networks.

Fix $L, N_0, \dots, N_L \in \mathbb{N}$. We define a **neural network** as the ordered set of matrix-vector tuples $\Phi = \{(A_l, b_l)\}_{l=1}^L$ where $A_l \in \mathbb{R}^{N_l \times N_{l-1}}$ and $b_l \in \mathbb{R}^{N_l}$. We call the ordered tuple $S = (L, N_0, \dots, N_L)$ the **architecture** of Φ. We define $\mathcal{NN}(S)$ to be the set of all neural networks with architecture S. We sometimes write $\mathcal{NN}_{d,1}(S)$ to make explicit the restriction that the $N_0 = d, N_L = 1$. That is two neural networks Φ_1, Φ_2 belong to the same $\mathcal{NN}(S)$ if and only if the dimensions of all the matrices and vectors defining them match. When a neural network Φ is given, we write $S(\Phi)$ to denote its architecture. In the rest of the paper, we will only be concerned with the case $N_0 = d, N_L = 1$.

Now let $\varrho : \mathbb{R} \to \mathbb{R}$ be the ReLU activation function $\varrho(x) := \max\{x, 0\}$. For a vector $v = (v_1, \dots, v_n) \in \mathbb{R}^n$, with a slight abuse of notation, we write $\varrho(v)$ to mean $\varrho(v) := (\varrho(v_1), \dots, \varrho(v_n)) \in \mathbb{R}^n$. Also, let $\mathcal{NN} := \bigcup_S \mathcal{NN}(S)$ where the union runs over all choices of valid architectures S. For a given set $\Omega \subset \mathbb{R}^{N_0}$, we can define the realization map of a neural network Φ as the map $R_\varrho^\Omega : \mathcal{NN} \to C(\Omega)$ where $R_\varrho(\Phi) : \Omega \to \mathbb{R}$ is defined in the following recursive fashion:

$$\begin{aligned} &R_\varrho(\Phi)(x) = x_L \text{ where} \\ &x_0 := x \\ &x_l := \varrho(A_l x_{l-1} + b_l), l = 1, \dots, L-1 \\ &x_L := A_L x_{L-1} + b_L. \end{aligned}$$

For a given architecture S, We will define the total number of neurons as $N(S) := \sum_{i=1}^L N_i$, and the number of layers as $L(S) := |S|$ where $|S|$ is the cardinality of S. Furthermore for a given $\Phi \in \mathcal{NN}(S)$, we define the following quantities that specify the complexity of Φ:

- the connectivity $\mathcal{M}(\Phi)$ denotes the total number of nonzero entries in the matrices $A_\ell, \ell \in \{1, 2, \dots, L\}$, and the vectors $b_\ell, \ell \in \{1, 2, \dots, L\}$,
- width $\mathcal{W}(\Phi) := \max_{\ell:0\leq\ell\leq L} N_\ell$,
- $\mathcal{L}(\Phi)$ is the total number of hidden layers in the architecture defining Φ,
- weight magnitude $\mathcal{B}(\Phi) := \max_{\ell:0\leq\ell\leq L} \max\left\{\|A_\ell\|_{\max}, \|b_\ell\|_{\max}\right\}$.

We also make $\mathcal{NN}(S)$ a finite-dimensional normed space by equipping it with the norm

$$\|\Phi\|_{\mathcal{NN}} := \max_{\ell:0\leq\ell\leq L} \|A_l\|_{\max} + \max_{\ell:0\leq\ell\leq L} \|b_l\|_{\max}.$$

For a fixed architecture S and a fixed choice of function $\pi : \mathbb{N} \to \mathbb{R}_{>0}$, we define $\mathcal{NN}_{d,1}^{\pi,S}(M)$ to be the class of neural networks with architecture S and whose weights are bounded by $\pi(M)$:

$$\mathcal{NN}_{d,1}^{\pi,S}(M) := \{\Phi \in \mathcal{NN}_{d,1}(S) : \mathcal{M}(\Phi) \leq M, \quad \mathcal{B}(\Phi) \leq \pi(M)\}. \tag{3}$$

for any $M > 0$.

It is possible to give a partial-order $\leq$ to the set of architectures by stipulating $S_1 \leq S_2$ for $S_1 = (L, N_1, \ldots, N_L)$ and $S_2 = (L', M_1, \ldots, M_{L'})$ if and only if $L \leq L'$ and $N_i \leq M_i$ for all $i = 1, \cdots, L$.

For the purposes of proving universal consistency in Section 2, we want to consider a method of sieves where we choose an estimator $\widehat{\theta}_n$ from $\mathcal{NN}_{d,1}^{\pi,S_n}(M_n)$ for a suitable choice of increasing sequence of architectures $\{S_n\}_{n\in\mathbb{N}}$ and real numbers $\{M_n\}_{n\in\mathbb{N}}$. Therefore, our neural networks will come from the set of a countable union:

$$\bigcup_{n=1}^{\infty} \mathcal{NN}_{d,1}^{\pi,S_n}(M_n).$$

We want to give this set a topology so that we have a compact space: this is necessary to apply Wald's method for proving consistency. Thus, we consider the following construction in the next two paragraphs.

For each $n \in \mathbb{N}$, let $d_n(\cdot,\cdot)$ be the metric on $\mathcal{NN}_{d,1}^{\pi,S_n}(M_n)$ induced by the norm $\|\cdot\|_{\mathcal{NN}}$. Then, define the disjoint union space:

$$\widetilde{\Theta} := \bigsqcup_{n=1}^{\infty} \mathcal{NN}_{d,1}^{\pi,S_n}(M_n) \tag{4}$$

with the disjoint union topology. This space is also metrizable and so normal. We can give an explicit metric that metrizes this topology: if we let D_n be the diameter of the space $\mathcal{NN}_{d,1}^{\pi,S_n}(M_n)$ for all $n \in \mathbb{N}$,

$$d(x,y) = \begin{cases} d_n(x,y), & \text{if } x, y \in \mathcal{NN}_{d,1}^{\pi,S_n}(M_n); \\ \max\{D_n, D_m\}, & \text{if } x \in \mathcal{NN}_{d,1}^{\pi,S_n}(M_n), \\ & y \in \mathcal{NN}_{d,1}^{\pi,S_m}(M_m), n \neq m \end{cases} \tag{5}$$

is such a metric (c.f. Example 2.6, Theorem 2.12 of (Sharma et al., 2020)). It is a second-countable, complete metric space. Since it is the disjoint union of countably many compact Hausdorff spaces, it is also a locally compact Hausdorff space.

The above construction ensures the existence of the Stone-Čech compactification of $\widetilde{\Theta}$, which we denote by Θ. Recall that the Stone-Čech compactification is characterized by the fact that Θ is a compact Hausdorff space containing $\widetilde{\Theta}$ as a dense subspace and that any continuous function $f : \widetilde{\Theta} \to C$ for any compact Hausdorff space C can be uniquely extended to a continuous function $\bar{f} : \Theta \to C$. This compactification is unique up to equivalence that identifies two compactifications Y_1, Y_2 of $\widetilde{\Theta}$ such that there exists a homeomorphism $h : Y_1 \to Y_2$ that is an identity when restricted to $\widetilde{\Theta}$. In fact, Θ is not metrizable because $\widetilde{\Theta}$ is non-compact. One point of caution is that while all points of $\Theta \backslash \widetilde{\Theta}$ are limit points of $\widetilde{\Theta}$ by definition of compactification, none of them are a (sequential) limit of any sequence of points from $\widetilde{\Theta}$.

It is not difficult to check that the realization mapping $R_\varrho : \widetilde{\Theta} \to C(\Omega)$ is continuous when $C(\Omega)$ is equipped with the uniform norm (for e.g., Proposition 4.1 of (Petersen et al., 2021)). For our analysis, we may assume without loss of generality that the realization mapping is followed by a projection to the unit ball in $C(\Omega)$, which we denote by $U(C(\Omega))$. This map is also continuous because the projection is achieved by mere scaling. Furthermore, we extend the domain of the realization mapping to Θ, which is possible by the characterizing property of Stone-Cech compactification.

1.3. Kolmogorov-Donoho approximation theory

In this subsection, we introduce the concepts from the Kolomogorov-Donoho approximation theory that appear in Section 3. In particular, we assume that the regression function belongs to a function class with an information-theoretic constraint.

Let $l \in \mathbb{N}$, $d \in \mathbb{N}$, $\Omega \subset \mathbb{R}^d$ such that Ω is Lebesgue measurable. In all that follows, we equip Ω with the Borel σ-algebra and the d-dimensional Lebesgue measure on it. Let $\mathcal{C}$ be a class of functions $\mathcal{C} \subset L^2(\Omega)$. First, define the set of encoders and the set of decoders as follows:

$$\mathcal{E}^l := \{E : \mathcal{C} \to \{0,1\}^l\},$$
$$\mathcal{D}^l := \{D : \{0,1\}^l \to \mathcal{C}\}.$$

Definition 1.1 (Kolmogorov-Donoho optimal exponent)**.** For each $\epsilon > 0$, let the minimax code length be defined as:

$$L(\epsilon, \mathcal{C}) := \min\{\ell \in \mathbb{N} : \exists (E,D) \in \mathcal{E}^\ell \times \mathcal{D}^\ell : . \sup_{f\in\mathcal{C}} \|D(E(f)) - f\|_{L^2(\Omega)} \leq \epsilon\}.$$

We define the (Kolmogorov-Donoho) optimal exponent of $\mathcal{C}$ as the real number

$$\gamma^*(\mathcal{C}) := \sup\{\gamma \in \mathbb{R} : L(\epsilon, \mathcal{C}) \in O(\epsilon^{-\frac{1}{\gamma}})\}.$$

The optimal exponent is known for L^p-Sobolev spaces, Besov spaces, modulation spaces, and Cartoon function classes as summarized in Table 1 of (Elbrächter et al., 2021).

There is a rich literature on the class of basis functions whose linear combinations can be used as approximators for these function spaces. That is, given a Hilbert space $\mathcal{H} = L^2(\Omega)$ for some bounded set $\Omega \subset \mathbb{R}^d$, we consider a countable family of functions in $\mathcal{H}$, called a dictionary and denoted $\mathcal{D} = \{\psi_i\}_{i\in\mathbb{N}}$, with which we approximate any function from $\mathcal{C} \subset \mathcal{H}$. We may measure the performance of $\mathcal{D}$ with respect to $\mathcal{C}$ with the following quantity:

$$\varepsilon^{\pi}_{\mathcal{C},\mathcal{D}}(M) := \sup_{f\in\mathcal{C}} \inf_{\substack{I_{f,M}\subset\{1,2,\ldots,\pi(M)\} \\ |I_{f,M}|=M, |c_i|\leq\pi(M)}} \left\| f - \sum_{i\in I_{f,M}} c_i\psi_i \right\|_{L^2(\Omega)} \tag{6}$$

where π denotes some given real polynomial. Then, one defines the effective best M-term approximation rate of $\mathcal{C}$ with dictionary $\mathcal{D}$ as:

Definition 1.2 (Effective best M-term approximation rate)**.**

$$\gamma^*(\mathcal{C},\mathcal{D}) := \sup\{\gamma \geq 0 : \exists \text{ polynomial } \pi \text{ such that } \varepsilon^{\pi}_{\mathcal{C},\mathcal{D}}(M) \in O(M^{-\gamma})\}.$$

A notable relationship between $\gamma^*(\mathcal{C})$ and $\gamma^*(\mathcal{C},\mathcal{D})$ is $\gamma^*(\mathcal{C},\mathcal{D}) \leq \gamma^*(\mathcal{C})$. Then, we say that $\mathcal{C}$ *is optimally representable by* $\mathcal{D}$ *if* $\gamma^*(\mathcal{C},\mathcal{D}) = \gamma^*(\mathcal{C})$. Many function spaces usually studied in the approximation theory literature are, in fact, optimally representable by well-known dictionaries such as those based on the Fourier/wavelet basis and the Haar basis.

There is a natural corresponding concept for the class of neural networks as a replacement for dictionaries. Recalling the definition (3), we will define the union of all neural networks whose architecture has depth bounded by $\pi(M)$ for a given function π. Specifically,

$$\mathcal{NN}^{\pi}_{d,1}(M) := \bigcup_{S:L(S)\leq\pi(\log M)} \mathcal{NN}^{\pi,S}_{d,1}(M).$$

Note for this definition, we don't care about the topology on this set at this point.

Similar to the effective best approximation error $\varepsilon^{\pi}_{\mathcal{C},\mathcal{D}}(M)$, defined with respect to the dictionary $\mathcal{D}$, we define the effective best approximation with neural networks as follows:

$$\varepsilon^{\pi}_{\mathcal{N}}(M) := \sup_{f\in\mathcal{C}} \inf_{\Phi\in\mathcal{NN}^{\pi}_{d,1}(M)} \|f - R_\varrho(\Phi)\|_{L^2(\Omega)}. \tag{7}$$

Just as we did for the dictionary $\mathcal{D}$, we define the best effective M-term approximation rate as follows:

Definition 1.3 (Effective best M-weight approximation rate)**.**

$$\gamma^*_{\mathcal{N}}(\mathcal{C}) := \sup\{\gamma \geq 0 : \exists \text{ polynomial } \pi \text{ such that } \varepsilon^{\pi}_{\mathcal{N}}(M) \in O(M^{-\gamma}), M \to \infty\}.$$

This means if $\gamma^*_{\mathcal{N}}(\mathcal{C}) > 0$, the L^2 approximation error decays at least polynomially in the connectivity of the approximating neural networks. Furthermore, it is shown in Theorem VI.4 of (Elbrächter et al., 2021) that $\gamma^*_{\mathcal{N}}(\mathcal{C}) \leq \gamma^*(\mathcal{C})$, which makes the following definition natural:

Definition 1.4. We say that $\mathcal{C} \subset L^2(\Omega)$ is optimally representable by neural networks if

$$\gamma^*_{\mathcal{N}}(\mathcal{C}) = \gamma^*(\mathcal{C}).$$

Quite general classes of functions are optimally representable by neural networks including the Besov spaces and the modulation spaces. These results follow from the "transference principle" which shows that $\gamma^*(\mathcal{C},\mathcal{D}) \leq \gamma^*_{\mathcal{N}}(\mathcal{C})$ for most useful dictionaries that optimally represent classical function spaces.

2. Universal consistency

In this section, we state our first result on the universal consistency of wide and deep ReLU neural network classifiers.

We will need the following lemma to establish that the empirical risk minimizer is well-defined as a classifier. Its proof is relegated to Appendix B.1.

Lemma 2.1. *Let $(A, \mathcal{A})$ be a measurable space and B a compact, metrizable topological space. Assume $m(\cdot,\cdot) : A \times B \to \mathbb{R}$ is measurable in the first argument and continuous in the second argument. Then, there exists a Borel measurable mapping $\widehat{f} : A \to B$ that satisfies $m(a, \widehat{f}(a)) = \sup_{b\in B} m(a, b)$ is Borel measurable.*

Now, we state our main theorem on the universal consistency of wide and deep ReLU neural networks. Its proof is relegated to Appendix B.2.

Theorem 2.2. *Let $\{S_n\}_{n\in\mathbb{N}}$ be an increasing sequence of architectures such that $W(S_n) \geq n$ or $L(S_n) \geq n$ for all $n \in \mathbb{N}$. There exists some increasing function π and constant c_d only depending on d such that the empirical risk minimizer of the logistic loss on $\mathcal{NN}^{\pi,S_n}_{d,1}(c_d n)$ for each $n \in \mathbb{N}$ defined as:*

$$\widehat{\theta}_n := \operatorname*{arg\,min}_{\theta\in\mathcal{NN}^{\pi,S_n}_{d,1}(n)} \frac{1}{n}\sum_{i=1}^{n} l(R_\varrho(\theta)(X_i), Y_i) \tag{8}$$

is universally strongly consistent:

$$\lim_{n\to\infty} L(R_\varrho(\widehat{\theta}_n) + 1/2) \to L^* \text{ with probability } 1.$$

Remark 2.3. As can be seen from the proof, the only property that we require of the surrogate loss is that its empirical minimizer in $\mathcal{NN}^{\pi,S_n}_{d,1}(c_d n)$ achieves 0 classification loss. The same conclusion holds for any other continuous loss function with such property.

As noted in the Introduction, this result answers the open problem mentioned in (Radhakrishnan et al., 2023). In fact, the classifiers in Theorem 2.2 are interpolating classifiers, i.e., they correctly classify all training points, and are also feasible as they are the minimizers of a convex surrogate loss.

3. Rates of convergence

The second question of interest, which is more practically relevant, is what upper bounds we can establish on the excess risk of the empirical risk minimizer (8) as a function of n that is independent of any individual choice of the underlying distribution, i.e., we want to establish a uniform (in the set of probability measures) rate of convergence. It is well known that no universal rates that hold for all probability distributions are possible (c.f. Theorem 7.2 of (Devroye et al., 2013)).

This means that we must have some restrictions on the set of possible P. Observing that the joint distribution of (X, Y) on $[0,1]^d \times \{0,1\}$ is fully determined by the specification of $E[Y|X]$ and the marginal measure μ_X on $[0,1]^d$, we take the view of considering all P such that the regression function belongs to some given model class of functions and the marginal law of X satisfies certain regularity conditions.

What model classes are suitable and interesting for practical relevance is in itself an important question. As noted in the Introduction, smoothness assumptions are most widely used. We generalize the landscape of classification theory by taking advantage of how well neural networks can approximate the most useful dictionaries.

Our program will work with the usual decomposition of the excess risk in terms of estimation and approximation error:

$$\mathcal{E}(\widehat{f}_n) = \underbrace{E[L_n] - \inf_{f \in \mathcal{F}_n} E[L(f)]}_{①} + \underbrace{\inf_{f \in \mathcal{F}_n} E[L(f)] - L^*}_{②}$$

where term ① comprises the estimation error, and we will rely on empirical risk minimization and more fundamentally, empirical process theory to control this error. Term ② comprises the approximation error, and we control it by proposing suitable classes of neural networks that well-approximate the regression function n L^p (c.f. Section A).

3.1. Distributional assumptions

For our results on uniform convergence rates, we will make the following three assumptions:

Assumption 3.1. (Tsybakov noise condition) We assume there exist constants $C_0 > 0$ and $\alpha \geq 0$ such that

$$P_X(0 < |\eta(x) - 1/2| \leq t) \leq C_0 t^\alpha, \quad \forall t > 0. \quad (9)$$

Remark 3.2. This assumption is used widely in the literature and controls the concentration of measure near the optimal decision boundary. The assumption becomes vacuous for $\alpha = 0$ and the case $\alpha = \infty$ corresponds to a strict margin condition.

Assumption 3.3. We assume that the distribution of X admits an L^2 density with respect to the n-dimensional Lebesgue measure restricted to $[0,1]^d$ that is uniformly bounded by some constant.

Remark 3.4. While we have adopted the Lebesgue measure as the dominating measure of P to take advantage of the known approximation results, we believe the approximation theory can be generalized to arbitrary σ-finite measures.

Assumption 3.5. We assume that the regression function belongs to some class of functions $\mathcal{F} \subset L^2([0,1]^d)$ with a finite Kolmogorov-Donoho optimal exponent $\gamma^*(\mathcal{F}) > 0$.

3.2. Convergence rates

In this section, we give our second main results that characterize sufficient conditions for a set of probability measures under which neural network classifiers achieve minimax optimality.

There is a somewhat subtle relationship between regression and classification, and we relegate a detailed discussion on this relationship to Appendix A. For now, we remark that while L^p consistency is a sufficient but not a necessary condition for the consistency of the corresponding plug-in classification rule (pointwise regime), the convergence rate for the pth power of L^p norm in the minimax sense for some classical function spaces may agree with the minimax rate of convergence for the classification risk.

We also remark that the observation of (Audibert & Tsybakov, 2007) in the paragraph after Lemma 5.2 is somewhat misleading: the paper claims that deriving convergence rates for classification risk based on L^2 risk is not the right tool in the presence of Tsybakov noise condition. Specifically, under a suitable regime, for some constant $c > 0$,

$$\liminf_{n \to \infty} \inf_{T_n} \sup_{f \in \Sigma(\beta, L)} \mathbf{E}_f \left[n^{\frac{2\beta}{2\beta+d}} \|T_n - f\|_2^2 \right] \geq c,$$

where the infimum is over all possible estimators and $\Sigma(\beta, L)$ is the L-Hölder ball of functions, which then implies that inequality (11) (in Appendix A) only leads to suboptimal rates for the classification risk. However, while $n^{-\frac{2\beta}{2\beta+d}}$ is certainly the best possible rate for the square of L^2 risk in the above sense, it is only so when the infimum is taken over an estimator sequence (T_n's), not deterministic functions.

The approach using estimation and approximation error decomposition, on the other hand, allows us to fully use the approximation power of realizations of neural networks that lead to minimax optimal rates even in the presence of the Tsybakov noise condition.

Now, we state our result on the convergence rates of neural network classifiers in the framework of function classes with finite Kolmogorov-Donoho optimal exponents. For the following result, the empirical risk minimization is taken with respect to the classification loss. Its proof is relegated to Appendix B.3.

Theorem 3.6. *Let $\mathcal{F}$ be a compact subset of $L^2([0,1]^d)$ with Kolmogorov-Donoho optimal exponent $\gamma^* > 0$ that is optimally representable by neural networks with polynomial π. Let $\mathcal{P}_{\mathcal{F}}$ be a given class of distributions satisfying Assumption 3.1, Assumption 3.3, and Assumption 3.5 (with above $\mathcal{F}$). Define the optimal minimax rate of convergence for $\mathcal{P}_{\mathcal{F}}$ as follows:*

$$m^* := \inf\Big\{ m \in \mathbb{R}_+ : m \text{ satisfies } \inf_{g_n} \sup_{P \in \mathcal{P}_{\mathcal{F}}} E[L(g_n)] - L^* = \Omega(n^{-m}) \Big\}.$$

Additionally, assume that α, γ^*, m^* satisfy

$$2(1+\alpha)\gamma^*(1-m^*) \geq (2+\alpha)m^*. \tag{10}$$

Define

$$\mathcal{NN}_n := \mathcal{NN}^{\pi}_{d,1}(C_{d,\alpha,m^*,\gamma^*} n^{\frac{(2+\alpha)m^*}{2(1+\alpha)\gamma^*}}).$$

where $C_{d,\alpha,m^*,\gamma^*}$ is a constant that only depends on d, α, m^*, γ^* (see Definition 1.1). Let $\widehat{\theta}_n$ be the empirical risk minimizer of the classification loss:

$$\widehat{\theta}_n := \arg\min_{\theta \in \mathcal{NN}_n} \frac{1}{n} \sum_{i=1}^{n} P(p_{R_\varrho(\theta)}(X_i) \neq Y_i)$$

Then the plug-in classification rule based on $\{R_\varrho(\widehat{\theta}_n)\}_{n\in\mathbb{N}}$ achieves minimax optimal (up to polylogarithmic factor) rate of convergence for the excess classification risk.

Remark 3.7. Note m^* may depend on α and $\mathcal{F}$. Condition (10) is not very stringent for many classical function spaces: Examples of classical regimes in which (10) holds will be provided in Section 3.3. In fact, the condition turns out to be vacuous for the space of Besov functions.

Remark 3.8. Suppose $\mathcal{P}_{\mathcal{F}}$ is such that the minimax rate of L^2-risk matches that of classification risk in the sense of (12) and the rate is given by n^{-m^*}. Theorem 3.6 shows that this optimal rate is still achieved if the infimum on the right-hand side of (12) is taken over all $f_n \in \mathcal{F}_n$ instead. Does this mean that we also get the same rate if we replace the left-hand side of (12) by $f_n \in \mathcal{F}_n$? Because $\mathcal{F}$ is optimally representable by neural networks with exponent γ^*, for any constant $C > 0$ and any $m > \frac{(2+\alpha)m^*}{2(1+\alpha)}$ (in particular, $m = m^*$),

$$Cn^{-m} < \sup_{P \in \mathcal{P}} \inf_{f \in \mathcal{F}_n} (E_n[\|f - \eta\|_2^2])^{\frac{1}{2}}$$

happens infinitely often as $n \to \infty$. Because we have

$$\sup_{P \in \mathcal{P}_{\mathcal{F}}} \inf_{f \in \mathcal{F}_n} (E_n[\|f - \eta\|_2^2])^{\frac{1}{2}} \leq \inf_{f_n \in \mathcal{F}_n} \sup_{P \in \mathcal{P}_{\mathcal{F}}} (E_n[\|f_n - \eta\|_2^2])^{\frac{1}{2}},$$

we conclude that the answer is no.

3.3. Two examples

We demonstrate two applications of Theorem 3.6 to classical function spaces whose Kolmogorov-Donoho optimal exponents are known and are optimally representable by neural networks.

3.3.1. Hölder functions

For a real number $\beta \geq 1$, let $m = \lfloor \beta \rfloor$. We define Hölder class $C^\beta([0,1]) := C^{m,\beta-m}([0,1])$ following the definition in Section 1.1. We take $\mathcal{F}$ to be the unit ball of Hölder functions. The Kolomogrov-Donoho optimal exponent is given by $\gamma^* = \beta$ and it is optimally representable by neural networks (Elbrächter et al., 2021). Under certain regularity conditions (Definition 2.2 of (Audibert & Tsybakov, 2007)) on the marginal distribution of X that is stronger than Assumption 3.3, the minimax optimal rate is given by $m^* = \frac{\beta(1+\alpha)}{2\beta+d}$. Then, it suffices to check assumptions (10) which translates to

$$\beta - 1 \geq \frac{\alpha}{2}(1 + 2\beta).$$

This shows that for "difficult" problems ($\alpha < 1, \beta > 1$), the proposed neural network classification rules from Theorem 3.6 achieves minimax optimal rate of convergence.

3.3.2. Besov functions

We take $\mathcal{F}$ to be the unit ball of the Besov class $B^m_{2,q}([0,1]^d) \subset L^2([0,1]^d)$ of Besov functions (see Chapter 4.3 of (Giné & Nickl, 2021) for a definition and basic properties). Then, $\gamma^* = \frac{m}{d}$ as shown in Theorem 1.3 of (Grohs et al., 2023). Under the assumption that the density of marginal distribution of X is upper bounded by a constant larger than 1, which is clearly implied by Assumption 3.3, we have $m^* = \frac{m}{2m+d}$ as long as $\alpha = 0$ (making Assumption 3.1 null), $\frac{m}{d} > \frac{1}{q} - \frac{1}{2}$ and $1 \leq q \leq \infty$ (see page 2278 of (Yang, 1999)). Assumption (10) translates to

$$2(m + d) \geq 2d.$$

Note Assumption (10) is vacuous in this case. This implies that the conclusion of Theorem 3.6 holds for all choices of α, d, q satisfying $\frac{m}{d} > \frac{1}{q} - \frac{1}{2}$.

Acknowledgements

We are grateful to the anonymous reviewers for their helpful comments that improved this paper. The authors are partially sponsored by NSF grants DMS 2015363, 2229876, and the A. Russell Chandler III Professorship at Georgia Tech.

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

References

Audibert, J.-Y. and Tsybakov, A. B. Fast learning rates for plug-in classifiers. *The Annals of Statistics*, 35(2):608 – 633, 2007. doi: 10.1214/009053606000001217. URL https://doi.org/10.1214/009053606000001217.

Barron, A. R. Universal approximation bounds for superpositions of a sigmoidal function. *IEEE Transactions on Information theory*, 39(3):930–945, 1993.

Bartlett, P. L., Harvey, N., Liaw, C., and Mehrabian, A. Nearly-tight vc-dimension and pseudodimension bounds for piecewise linear neural networks. *The Journal of Machine Learning Research*, 20(1):2285–2301, 2019.

Bos, T. and Schmidt-Hieber, J. Convergence rates of deep relu networks for multiclass classification. *Electronic Journal of Statistics*, 16(1):2724–2773, 2022.

Boucheron, S., Bousquet, O., and Lugosi, G. Theory of classification: A survey of some recent advances. *ESAIM: probability and statistics*, 9:323–375, 2005.

Cao, Y., Chen, Z., Belkin, M., and Gu, Q. Benign overfitting in two-layer convolutional neural networks. *Advances in neural information processing systems*, 35:25237–25250, 2022.

DeVore, R., Hanin, B., and Petrova, G. Neural network approximation. *Acta Numerica*, 30:327–444, 2021.

Devroye, L., Györfi, L., and Lugosi, G. *A probabilistic theory of pattern recognition*, volume 31. Springer Science & Business Media, 2013.

Donoho, D. L., Vetterli, M., DeVore, R. A., and Daubechies, I. Data compression and harmonic analysis. *IEEE transactions on information theory*, 44(6):2435–2476, 1998.

Elbrächter, D., Perekrestenko, D., Grohs, P., and Bölcskei, H. Deep neural network approximation theory. *IEEE Transactions on Information Theory*, 67(5):2581–2623, 2021.

Faragó, A. and Lugosi, G. Strong universal consistency of neural network classifiers. *IEEE Transactions on Information Theory*, 39(4):1146–1151, 1993.

Frei, S., Chatterji, N. S., and Bartlett, P. Benign overfitting without linearity: Neural network classifiers trained by gradient descent for noisy linear data. In Loh, P.-L. and Raginsky, M. (eds.), *Proceedings of Thirty Fifth Conference on Learning Theory*, volume 178 of *Proceedings of Machine Learning Research*, pp. 2668–2703. PMLR, 02–05 Jul 2022. URL https://proceedings.mlr.press/v178/frei22a.html.

Giné, E. and Nickl, R. *Mathematical foundations of infinite-dimensional statistical models*. Cambridge university press, 2021.

Grohs, P., Klotz, A., and Voigtlaender, F. Phase transitions in rate distortion theory and deep learning. *Foundations of Computational Mathematics*, 23(1):329–392, 2023.

Hinrichs, A., Piotrowska, I., Piotrowski, M., et al. On the degree of compactness of embeddings between weighted modulation spaces. *Journal of Function Spaces*, 6:303–317, 2008.

Kerkyacharian, G., Tsybakov, A. B., Temlyakov, V., Picard, D., and Koltchinskii, V. Optimal exponential bounds on the accuracy of classification. *Constructive Approximation*, 39:421–444, 2014.

Kim, Y., Ohn, I., and Kim, D. Fast convergence rates of deep neural networks for classification. *Neural Networks*, 138:179–197, 2021.

Kohler, M. and Langer, S. Statistical theory for image classification using deep convolutional neural networks with cross-entropy loss. *arXiv preprint arXiv:2011.13602*, 2020.

Kohler, M. and Walter, B. Analysis of convolutional neural network image classifiers in a rotationally symmetric model. *IEEE Transactions on Information Theory*, 2023.

Kohler, M., Krzyzak, A., and Walter, B. On the rate of convergence of image classifiers based on convolutional neural networks. *arXiv preprint arXiv:2003.01526*, 2020.

Kohler, M., Krzyżak, A., and Walter, B. On the rate of convergence of image classifiers based on convolutional neural networks. *Annals of the Institute of Statistical Mathematics*, 74(6):1085–1108, 2022.

Koltchinskii, V. *Oracle inequalities in empirical risk minimization and sparse recovery problems: École D'Été de Probabilités de Saint-Flour XXXVIII-2008*, volume 2033. Springer Science & Business Media, 2011.

Kou, Y., Chen, Z., Chen, Y., and Gu, Q. Benign overfitting in two-layer relu convolutional neural networks. In *International Conference on Machine Learning*, pp. 17615–17659. PMLR, 2023.

Mammen, E. and Tsybakov, A. B. Smooth discrimination analysis. *The Annals of Statistics*, 27(6):1808–1829, 1999.

Petersen, P. and Voigtlaender, F. Optimal approximation of piecewise smooth functions using deep relu neural networks. *Neural Networks*, 108:296–330, 2018.

Petersen, P., Raslan, M., and Voigtlaender, F. Topological properties of the set of functions generated by neural networks of fixed size. *Foundations of computational mathematics*, 21:375–444, 2021.

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

Radhakrishnan, A., Belkin, M., and Uhler, C. Wide and deep neural networks achieve consistency for classification. *Proceedings of the National Academy of Sciences*, 120 (14):e2208779120, 2023.

Sharma, R. P., Goyal, N., et al. Disjoint union metric and topological spaces. *Southeast Asian Bulletin of Mathematics*, 44(5), 2020.

Suh, N. and Cheng, G. A survey on statistical theory of deep learning: Approximation, training dynamics, and generative models, 2024.

Tsybakov, A. B. Optimal aggregation of classifiers in statistical learning. *The Annals of Statistics*, 32(1):135–166, 2004.

Walter, B. Analysis of convolutional neural network image classifiers in a hierarchical max-pooling model with additional local pooling. *Journal of Statistical Planning and Inference*, 224:109–126, 2023.

Yang, Y. Minimax nonparametric classification. I. rates of convergence. *IEEE Transactions on Information Theory*, 45(7):2271–2284, 1999.

A. Discussion on the relationship between regression and classification

Here we review some well-known results on the connection between regression and classification and discuss some subtleties in the minimax regime. In this discussion, the domain of X will be $\mathbb{R}^d$ instead of $[0,1]^d$.

Denote by E_n the expectation with respect to the distribution of $Z_1,\dots,Z_n$ and μ_X the distribution on $\mathbb{R}^d$ induced by P and X. In the following, assume that $\{g_n\}_{n\in\mathbb{N}}$ is a plug-in classification rule based on real-valued function sequence $\{f_n\}_{n\in\mathbb{N}}$. We can appeal to Fubini's theorem since all measures are finite and functions are bounded and deduce the following:

$$\begin{aligned}
E[L(g_n)]-L^* &= E_n[L(g_n)-L(g^*)] \\
&= E_n[E[\mathbb{1}_{g_n(X,Z_1,\dots,Z_n)\neq Y}-\mathbb{1}_{g^*(X)\neq Y}|Z_1,\dots,Z_n]] \\
&= E_n\left[\int_{\mathbb{R}^d}\eta(x)\left(\mathbb{1}_{g_n(\cdot,Z_1,\dots,Z_n)=0}(x)-\mathbb{1}_{g^*(\cdot)=0}(x)\right)\mu_X(dx)\right] \\
&\quad + E_n\left[\int_{\mathbb{R}^d}(1-\eta(x))\left(\mathbb{1}_{g_n(\cdot,Z_1,\dots,Z_n)=1}(x)-\mathbb{1}_{g^*(\cdot)=1}(x)\mu_X(dx)\right)\right] \\
&= E_n\left[\int_{\mathbb{R}^d}|2\eta(x)-1|\mathbb{1}_{g_n(\cdot,Z_1,\dots,Z_n)\neq g^*(\cdot)}(x)\mu_X(dx)\right] \\
&\le E_n\left[\int_{\mathbb{R}^d}2|\eta(x)-f_n(x,Z_1,\dots,Z_n)|\mu_X(dx)\right] \\
&\le 2E_n\left[\sqrt[p]{\int_{\mathbb{R}^d}|\eta(x)-f_n(x,Z_1,\dots,Z_n)|^p\mu_X(dx)}\right]
\end{aligned}$$

for any $p\ge 1$ where the second to last inequality follows from the observation that for x such that $g_n(x,X_1,\dots,X_n)\neq g^*(x)$, we must have either $f_n(x,X_1,\dots,X_n)<\frac{1}{2}\le\eta(x)$ or $\eta(x)<\frac{1}{2}\le f_n(x,X_1,\dots,X_n)$ so that $|\eta(x)-\frac{1}{2}|\le|\eta(x)-f_n(x,X_1,\dots,X_n)|$, and the last inequality follows from Hölder's inequality. In view of the above inequality, fixing $z_1,\dots,z_n$, we may consider $\widehat{f}_n := f_n(\cdot,z_1,\dots,z_n):\mathbb{R}^d\to\mathbb{R}$ as an approximating function of true η corresponding to some unknown P in the L^p sense, and obtain a convergence rate for the excess risk from that of $E_n[\|\eta-f_n(\cdot,Z_1,\dots,Z_n)\|_{L^p(\mathbb{R}^d,\mu)}]$ for some integer $p\ge 1$. By abuse of notation, we will write this also as $E_n[\|\eta-f_n\|_p]:=E_n[\|\eta-f_n\|_{L^p(\mathbb{R}^d,\mu)}]$, omitting the dependence of f_n on $Z_1,\dots,Z_n$.

More can be said if $p>1$. For a fixed P, if $\|\eta-f_n\|_p\to 0$ in probability, we have $\rho_n(P):=\frac{E[L(g_n)]-L^*}{E_n[\|\eta-f_n\|_p]}\to 0$ as $n\to\infty$, which means the excess risk converges to 0 faster than the L^p-risk (Theorem 6.5 of (Devroye et al., 2013)). In this sense, classification is easier than regression. Then, a natural question to ask is what can be said about the convergence rate of this ratio. One answer is that no universal (in both P and estimator sequence) bound is possible on this ratio: precisely, given any sequence of numbers converging to 0 arbitrarily slowly, one can construct some P, and a rule g_n based on f_n such that $\|\eta-f_n\|_p\to 0$ in probability holds, but the ratio $\frac{E[L(g_n)]-L^*}{E_n[\|\eta-f_n\|_p]}$ approaches 0 as slow as the given sequence (see Chapter 6 of (Devroye et al., 2013)).

On the other hand, if one assumes that either η is bounded away from $\frac{1}{2}$ or $L^*=0$, which is a favorable situation for classification, the excess risk can be shown to converge to 0 at least as fast the pth power of the L^p-risk (which is smaller than the L^p-risk):

$$\frac{E[L(g_n)]-L^*}{E_n[\|\eta-f_n\|_p^p]}=O(1).$$

Under a less stringent condition than requiring η be bounded away from $1/2$, known as the Tsybakov noise condition parametrized by C_0,α (see (9)), we have for $1\le p<\infty$,

$$\frac{E[L(g_n)]-L^*}{E_n\left[\|\eta-f_n\|_p^{\frac{p(1+\alpha)}{p+\alpha}}\right]}\le C \tag{11}$$

where C only depends on C_0,α,p. See Lemma 5.2 of(Audibert & Tsybakov, 2007) for a proof.

The discussion in the previous paragraph allows one to derive uniform convergence rates from the approximation properties of f_n for η. While it is well-known that no universal convergence rates are possible, if we restrict η to belong to some

known family of functions that can be uniformly approximated by a certain class of functions, uniform convergence rates are attainable. This is the view we worked with when deriving convergence rate results in Section 3.

There is one sense in which the convergence rate of $E[L(g_n)] - L^*$ matches that of $(E_n[\|\eta - f_n\|_p^p])^{\frac{1}{p}}$. It is shown in (Yang, 1999) that the minimax rates of L^2 risk for a certain class of distributions (characterized by nonparametric classes of functions and regularity conditions on the marginal distribution of X) decay to 0 at the same rate as the minimax rate of the excess risk. Precisely, for some class of probability measures, denoted $\mathcal{P}$,

$$\inf_{f_n} \sup_{P \in \mathcal{P}} (E_n[\|f_n - \eta\|_2^2])^{\frac{1}{2}} \approx \inf_{g_n} \sup_{P \in \mathcal{P}} E[L(g_n)] - L^*. \tag{12}$$

where the infimum on the left-hand side is taken over all measurable real-valued functions and the infimum on the right-hand side is taken over all plug-in classifiers.

It is important to observe a key difference from the discussion of the preceding paragraph where we compared the L^2-risk associated with a real-valued function f with the classification risk of the plug-in rule associated with the same f (pointwise comparison): in contrast, the classifier that achieves (or nearly so) the infimum of the right-hand side of (12) is not necessarily that formed as a plug-in rule of the function that achieves (or nearly so) the infimum of the left-hand side of (12).

The lesson is that in this uniform regime of minimax risk, we observe a different asymptotic connection between classification and regression than in the pointwise regime: while in the pointwise regime, $E[L(g_n)] - L^*$ converges at least as fast as $E_n[\|f_n - \eta\|_2^2]$, which implies faster rate than $(E_n[\|f_n - \eta\|_2^2])^{\frac{1}{2}}$ since f_n, η can be assumed to be bounded by 1, in the minimax sense, $E[L(g_n)] - L^*$ converges at the same speed as $(E_n[\|f_n - \eta\|_2^2])^{\frac{1}{2}}$.

B. Proofs

B.1. Proof of Lemma 2.1

Proof. Under the axiom of countable choice, B is second-countable. Furthermore, it is normal as it is metrizable. Then, we use the fact that a regular, second-countable space can be embedded as a subspace of $\mathbb{R}^{\mathbb{N}}$ with the product topology. The image of this embedding is compact since B is. From now on, we make this identification up to homeomorphism.

Let $\{f_1, f_2, \dots\}$ be a dense set in B and fix $a \in A$. Define $\tilde{m} : B \to \mathbb{R}$ as $\tilde{m}(f) := \inf\{m(a, f) - m(a, f_n), n \in \mathbb{N}\}$, which is upper-semicontinuous. Then, any $\tilde{f}$ satisfies $m(a, \tilde{f}) = \sup_{f \in B} m(a, f)$ if and only if $\tilde{m}(\tilde{f}) = 0$. This shows that for each fixed a, the set of maximizers of $m(a, \cdot)$ is given by $B_0 := \tilde{m}^{-1}(0) = \tilde{m}^{-1}([0, \infty))$, which is closed and hence compact in $\mathbb{R}^{\mathbb{N}}$. Now let $\pi_n : \mathbb{R}^{\mathbb{N}} \to \mathbb{R}$ be the projection onto the nth coordinate. Then, $\pi_1(B_0)$ is compact in $\mathbb{R}$ so it has a maximum element, say v_1. Let $B_1 := \pi_1^{-1}(v_1) \cap B_0$, which is clearly non-empty and compact. Then proceed inductively, so that we obtain we obtain a sequence of decreasing sets $B_1 \supset B_2 \supset \dots$ Then, the set $\cap_{n=1}^{\infty} B_n$ is non-empty since each finite intersection is non-empty. Now, if any two elements are in this set, by construction they agree on all the coordinates so they are equal. This shows there is a maximum element $\widehat{f}(a) \in B_0$ in the dictionary order over $\mathbb{R}^{\mathbb{N}}$. Thus, for each $a \in A$, we can assign such $\widehat{f}(a)$ to obtain a well-defined mapping from A to B. It only remains to show this map is measurable.

It suffices to show that each $a \mapsto \pi_i(\widehat{f}(a))$ is measurable for all $i \in \mathbb{N}$. Fix a closed interval $[u, v] \subset \mathbb{R}$ for this. We can consider the function $g_{[u,v]} : A \to \mathbb{R}$ defined by $g_{[u,v]}(a) = \sup\{m(a, f_n) : n \in \mathbb{N}\} - \sup\{m(a, f_n) : \pi_i(f_n) \in [u, v], n \in \mathbb{N}\}$. This function is Borel measurable as both infimums are taken only over countably many measurable functions. Then, from the observation that $(\pi_i \circ \widehat{f})^{-1}([u, v]) = g_{[u,v]}^{-1}(0)$ we can conclude that indeed $\pi_i \circ \widehat{f}$ is Borel measurable. □

B.2. Proof of Theorem 2.2

Proof. First, we check there are no existence and measurability issues in (8). Suppose some π, c_d are given (for now). If we regard $z^n = \{X_i, Y_i\}_{i=1,\dots,n}$ as fixed numbers, clearly there is some $\theta \in \mathcal{NN}_{d,1}^{\pi, S_n}(c_d n)$ achieving the minimum in (8) by continuity of the associated maps and compactness of $\mathcal{NN}_{d,1}^{\pi, S_n}(c_d n)$. Denote any choice of such θ for z^n as θ_{z^n}. Moreover, for $\mathcal{Z} = [0, 1]^d \times \{0, 1\}$, Lemma 2.1 gives the existence of a Borel-measurable function $\widehat{\theta}_n : (\mathcal{Z}^n, \mathcal{B}(\mathcal{Z}^n)) \to \mathcal{NN}_{d,1}^{\pi, S_n}(c_d n)$ such that $\widehat{\theta}_n(z^n) = \theta_{z^n}$.

We now claim that there are some π, c_d such that $P_n M_{\widehat{\theta}_n} = 0$ so that $P_n M_{\widehat{\theta}_n} \geq P_n M_\theta$ for all $\theta \in \mathcal{NN}_{d,1}^{\pi, S_n}(c_d n)$. In other words, for each n, the empirical risk minimizer of the logistic loss achieves perfect classification accuracy for the n points.

This follows from the fact that there exists some π, which may be assumed to be increasing, such that there is a realization of some $\tilde{\theta} \in \mathcal{NN}_{d,1}^{\pi,S_n}(c_d n)$ such that

$$l(R_\varrho(\tilde{\theta})(X_i), Y_i) \leq \frac{\log 2}{n}, \quad i = 1, \ldots, n. \tag{13}$$

Such $\widetilde{\theta}$ can be taken to be either a 1 hidden-layer ReLU neural network with width n (see Theorem 5.1 of (Pinkus, 1999)) or a ReLU neural network with width 3 and $n-1$ hidden-layers (see Proposition 3.10 of (DeVore et al., 2021)). This observation and the definition of $\widehat{\theta}_n$ implies the claim $P_n M_{\widehat{\theta}_n} = 0$.
Fix any $\epsilon > 0$. Let

$$\mathcal{F}_0 := \{f : [0,1]^d \to \mathbb{R} : f \text{ is measurable and } f(X) \text{ is a version of } E[Y|X]\}.$$

Let $M^* := -L^*$, be the negative of the Bayes optimal classification risk. Choose any $f_0 \in \mathcal{F}_0$. We may assume $\|f_0\|_u \leq 1$. By Lusin's theorem, there exists a continuous function $\widetilde{f}_0$ and a measurable set E with $P(E) < \frac{\epsilon}{2}$ such that on E^c, $\widetilde{f}_0 = f_0$ and $\left\|\widetilde{f}_0\right\|_u \leq \|f_0\|_u$. This guarantees that

$$\widetilde{\mathcal{F}}_0 := \{\widetilde{f} \in U(C([0,1]^d)) : \exists f \in \mathcal{F}_0 \text{ such that outside a set of measure less than } \frac{\epsilon}{2}, f = \widetilde{f}\}$$

is non-empty, and the classification risk associated with functions in this class differs from L^* by at most $\frac{\epsilon}{2}$. Fix any $\widetilde{f}_0 \in \widetilde{\mathcal{F}}_0$. Define the set $A := \{\theta \in \Theta : M^* - PM_\theta \geq \epsilon\}$. Because the mapping $R_\varrho : \Theta \to U(C([0,1]^d))$ is surjective (by for e.g., Theorem 3.1 of (Pinkus, 1999)), there exists $\theta_0 \in \Theta$ such that $R_\varrho(\theta_0) = \widetilde{f}_0$ and so $PM_{\theta_0} > PM_{\theta'}$ for all $\theta' \in A$. Then,

$$\limsup_{n\to\infty}\{\widehat{\theta}_n \in A\} \subseteq \left\{\limsup_{n\to\infty} \sup_{\theta\in A} P_n M_\theta \geq PM_{\theta_0}\right\}. \tag{14}$$

Note the $\limsup$ on the left-hand side of (14) is for a sequence of sets while the $\limsup$ on the right-hand side is for a sequence of real numbers. (14) follows from the fact that $\widehat{\theta}_n \in A$ infinitely often implies $\sup_{\theta\in A} P_n M_\theta \geq P_n M_{\widehat{\theta}_n} \geq P_n M_{\theta_0}$ infinitely often. But, $P_n M_{\theta_0} \to PM_{\theta_0}$ almost surely by the strong law of large numbers.

Before moving further, we show that the map $\theta \to PM_\theta$ is upper-semicontinuous. We use the following convention for the sign function, which is upper-semicontinuous:

$$sgn(x) = \begin{cases} -1, & \text{if } x < 0; \\ 1, & \text{if } x \geq 0. \end{cases}$$

The map defined by $t \mapsto -\mathbb{1}_{(\infty,0)}(t)$ is also upper-semicontinuous. Let $\mathcal{Z} := [0,1]^d \times \{0,1\}$. Then, the mapping defined by the following sequence of compositions is seen to be upper-semicontinuous:

$$\begin{aligned} &M : \mathcal{Z} \times \Theta \to \{-1, 0\}, \\ &(z, \theta) \mapsto (z, R_\varrho(\theta)) \mapsto (R_\rho(\theta)(x), y) \mapsto (sgn(R_\varrho(\theta)(x)), y) \\ &\mapsto sgn(R_\varrho(\theta)(x))(2y-1) \mapsto -\mathbb{1}_{(-\infty,0)}(sgn(R_\rho(\theta)(x))(2y-1)). \end{aligned}$$

The claimed upper-semicontinuity follows from the fact that the composition $f \circ g$ is upper-semicontinuous if either f is upper-semicontinuous and g is continuous or both f, g are upper-semicontinuous with f non-decreasing. In what follows, we will use the notation $M_\theta(z) := M(z, \theta)$.

Denote $\Theta_0 := \{\theta \in \Theta : P(M_\theta) = \sup_{\theta'\in\Theta} P(M_{\theta'})\}$. Here $P(M_\theta)$ denotes the integral of M_θ as a function of z when z is distributed according to P, and from the construction of M_θ, it follows that $P(M_\theta) = -P(sgn(R_\varrho(\theta)(X)) \neq 2Y - 1)$. Note this is the negative of the classification risk associated with the plug-in classifier based on $R_\varrho(\theta) + 1/2$. We also note this set is non-empty because Θ is compact and the map $\theta \to PM_\theta$ is upper-semicontinuous, essentially by Fatou's lemma.

Now returning to the proof, Denote by $M_U(z) := \sup_{\theta\in U} M_\theta(z)$ for any set $U \subseteq \Theta$. In our case, $M_U(\cdot)$ is also measurable because $R_\varrho(U)$ is contained in $U(C(\Omega))$, which is separable. For each $\theta \in A$, there exists some small enough open

neighborhood U^θ of θ, such that $PM_{U^\theta} < PM_{\theta_0}$ by upper-semicontinuity of the map $\theta \to PM_\theta$ (checked at the beginning of Section 2) and the definition of θ_0 and A. Consider the open cover of A by the open sets $\{U^\theta : \theta \in A\}$ with the aforementioned property. Since A is a compact subset of Θ, we have a finite subcover, which we denote by $\{U^{\theta_1}, \dots, U^{\theta_m}\}$ for some $\theta_1, \dots, \theta_m \in A$, $m \in \mathbb{N}$. With this construction,

$$\sup_{\theta \in A} P_n M_\theta \leq \max_{i:1 \leq i \leq m} P_n M_{U^{\theta_i}} \xrightarrow[a.s.]{n \to \infty} \max_{i:1 \leq i \leq m} PM_{U^{\theta_i}} < PM_{\theta_0}.$$

from which we conclude

$$P\left(\limsup_{n \to \infty} \sup_{\theta \in A} P_n M_\theta < PM_{\theta_0}\right) = 1 \tag{15}$$

Thus, the right-hand side of (14) has probability 0 because of (15), which implies $\widehat{\theta}_n \in A^c$ eventually with probability 1. Since ϵ was arbitrary, we conclude that

$$\lim_{n \to \infty} L(R_\varrho(\widehat{\theta}_n) + 1/2) \to L^* \text{ with probability } 1.$$

□

B.3. Proof of Theorem 3.6

Proof. In the definition of $\mathcal{NN}_n$, we may assume without loss of generality that all architectures have bounded widths, which ensures that $\mathcal{NN}_n$ may be viewed as a compact, completely metrizable space. A similar argument as in the proof of Theorem 2.2 shows that $\widehat{\theta}_n$ is well-defined as a measurable mapping from $\mathcal{Z}^n \to \mathcal{NN}_n$. Take the standard estimation and approximation error decomposition:

$$\underbrace{E[L(R_\varrho(\widehat{\theta}_n))] - \inf_{f \in R_\varrho(\mathcal{NN}_n)} E[L(f)]}_{①} + \underbrace{\inf_{f \in R_\varrho(\mathcal{NN}_n)} E[L(f)] - L^*}_{②}.$$

Because $\mathcal{F}$ is optimally representable by neural networks, we have

$$\sup_{f \in \mathcal{F}} \inf_{\Phi \in \mathcal{NN}_n} \|f - R_\varrho(\Phi)\|_{L^2([0,1]])} \leq C_d n^{-\frac{(2+\alpha)m^*}{2(1+\alpha)}}.$$

Comparison inequality (11) and Assumption 3.3 then implies that

$$\inf_{\Phi \in \mathcal{NN}_n} E[L(R_\varrho(\Phi))] - L^* \leq C_{\alpha,d} n^{-m^*}$$

where $C_{\alpha,d}$ only depends on α, d. This bounds ②. For ①, we directly appeal to a suitable modification of Theorem 5.8 of (Koltchinskii, 2011) using the fact that the VC-dimension of $R_\varrho(\mathcal{NN}_n)$ is bounded by $C_d n^{\frac{(2+\alpha)m^*}{2(1+\alpha)\gamma^*}} \log^p(n+1)$ where p is the degree of π ((Bartlett et al., 2019)). Then assumption (10) ensures that ① $\leq C_d n^{-m^*} \log^p(n+1)$ where p is the degree of π. Therefore, we conclude that the plug-in classification rule corresponding to $\{R_\varrho(\widehat{\theta}_n)\}_{n \in \mathbb{N}}$ achieves minimax optimal rate of convergence for $\mathcal{P}_\mathcal{F}$ up to a polylogarithmic factor. □