Minimax Optimal Deep Neural Network Classifiers Under Smooth Decision Boundary

Tianyang Hu   HUTIANYANG.UP@OUTLOOK.COM
The Chinese University of Hong Kong, Shenzhen, China

Ruiqi Liu   RUIQLIU@TTU.EDU
Texas Tech University, Lubbock, TX

Zuofeng Shang   ZSHANG@NJIT.EDU
New Jersey Institute of Technology, Newark, NJ

Guang Cheng   GUANGCHENG@UCLA.EDU
University of California, Los Angeles, Los Angeles, CA

Abstract

Deep learning has gained huge empirical successes in large-scale classification problems. In contrast, there is a lack of statistical understanding about deep learning methods, particularly in the minimax optimality perspective. For instance, in the classical smooth decision boundary setting, existing deep neural network (DNN) approaches are rate-suboptimal, and it remains elusive how to construct minimax optimal DNN classifiers. Moreover, it is interesting to explore whether DNN classifiers can circumvent the "curse of dimensionality" in handling high-dimensional data. The contributions of this paper are two-fold. First, based on a localized margin framework, we discover the source of suboptimality of existing DNN approaches. Motivated by this, we propose a new deep learning classifier using a divide-and-conquer technique: DNN classifiers are constructed on each local region and then aggregated to a global one. We further propose a localized version of the classical Tsybakov's noise condition, under which statistical optimality of our new classifier is established. Second, we show that DNN classifiers can adapt to low-dimensional data structures and circumvent the "curse of dimensionality" in the sense that the minimax rate only depends on the effective dimension, potentially much smaller than the actual data dimension. Numerical experiments are conducted on simulated data to corroborate our theoretical results.

Keywords: neural network classifier, smooth boundary, divide-and-conquer, statistical optimality, curse of dimensionality

1. Introduction

Deep learning has achieved many breakthroughs in modern classification tasks, especially for natural images (Deng et al., 2009; Krizhevsky et al., 2012; He et al., 2016; Nguyen et al., 2017). In contrast to the huge empirical success, statistical understanding of DNN classifiers is still lacking as to why neural networks perform better than traditional classification methods, particularly for high-dimensional structured data. Most theoretical works revolving around neural network classifiers focus on the generalization perspective, developing error bounds for trained classifiers on unseen data (Vapnik, 1999; Bousquet and Elisseeff,

2002; Zhang et al., 2016; Cao and Gu, 2019). However, derivation of generalization bounds mostly relies on the complexity of the DNN class, which is often independent of the data distribution. Deep learning is not better than traditional methods for every data set and the success of DNN classifiers should not only be contributed to the effectiveness of neural networks, but also task-specific properties such as data structures, noise distributions, etc. As a result, many existing generalization error bounds are not tight enough and may not fully capture the observed generalization performance (Dziugaite and Roy, 2017; Jiang et al., 2019). Without involving the concept of minimax optimality, such bounds may provide little guidance for practical applications.

There is a rich literature on minimax optimal estimation using DNNs (Schmidt-Hieber, 2020, 2019; Liu et al., 2022; Kohler et al., 2022; Liu et al., 2020), under which *task-specific* and *statistical optimal* results can be derived. By making specific assumptions on the data, the performance of an estimation method can be sharply characterized by the convergence rate of the estimation error (upper bound). In the mean time, a task-specific lower bound on the convergence rate can also be derived, independent of the estimation method. When the convergence rate of a certain method matches the lower bound, statistical optimality is achieved. Furthermore, for high-dimensional structured data, if the convergence rate is "dimension-free" in the sense that it doesn't depend on the original data dimension d, but on an effective dimension $d^* \ll d$, the "curse of dimensionality" can be alleviated. It is interesting to explore whether minimax optimal DNN approaches with dimension-free properties can be established in the classification setting. Existing literature on this front is limited, and most works either treat classification as regression by estimating the conditional class probability instead of the decision boundary (Kohler et al., 2020; Kohler and Langer, 2020; Bos and Schmidt-Hieber, 2021; Hu et al., 2021; Wang et al., 2022b,a; Wang and Shang, 2022) or settle for an upper bound on the misclassification risk (Kim et al., 2021; Steinwart et al., 2007; Hamm and Steinwart, 2020). Unlike regression problems where one intends to estimate the unknown regression functions, the goal of classification is to recover the unknown *decision boundary* separating different classes. Properties of the decision boundary are critical and the analysis is more demanding. In literature, the smooth decision boundary assumption and Tsybakov's low noise condition (Mammen and Tsybakov, 1999) are commonly adopted, whereas minimax optimal DNN classifiers under these assumptions are nonexistent, not even for simpler classification methods such as support vector machine (SVM) (Hamm and Steinwart, 2020). There is a clear gap between DNN classifier's empirical success and its suboptimality. In this work, we aim to address the issue of rate suboptimality in the smooth boundary setting and also provide new insights to how DNNs avoid the "curse of dimensionality". The main contributions of this paper are:

- By dissecting the 0-1 loss excess risk of DNN classifiers, we identify the potential origin of the suboptimality. Motivated by this, we propose a new deep learning classifier using a divide-and-conquer technique: DNN classifiers are constructed on each local region and then aggregated to a global one. We further propose a novel *localized* version of the classical Tsybakov's noise condition, which enables us to develop both lower and upper bounds on the convergence rate and establish statistical optimality for DNN classifiers with proper architectures, i.e., depth, width, sparsity, etc.

- In the proposed localized smooth decision boundary setting, we show that DNN classifiers can adapt to low-dimensional structures underlying the high-dimensional data and circumvent the "curse of dimensionality" in the sense that the optimal rate of convergence only depends on some effective dimension d^*, potentially much smaller than the data dimension.

This work is *the first* to establish the *minimax optimal* convergence rate for DNN classifiers under the classical smooth decision boundary setting, with a novel localized Tsybakov's noise condition. The proposed separation condition facilitates a finer-grained understanding of classification, which, DNN classifiers can take full advantage of by utilizing the divide-and-conquer strategy. The optimality proof relies on delicate constructions of DNNs where the representation power, structural flexibility, composition nature all play critical roles. The ability to adapt to low-dimensional data structures while achieving the optimal convergence rate showcases the power of deep learning and provides insights for understanding DNN's empirical success in classifying high-dimensional data.

2. Preliminary

Let $\mathbb{Q}$ be the Lebesgue measure on $\mathbb{R}^d$. For any function $f(\boldsymbol{x}) : \mathcal{X} \to \mathbb{R}$, denote $\|f\|_\infty = \sup_{\boldsymbol{x} \in \mathcal{X}} |f(\boldsymbol{x})|$ and $\|f\|_p = (\int_{\mathcal{X}} |f(\boldsymbol{x})|^p d\boldsymbol{x})^{1/p}$ for $p \in \mathbb{N}$. For a vector $\boldsymbol{x}$, $\|\boldsymbol{x}\|_p$ denotes its p-norm, for $1 \le p \le \infty$. L_p and l_p are used to distinguish function norms and vector norms. For two given sequences of real numbers $\{a_n\}_{n \in \mathbb{N}}$ and $\{b_n\}_{n \in \mathbb{N}}$, we write $a_n \lesssim b_n$ if there exists a constant $C > 0$ such that $a_n \le C b_n$ for all sufficiently large n. Let $\Omega(\cdot)$ be the counterpart of $O(\cdot)$ that $a_n = \Omega(b_n)$ means $a_n \gtrsim b_n$. In addition, we write $a_b \asymp b_n$ if $a_n \lesssim b_n$ and $a_n \gtrsim b_n$. For $a, b \in \mathbb{R}$, denote $a \vee b = \max\{a, b\}$ and $a \wedge b = \min\{a, b\}$. We use $\mathbb{I}$ to denote the indicator function. For a set $G \subset \mathbb{R}^d$, denote ∂G as its boundary and G° as its interior. For two sets $G_1, G_2 \subset \mathbb{R}^d$, let their symmetric difference be $d_\triangle(G_1, G_2) = \mathbb{Q}(G_1 \triangle G_2) = \mathbb{Q}((G_1 \backslash G_2) \cup (G_2 \backslash G_1))$. We shall also use differential operator $\partial^{\boldsymbol{\alpha}} = \partial^{\alpha_1} \dots \partial^{\alpha_r}$ with multi-index $\boldsymbol{\alpha} = (\alpha_1, \dots, \alpha_r) \in \mathbb{N}^r$ and $|\boldsymbol{\alpha}| := |\boldsymbol{\alpha}|_1$. For $\beta > 0$, let $\lfloor \beta \rfloor$ denote the largest integer strictly smaller than β.

2.1 Binary classification

Denote $\boldsymbol{x} \in \mathcal{X} \subset \mathbb{R}^d$ as the feature vector and $y \in \{-1, 1\}$ as the label. Assume the classes are balanced, i.e., $\mathbb{P}(y = 1) = 1/2$, and

$$\boldsymbol{x} \mid y = 1 \sim p(\boldsymbol{x}), \quad \boldsymbol{x} \mid y = -1 \sim q(\boldsymbol{x}),$$

where p, q are two bounded densities on $\mathcal{X}$. The goal of classification is to find the optimal classifier C^* with the lowest expected *0-1 loss*

$$R(C) = \mathbb{E}_{(\boldsymbol{x}, y)}[\mathbb{I}\{C(\boldsymbol{x}) \ne y\}].$$

C^* corresponds to the optimal decision region $G^* := \{\boldsymbol{x} \in \mathcal{X}, p(\boldsymbol{x}) - q(\boldsymbol{x}) \ge 0\}$ by assigning label 1 if $\boldsymbol{x} \in G^*$ and label -1 otherwise. Equivalently, we can consider real-valued functions f and classification can be done by $C(\boldsymbol{x}) = \text{sign}(f(\boldsymbol{x}))$. With a slight abuse of notation, we use $R(C), R(G)$ and $R(f)$ interchangeably.

Without knowing the data distribution, C^* is not directly calculable. Instead, we observe data set $\mathcal{D} = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$ and estimate C^* by the empirical risk minimizer, i.e.,

$$\widehat{C} = \operatorname{argmin} R_n(C), \text{ where } R_n(C) = \sum_{i=1}^n \mathbb{I}\{C(\boldsymbol{x}_i) \neq y_i\}.$$

In practice, the empirical 0-1 loss minimizer is not computationally feasible because minimizing such loss over $\mathcal{C}_n$ is NP hard (Bartlett et al., 2006). An alternative approach is to replace the 0-1 loss with other computationally friendly surrogate losses $\phi : \mathbb{R} \to \mathbb{R}_+$, e.g. hinge loss $\phi(z) = (1 - z)_+ = \max\{1 - z, 0\}$, logistic loss $\phi(z) = \log(1 + \exp(-z))$, etc. Even though surrogate losses are used in practice, 0-1 loss gives the most fundamental characterization of the classification problem and is of great theoretical interest. This work focuses on 0-1 loss and surrogate losses such as hinge loss, cross-entropy are reserved for future work.

The performance of an estimated classifier $\widehat{C}$ can be evaluated by its *excess risk* defined as

$$\mathcal{E}(\widehat{C}, C^*) = R(\widehat{C}) - R(C^*),$$

which measures the closeness between $\widehat{C}$ and C^* in terms of the expected 0-1 loss. Convergence rate refers to how fast $\mathcal{E}(\widehat{C}, C^*)$ converges to zero with respect to n and it is the focus on this paper. Knowing the relationship between C^* and G^*, classification can also be seen as nonparametric estimation of sets, and the 0-1 excess risk of an estimator $\widehat{G}$ can be written as

$$\mathcal{E}(\widehat{G}, G^*) = d_{p,q}(\widehat{G}, G^*) = \int_{\widehat{G} \triangle G^*} |p(\boldsymbol{x}) - q(\boldsymbol{x})| \mathbb{Q}(d\boldsymbol{x}).$$

There are two key factors governing the rate of convergence in classification: the complexity of the decision boundary and how concentrated the data are near the decision boundary, which we will introduce in Section 2.2 and 2.3 respectively.

2.2 Smooth boundary assumption

In statistics literature, one of the most common boundary assumptions is called the smooth boundary fragments (Mammen and Tsybakov, 1999; Tsybakov, 2004; Korostelev and Tsybakov, 2012). For $d \geq 2$ and $\boldsymbol{x} = (x_1, \ldots, x_d) \in \mathcal{X}$, let $\boldsymbol{x}_{-d} = (x_1, \cdots, x_{d-1})$ be the vector of the first $d - 1$ coordinates. The smooth boundary fragment setting assumes that the optimal set G^* can be expressed as

$$G_{f^*} := \{\boldsymbol{x} \in \mathcal{X} : f^*(\boldsymbol{x}_{-d}) \geq x_d\}, \tag{1}$$

where f^* is an unknown $(d - 1)$-variate smooth function. Throughout this article, assume that f^* belongs to the following unit ball of β-Hölder space:

$$\mathcal{H}_\beta = \left\{ f : \sum_{\boldsymbol{\alpha}:|\boldsymbol{\alpha}|<\beta} \|\partial^{\boldsymbol{\alpha}} f\|_\infty + \sum_{\boldsymbol{\alpha}:|\boldsymbol{\alpha}|=\lfloor\beta\rfloor} \sup_{\substack{\boldsymbol{x},\boldsymbol{x}'\in\mathcal{X} \\ \boldsymbol{x}_{-d}\neq\boldsymbol{x}'_{-d}}} \frac{|\partial^{\boldsymbol{\alpha}} f(\boldsymbol{x}_{-d}) - \partial^{\boldsymbol{\alpha}} f(\boldsymbol{x}'_{-d})|}{|\boldsymbol{x}_{-d} - \boldsymbol{x}'_{-d}|^{\beta-\lfloor\beta\rfloor}} \leq 1 \right\}. \tag{2}$$

Clearly, $f^* \in \mathcal{H}_\beta$ if all partial derivatives up to order $\lfloor \beta \rfloor$ exist and are bounded, and the partial derivatives of order $\lfloor \beta \rfloor$ are $\beta - \lfloor \beta \rfloor$ Hölder continuous. Let $\mathcal{G}_\beta^* = \{G_{f^*} : f^* \in \mathcal{H}_\beta\}$ be the collection of all sets in (1), namely, the hypothesis class.

Due to the smoothness of f^*, the complexity of the hypothesis class $\mathcal{G}_\beta^*$, which is usually measured by *bracketing entropy*, can be controlled. To be more specific, for any $\delta > 0$, the bracketing number $\mathcal{N}_B(\delta, \mathcal{G}, d_\triangle)$ is the minimal number of set pairs (U_j, V_j) such that

(a) For each j, $U_j \subset V_j$ and $d_\triangle(U_j, V_j) \leq \delta$;

(b) For any $G \in \mathcal{G}$, there exists a pair (U_j, V_j) such that $U_j \subset G \subset V_j$.

Simply denote $\mathcal{N}_B(\delta) = \mathcal{N}_B(\delta, \mathcal{G}, d_\triangle)$ if no confusion arises. The bracketing entropy is defined as $H_B(\delta) = \log \mathcal{N}_B(\delta, \mathcal{G}, d_\triangle)$. In the smooth boundary fragment case, Mammen and Tsybakov (1999) showed that

$$H_B(\delta, \mathcal{G}_\beta^*, d_\triangle) \lesssim \delta^{-\frac{d-1}{\beta}}.$$

For more flexibility, the boundary fragments assumption can be extended to be more general that G^* consists of unions and intersections of smooth hyper-surfaces defined as in (1) (Kim et al., 2021; Tsybakov, 2004). In this work, we focus on the basic case (1) for simplicity.

2.3 Separation assumption

The following Tsybakov's noise condition (Mammen and Tsybakov, 1999) quantifies how separated p and q are near the decision boundary:

(N) There exists constants $c, T > 0$ and $\kappa \in [0, \infty]$ such that for any $0 \leq t \leq T$,

$$\mathbb{Q}\left(\{\boldsymbol{x} : |p(\boldsymbol{x}) - q(\boldsymbol{x})| \leq t\}\right) \leq ct^\kappa.$$

(N) is a population-level assumption and κ is referred to as the noise exponent. (N) trivially holds for $\kappa = 0$. The bigger the κ, the more separated the data around the decision boundary, and hence, the easier the classification. In the extreme case where p, q have different supports, κ can be arbitrarily large (∞) and the boundary recovery is the easiest. To another extreme where $\mathbb{Q}(\{\boldsymbol{x} \in \mathcal{X} : p(\boldsymbol{x}) = q(\boldsymbol{x})\}) > 0$, there exists a region where different classes are indistinguishable. In this case, $\kappa = 0$ and the optimal decision boundary is hard to estimate in that region.

Under the smooth boundary fragment assumption (1) with smoothness β and the Tsybakov's noise condition (N) with noise exponent κ, Mammen and Tsybakov (1999) showed that the optimal rate of convergence for the 0-1 loss excess risk is

$$\inf_{C_n \in \mathcal{C}} \sup_{G^* \in \mathcal{G}_\beta^*} \mathcal{E}(C_n, G^*) = \Omega\left(n^{-\frac{\beta(\kappa+1)}{\beta(\kappa+2)+(d-1)\kappa}}\right), \tag{3}$$

where $\mathcal{C}$ can be any classifier family. Note that the "curse of dimensionality" does occur in this bound. As d gets larger, the rate becomes extremely slow.

5

2.4 DNN in classification

We consider DNN with rectified linear unit (ReLU) activation that $\sigma(z) = \max\{z, 0\}$. For a ReLU neural network f_n indexed by the sample size, let L_n be the number of layers, N_n be the maximum width of all the layers, S_n be the total number of non-zero weights, B_n be maximum absolute values of all the weights. Denote $\mathcal{F}_n = \mathcal{F}(L_n, N_n, S_n, B_n)$ as a DNN family with structural constraints specified by L_n, N_n, S_n, B_n. In this work, depth L_n and width N_n are the primary focuses.

Convergence rate of DNN classifiers has been investigated in literature. Kim et al. (2021) derived fast convergence rates of ReLU DNN classifiers learned using the hinge loss ($\phi(z) = \max(0, 1 - z)$). Under the smooth boundary fragment assumption (1) and Tsybakov's noise condition (N), the empirical hinge loss minimizer

$$\widehat{f}_{\phi,n} = \underset{f \in \mathcal{F}_n}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^{n} \phi(y_i f(\boldsymbol{x}_i)),$$

within some DNN family with carefully selected L_n, N_n, S_n, B_n satisfies

$$\sup_{G^* \in \mathcal{G}_\beta^*} \mathbb{E}\left[\mathcal{E}(\widehat{f}_{\phi,n}, G^*)\right] \lesssim \left(\frac{\log^3 n}{n}\right)^{\frac{\beta(\kappa+1)}{\beta(\kappa+2)+(d-1)\kappa+(d-1)}}. \tag{4}$$

Their result is rate *suboptimal* comparing to the minimax lower bound (3). To the best of the authors' knowledge, this is the only existing convergence rate result for DNN classifiers under the smooth boundary setting. The same suboptimal rate is referred to as "best known rates" for SVM (Hamm and Steinwart, 2020). Other theoretical works on convergence rate of DNN classifiers are carried out under other assumptions, e.g., separable (Zhang, 2000; Hu et al., 2021), teacher student setting (Hu et al., 2020), smooth conditional probability (Audibert et al., 2007; Steinwart et al., 2007; Kohler et al., 2020), etc. Classification by estimating the conditional probability is usually referred to as "plug-in" classifiers and it's worth noting that it essentially reduces classification to regression. In comparison, estimating the decision boundary is a more fundamental setting for classification (Hastie et al., 2009) and it is the focus of this work.

3. Main results – optimal rate of convergence

Under the classical smooth boundary fragment setting, DNN classifiers have not been proven optimal in recovering the decision boundary, which casts doubt on the empirical success of DNNs in classification. In this section, we first investigate the potential reason behind the suboptimality.

The noise exponent κ in the Tsybakov's noise condition (N) plays a critical role in determining the convergence rate. In general, bigger κ implies better separation, and hence easier boundary recovery. However, if we take a closer look at the excess risk by decomposing it into the stochastic (empirical) error and the approximation error (using DNN to approximate smooth functions), bigger κ benefits the former but is harmful to the latter (the approximation error $d_\triangle$ is amplified by a power of $1 + 1/\kappa$ as in Lemma 13. As $d_\triangle$ goes to zero, a larger κ will result in a smaller $1 + 1/\kappa$ so $d_\triangle^{1+1/\kappa}$ will shrink to zero slower.).

In other words, the bottleneck for the stochastic error is the region with the smallest separation while that for the approximation error is the region with the largest separation. If the separation along the decision boundary is inconsistent with sharp changes, the overall estimation performance of DNN classifiers could be harmed. The optimal rate of convergence can be achieved under extra separation consistency conditions. Complementary to the classical Tsybakov's noise condition (N), which indicates that the separation is not too small, the following condition ensures that the separation cannot be too large as well.

(N$^+$) There exist constants $c_1, T > 0$ and $\bar{\kappa} \in [0, \infty]$ such that for any $0 \leq t \leq T$,

$$\mathbb{Q}\left(\{\boldsymbol{x} \in G : |p(\boldsymbol{x}) - q(\boldsymbol{x})| \leq t\}\right) \geq c_1 t^{\bar{\kappa}}$$

holds for any positive-measure set $G \subset \mathcal{X}$ containing the decision boundary, i.e., $\partial G^* \cap G^\circ$ is not empty.

If (N) and (N$^+$) both hold with the same exponent κ, the separation is guaranteed to be consistent along the decision boundary. The following theorem utilizes this extra condition and shows that optimal rate of convergence can be achieved for DNN classifiers.

Lemma 1 *(Informal) Under assumptions (N) and the smooth boundary fragment assumption (1), if we further assume (N$^+$) with $\bar{\kappa} = \kappa$, then the empirical 0-1 loss minimizer within a ReLU DNN family with proper size achieves the optimal 0-1 loss excess risk convergence rate of $n^{-\frac{\beta(\kappa+1)}{\beta(\kappa+2)+(d-1)\kappa}}$.*

Lemma 1 is a special case of Theorem 5 and the proof can be found in the Section 7.3. The improvement to optimality stems from a better approximation error bound, which follows from the separation consistency guaranteed by condition (N$^+$). However, assuming identical separation over the whole support, i.e., $\kappa = \bar{\kappa}$ for (N) and (N$^+$), is too strong. It becomes much more realistic if we put this assumption on a small neighborhood, or a small segment of the decision boundary. If the densities p, q are not too irregular, the separation is expected to be locally consistent.

To this end, we conduct convergence analysis in a local region of the decision boundary and consider a localized version of the classical Tsybakov's noise condition. For simplicity, $\mathcal{X}$ is assumed to be $[0,1]^d$ in the remaining part of the paper. Extension to compact $\mathcal{X}$ is straightforward.

3.1 Localized separation condition

Recall the boundary assumption that $G^* = \{\boldsymbol{x} \in [0,1]^d : f^*(\boldsymbol{x}_{-d}) - x_d \geq 0\}$ where f^* is some smooth function from $\mathbb{R}^{d-1}$ to $\mathbb{R}$. Denote the decision boundary to be $\partial G^* := \{\boldsymbol{x} \in [0,1]^d : f^*(\boldsymbol{x}_{-d}) = x_d\}$. Every point $\boldsymbol{x} \in \partial G^*$ can be written as $\boldsymbol{x} = (\boldsymbol{x}_{-d}, f^*(\boldsymbol{x}_{-d}))$. Without loss of generality, assume $\mathbb{Q}(\partial G^*) = 0$. Notice that even though f^* defines the decision boundary, it has nothing to do with the separation. The separation condition, e.g., (N), depends on properties of the data densities p, q around the decision boundary. To quantify the separation, define for each $\boldsymbol{x}_{-d} \in [0,1]^{d-1}$,

$$m_{\boldsymbol{x}_{-d}}(t) := |p((\boldsymbol{x}_{-d}, f^*(\boldsymbol{x}_{-d}) + t)) - q((\boldsymbol{x}_{-d}, f^*(\boldsymbol{x}_{-d}) + t))|,$$
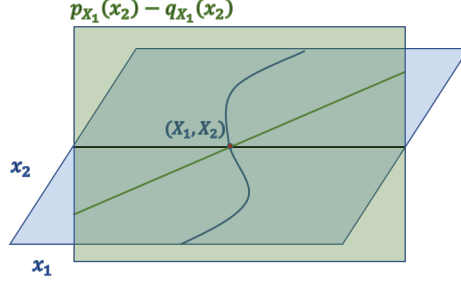
Figure 1: Illustration of the localized separation condition (M1) when $d = 2$. The blue plane represents $\mathcal{X}$ and the curved blue line is the optimal decision boundary. Fix some (x_1, x_2) in the decision boundary. Along the x_1 direction (black line), the density difference $p - q$ is plotted in the green plane as the green line. (M1) defines the noise exponent $K(x_1)$ locally at x_1. If $p - q$ is linear as shown, $K(x_1) = 1$.

which captures the how $|p(\boldsymbol{x}) - q(\boldsymbol{x})|$ changes along the direction of x_d on each point of the decision boundary. For ease of notation, we write $m_{\boldsymbol{x}_{-d}}(t)$ and $m_{\boldsymbol{x}}(t)$ when no confusion arises. Notice that $m_{\boldsymbol{x}}(0) = 0$ by definition and we want to characterize how $m_{\boldsymbol{x}}(t)$ behaves when t is close to 0, i.e., close to the decision boundary. Moreover, for every $\boldsymbol{x} \in \partial G^*$, define the localized separation exponent

$$K(\boldsymbol{x}) = \sup\{k \geq 0 : \lim_{t \to 0} \frac{m_{\boldsymbol{x}}(t)}{|t|^{1/k}} > 0\}.$$

$K(\boldsymbol{x})$ characterizes the separation condition locally at $\boldsymbol{x}_{-d}$, i.e. how separated are p and q on each point of the decision boundary, along the direction of x_d. For some $\boldsymbol{x} \in \partial G^*$, if the directional derivative of $p - q$ tangent to the decision boundary is finite and non-zero, then $K(\boldsymbol{x}) = 1$. Note that if $K(\boldsymbol{x})$ is regular enough, e.g., continuous, Tsybakov's noise condition (N) with exponent κ implies that $\kappa \leq \inf_{\boldsymbol{x} \in \partial G^*} K(\boldsymbol{x})$ and (N$^+$) implies $\bar{\kappa} \geq \sup_{\boldsymbol{x} \in \partial G^*} K(\boldsymbol{x})$. Consider the following separation conditions which form a localized version of the Tsybakov's noise condition.

(M1) There exists $\epsilon_0 > 0$ small enough and a constant $0 < C_{\epsilon_0} < \infty$ such that for all $\boldsymbol{x} \in \partial G^*$ and any $0 < t < \epsilon_0$,

$$\frac{1}{C_{\epsilon_0}} \leq \frac{m_{\boldsymbol{x}}(t)}{|t|^{1/K(\boldsymbol{x})}} \leq C_{\epsilon_0}.$$

(M2) $K(\boldsymbol{x})$ is α-Hölder continuous for some $0 < \alpha \leq 1$, i.e. there exists constant C_K such that for any $\boldsymbol{x}_1, \boldsymbol{x}_2 \in \partial G^*$,

$$|K(\boldsymbol{x}_1) - K(\boldsymbol{x}_2)| \leq C_K \|\boldsymbol{x}_1 - \boldsymbol{x}_2\|_2^\alpha.$$

Remark 2 (Justification of (M1,M2)) *The proposed separation condition is a localized version of the classical Tsybakov's noise condition (N) where (M1) corresponds to (N) in*

*a local region and (M2) ensures consistency among nearby regions. (M1,M2) is stronger than (N), only because it's a finer characterization of (N) and we expect it to hold if the data distribution is not too irregular. **Many existing settings are special examples of (M1,M2)**, e.g., in the separable case, $K(\boldsymbol{x}) \equiv \infty$; if p, q are piecewise linear, then (M) holds with $K(\boldsymbol{x}) \equiv 1$; if p, q are Hölder smooth with smoothness more than 2, e.g., Gaussian mixtures, then (M1,M2) holds and $K(\boldsymbol{x}) \leq 1$. Like (N), our condition is more of theoretical interest and hard to verify for real data. Still, the developed results can contribute to the understanding of DNN's empirical success in classification, e.g., how they can take advantage of better class separation and low-dimensional data structures.*

(M1,M2) is a slightly stronger characterization of the boundary separation. The following lemma shows that (M1) implies (N) and (N$^+$). The proof can be found in Section 7.

Lemma 3 *Denote $\kappa^- = \inf_{\boldsymbol{x} \in \partial G^*} K(\boldsymbol{x})$ and $\kappa^+ = \sup_{\boldsymbol{x} \in \partial G^*} K(\boldsymbol{x})$. Then condition (M1) implies that (N) holds with $\kappa = \kappa^-$ and (N$^+$) holds with $\bar{\kappa} = \kappa^+$.*

Although (M1,M2) and (N) are closely related, the fundamentals of the classification problem, measured by the convergence lower bound, might be different. Below we investigate the 0-1 loss excess risk lower bound under our proposed separation condition.

Theorem 4 *Under the smooth boundary fragments setting (1) with smoothness β. Assume condition (M1) and let $\kappa^- = \inf_{\boldsymbol{x} \in \partial G^*} K(\boldsymbol{x})$. For any function space $\mathcal{F}$, the 0-1 loss excess risk has the following lower bound,*

$$\inf_{\widehat{f} \in \mathcal{F}} \sup_{G^* \in \mathcal{G}_\beta^*} \mathbb{E}[\mathcal{E}(\widehat{f}, G^*)] \gtrsim \left(\frac{1}{n}\right)^{\frac{\beta(\kappa^- + 1)}{\beta(\kappa^- + 2) + (d-1)\kappa^-}}.$$

The convergence lower bound under (M1) only depends on κ^- and has the same form compared to (3) under (N). This is expected since (M1,M2) is a localized version of (N).

3.2 Localized convergence analysis

Defining function $K(\boldsymbol{x})$ enables us to consider local convergence behaviours. As can be seen from Lemma 1 and Lemma 3, if the separation described by $K(\boldsymbol{x})$ is consistent, i.e., $\kappa^+ \approx \kappa^-$, the convergence rate can be improved to be optimal. Such consistency is expected within any small-enough local region due to condition (M2). To further investigate the interplay between κ^+ and κ^-, we conduct localized convergence analysis in this section.

There are many ways to construct local regions. Since we are considering the smooth boundary fragment setting (1), where x_d is a special dimension, a natural choice for the local regions is equal-sized grids in the $\boldsymbol{x}_{-d}$ space. Choose integer $M > 0$ and divide $[0,1]^d$ into M^{d-1} regions

$$[0,1]^d = \bigcup_{j_1,\ldots,j_{d-1}=1}^{M} D_{(j_1,\ldots,j_{d-1})},$$

where $D_{(j_1,\ldots,j_{d-1})} := \{\boldsymbol{x} \in [0,1]^d : x_1 \in [\frac{j_1-1}{M}, \frac{j_1}{M}), \cdots, x_{d-1} \in [\frac{j_{d-1}-1}{M}, \frac{j_{d-1}}{M}]\}$. Denote the grid points to be $\bar{\boldsymbol{x}}_{j_1,\ldots,j_{d-1}}$. For ease of notation, let $\boldsymbol{j}_{-d} = (j_1, \cdots, j_{d-1})$ and denote J_M

as all M^{d-1} combinations of $\boldsymbol{j}_{-d}$'s described above. Correspondingly, divide the data set as $\mathcal{D} = \cup_{\boldsymbol{j}_{-d} \in J_M} \mathcal{D}_{\boldsymbol{j}_{-d}}$ where $\mathcal{D}_{\boldsymbol{j}_{-d}} = \{(\boldsymbol{x}, y) \in \mathcal{D} : \boldsymbol{x} \in D_{\boldsymbol{j}_{-d}}\}$. Similarly, the 0-1 loss can be decomposed into

$$
\begin{aligned}
d_{p,q}(\widehat{G}_n, G^*) &= \int_{\widehat{G}_n \triangle G^*} |p(\boldsymbol{x}) - q(\boldsymbol{x})| d\boldsymbol{x} = \sum_{\boldsymbol{j}_{-d} \in J_M} \int_{(\widehat{G}_n \triangle G^*) \cap D_{\boldsymbol{j}_{-d}}} |p(\boldsymbol{x}) - q(\boldsymbol{x})| d\boldsymbol{x} \\
&:= \sum_{\boldsymbol{j}_{-d} \in J_M} d_{\boldsymbol{j}_{-d}}(\widehat{G}_n, G^*).
\end{aligned}
$$

The empirical 0-1 loss, $R_n(f) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}\{f(\boldsymbol{x}_i)y_i < 0\}$, can also be decomposed into M^{d-1} parts, i.e., $R_n(G) = \sum_{\boldsymbol{j}_{-d} \in J_M} R_{n,\boldsymbol{j}_{-d}}$ where

$$
R_{n,\boldsymbol{j}_{-d}} = \frac{1}{|\mathcal{D}_{\boldsymbol{j}_{-d}}|} \sum_{i=1}^{n} \mathbb{I}\{\boldsymbol{x}_i \in D_{\boldsymbol{j}_{-d}} : f(\boldsymbol{x}_i)y_i < 0\}.
$$

Now we focus on each local region $D_{\boldsymbol{j}_{-d}}$. Similar to that in the whole region $[0,1]^d$, we have the following theorem on the local 0-1 loss excess risk convergence rate.

Theorem 5 *Under assumption (M1), further assume that for some $\boldsymbol{j}_{-d} \in J_M$, $\kappa^- \leq K(\boldsymbol{x}) \leq \kappa^+$ for all $\boldsymbol{x} \in D_{\boldsymbol{j}_{-d}}$. Let $\widetilde{\mathcal{F}}_n$ be a ReLU DNN family with size in the order of*

$$
\widetilde{N}_n \widetilde{L}_n \asymp n^{\frac{\kappa^+(\kappa^-+1)(d-1)/2}{(\kappa^-+2)(\kappa^++1)\beta + (d-1)\kappa^+(\kappa^-+1)}} \cdot \log^2(n).^{[1]}
$$

Let the empirical 0-1 loss minimizer be

$$
\widehat{f}_{n,\boldsymbol{j}_{-d}} := \operatorname*{argmin}_{f \in \widetilde{\mathcal{F}}_n} R_{n,\boldsymbol{j}_{-d}}(f). \tag{5}
$$

Then the 0-1 loss excess risk satisfies

$$
\sup_{G^* \in \mathcal{G}_\beta^*} \mathbb{E}(R_{\boldsymbol{j}_{-d}}(\widehat{f}_{n,\boldsymbol{j}_{-d}}) - R_{\boldsymbol{j}_{-d}}(G^*)) = \widetilde{O}\left(n^{-\frac{(\kappa^-+1)\beta}{(\kappa^-+2)\beta + \left(\frac{\kappa^-+1}{\kappa^++1}\right)(d-1)\kappa^+}}\right),
$$

*where $\widetilde{O}(\cdot)$ hides the $\log(n)$ terms.*

Note that (5) with $M = 1$ corresponds to regular DNN classifiers considered by Kim et al. (2021) and Hamm and Steinwart (2020). Our result is sharper in the sense that κ^+ appears in both the convergence rate and the classifier size. The requirement for the network size comes from the DNN approximation literature can can be quite flexible. In Theorem 5, no constraint is put on B_n and S_n since we are using the results from Lu et al. (2020).

Remark 6 (Origin of sub-optimality) *Our local convergence rate in Theorem 5 closely resembles those established under the original Tsybakov's noise condition (N). On one hand,*

---

1. $\widetilde{\mathcal{F}}_n$ is used to denote DNN family for local estimation. $\mathcal{F}_n$ is reserved for the global estimator DNN family.
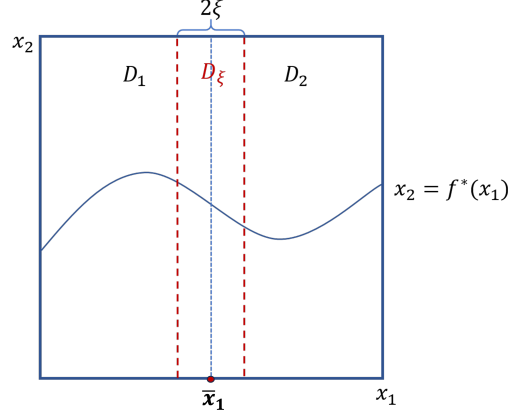
Figure 2: Illustration of region D_ξ in $d = 2, M = 2$ case.

the convergence rate bottleneck is indeed the minimum value of $K(\boldsymbol{x})$ in that region and κ^- plays the same role as κ in (N). On the other hand, the extra term in the denominator $(\kappa^- + 1)/(\kappa^+ + 1)$ reveals the potential source of the suboptimality of existing results. If no assumption is made on κ^+, one allows $\kappa^+ \to \infty$ so that $\kappa^+/(\kappa^+ + 1) \to 1$, the convergence rate reduces to the suboptimal one (4) obtained in Kim et al. (2021) and Hamm and Steinwart (2020). However, if $\kappa^+ = \kappa^-$, i.e., $K(\boldsymbol{x})$ is constant, optimal rate is attainable. This is consistent with our numerical findings in which the regular DNN classifier performs well when $K(\boldsymbol{x})$ is constant; see Figure 5(a). Condition (M2) says that $K(\boldsymbol{x})$ is locally smooth, i.e., $\kappa^+ \approx \kappa^-$ as long as the local region is small.

In order to achieve the fastest convergence rate, Theorem 5 requires $\widetilde{\mathcal{F}}_n$ to have proper size. If $\kappa^- = \kappa^+ = \kappa$, the size constraint is

$$\widetilde{N}_n \widetilde{L}_n = \widetilde{O}\left(n^{\frac{\kappa(d-1)/2}{(\kappa+2)\beta+(d-1)\kappa}}\right).$$

The bigger the κ, the larger the required size, and the faster the convergence rate. In case the size constraint is not met, the the following corollary gives the corresponding convergence results.

Corollary 7 *Under the same setting as Theorem 5, denote $r_0 = \frac{\kappa^+(\kappa^-+1)(d-1)}{(\kappa^-+2)(\kappa^++1)\beta+(d-1)\kappa^+(\kappa^-+1)}$ and let $\widetilde{\mathcal{F}}_n$'s size satisfy $\widetilde{N}_n \widetilde{L}_n = \Omega(n^{r/2})$ for some $0 < r < 1$. If $r < r_0$, the approximation error dominates and the 0-1 loss excess risk convergence rate becomes $\widetilde{O}\left(n^{-\frac{r\beta(\kappa^++1)}{(d-1)\kappa^+}}\right)$. If $r > r_0$, the stochastic error dominates and the convergence rate is $\widetilde{O}\left(n^{-\frac{(1-r)(\kappa^-+1)}{\kappa^-+2}}\right)$.*

3.3 Construction of the global estimator

In this section, we proceed from a localized analysis to the global one and evaluate the overall excess risk convergence rate. The goal is to construct a global classifier that takes advantage

of the developed results for each region $D_{\boldsymbol{j}_{-d}}$, such that Theorem 5 can be applied locally. To be more specific, denote the properly sized network family (according to Theorem 5) for each region as $\widetilde{\mathcal{F}}_{n,\boldsymbol{j}_{-d}}$, with width $\widetilde{N}_{n,\boldsymbol{j}_{-d}}$ and depth $\widetilde{L}_{n,\boldsymbol{j}_{-d}}$. Let $\mathcal{F}_n$ be the DNN family for global estimation, with depth L_n and width N_n. Denote the global empirical minimizer within $\mathcal{F}_n$ to be

$$\widehat{f}_n := \underset{f \in \mathcal{F}_n}{\operatorname{argmin}} R_n(f). \tag{6}$$

Then, it's ideal if for any $\boldsymbol{j}_{-d} \in J_M$, $\widehat{f}_n$ satisfies

$$R_{n,\boldsymbol{j}_{-d}}(\widehat{f}_n) = \min_{f \in \widetilde{\mathcal{F}}_{n,\boldsymbol{j}_{-d}}} R_{n,\boldsymbol{j}_{-d}}(f). \tag{7}$$

Note that $\widehat{f}_n$ degenerates to the regular DNN classifier when $M = 1$. For general $M > 1$, we aim to construct $\mathcal{F}_n$ in a divide-and-conquer fashion such that (7) is satisfied with high probability. Below we present the construction of $\mathcal{F}_n$ and the probabilistic argument.

Recall the definition of $D_{\boldsymbol{j}_{-d}}$, where we divide $[0,1]^d$ along $\boldsymbol{x}_{-d}$ into M^{d-1} equally sized regions. For some $0 < \xi \ll 1/M$, define a small region around all grid points $\bar{\boldsymbol{x}}_{\boldsymbol{j}_{-d}}$'s as

$$D_\xi = \{\boldsymbol{x} \in [0,1]^d : \min_{i=1,\ldots,d-1} |x_i - \bar{\boldsymbol{x}}_{\boldsymbol{j}_{-d},i}| \leq \xi, \ \boldsymbol{j}_{-d} \in J_M\},$$

where $\bar{\boldsymbol{x}}_{\boldsymbol{j}_{-d},i}$ denotes the i-th element of $\bar{\boldsymbol{x}}_{\boldsymbol{j}_{-d}}$. D_ξ with $d = 2, M = 2$ is illustrated in Figure 2. We aim to show that (7) holds outside D_ξ for some carefully chosen ξ and $\mathcal{F}_n$. To this end, define event

$$E_\xi := \{\boldsymbol{x}_i \notin D_\xi : \forall i = 1, 2, \ldots, n\}.$$

Since $p(\boldsymbol{x})$ and $q(\boldsymbol{x})$ are both bounded densities (by c_0), we have

$$\mathbb{P}(x \in D_\xi) \leq c_0 \mathbb{Q}(D_\xi) \leq 2c_0 M \xi(d-1),$$

The last inequality is due to the fact that D_ξ is covered by the union of $M(d-1)$ rectangles in $\mathbb{R}^d$, each with volume 2ξ. Therefore, if we choose M such that $nM\xi(d-1) \to 0$ as $n \to \infty$, then

$$\mathbb{P}(E_\xi) \geq (1 - 2c_0 M \xi(d-1))^n \to 1.$$

In the remaining of the analysis, we assume that E_ξ happens. For any $f_{n,\boldsymbol{j}_{-d}} \in \widetilde{\mathcal{F}}_{n,\boldsymbol{j}_{-d}}$, we make modifications and further construct $f^+_{n,\boldsymbol{j}_{-d}}$ that satisfies the following properties:

(P1) On $D_{\boldsymbol{j}_{-d}} \backslash D_\xi$, $f^+_{n,\boldsymbol{j}_{-d}} = f_{n,\boldsymbol{j}_{-d}}$;

(P2) Outside $D_{\boldsymbol{j}_{-d}}$, $f^+_{n,\boldsymbol{j}_{-d}} = 0$;

(P3) $f^+_{n,\boldsymbol{j}_{-d}} \in \widetilde{\mathcal{F}}^+_{n,\boldsymbol{j}_{-d}}$ where $\widetilde{\mathcal{F}}^+_{n,\boldsymbol{j}_{-d}}$ is slightly larger than $\widetilde{\mathcal{F}}_{n,\boldsymbol{j}_{-d}}$ with depth $\widetilde{L}^+_{n,\boldsymbol{j}_{-d}} = \widetilde{L}_{n,\boldsymbol{j}_{-d}} + O(1)$ and width $\widetilde{N}^+_{n,\boldsymbol{j}_{-d}} = 2\widetilde{N}_{n,\boldsymbol{j}_{-d}}$.
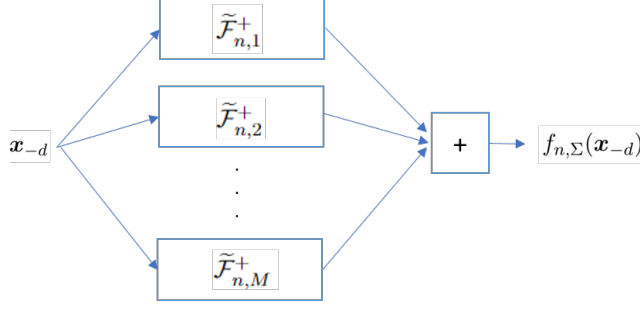
Figure 3: Illustration of the estimator DNN family $\mathcal{F}_n$ with $d = 2$.

The construction details and verification of (P1) to (P3) are deferred to Section 7.4. Let's proceed with the properties of $f_{n,\boldsymbol{j}_{-d}}^+$ and $\widetilde{\mathcal{F}}_{n,\boldsymbol{j}_{-d}}^+$. By (P2), $f_{n,\boldsymbol{j}_{-d}}^+$ is zero outside $D_{\boldsymbol{j}_{-d}}$ and we can combine them together to define

$$f_{n,\Sigma}(\boldsymbol{x}_{-d}) = \sum_{\boldsymbol{j}_{-d} \in J_M} f_{n,\boldsymbol{j}_{-d}}^+(\boldsymbol{x}_{-d}). \tag{8}$$

Easy to see that $f_{n,\Sigma}(\boldsymbol{x})$ is still a ReLU network. Correspondingly, define such structured DNN family to be $\mathcal{F}_n$, which is $\widetilde{\mathcal{F}}_{n,\boldsymbol{j}_{-d}}^+$ stacked in parallel for $\boldsymbol{j}_{-d} \in J_M$. See Figure 3 for illustration. Overall, the size of $\mathcal{F}_n$ satisfies $L_n \lesssim \max_{\boldsymbol{j}_{-d}}\{\widetilde{L}_{n,\boldsymbol{j}_{-d}}\}$, $N_n \lesssim M^{d-1} \cdot \max_{\boldsymbol{j}_{-d}}\{\widetilde{N}_{n,\boldsymbol{j}_{-d}}\}$. Recall that a larger κ demands a larger network size for approximation. Thus the requirement on L_n, N_n is mainly determined by the region with the biggest $K(\boldsymbol{x})$.

Due to the formulation of $\mathcal{F}_n$, $\widehat{f}_n$ can be written in the form of (8) as $\widehat{f}_n = \sum_{\boldsymbol{j}_{-d} \in J_M} \widehat{f}_{n,\boldsymbol{j}_{-d}}$. Denote $\widetilde{f}_{n,\boldsymbol{j}_{-d}}$ as the projection of f^* to $\mathcal{F}_{n,\boldsymbol{j}_{-d}}$, i.e., $\widetilde{f}_{n,\boldsymbol{j}_{-d}} = \operatorname{argmin}_{f \in \mathcal{F}_{n,\boldsymbol{j}_{-d}}} \|f - f^*\|_\infty$. Under event E_ξ, we have that for any $\boldsymbol{j}_{-d} \in J_M$,

$$R_{n,\boldsymbol{j}_{-d}}(\widehat{f}_n) = R_{n,\boldsymbol{j}_{-d}}(\widehat{f}_{n,\boldsymbol{j}_{-d}}) = \min_{f \in \widetilde{\mathcal{F}}_{n,\boldsymbol{j}_{-d}}} R_{n,\boldsymbol{j}_{-d}}(f) \le R_{n,\boldsymbol{j}_{-d}}(\widetilde{f}_{n,\boldsymbol{j}_{-d}}). \tag{9}$$

The second equality is guaranteed by event E_ξ and property (P1). The last inequality is due to empirical risk minimization and the fact that $\widetilde{f}_{n,\boldsymbol{j}_{-d}} \in \widetilde{\mathcal{F}}_{n,\boldsymbol{j}_{-d}}$. Equation (9) indicates that the global empirical minimizer within $\mathcal{F}_n$ also gives rise to the empirical minimizer locally within each $D_{\boldsymbol{j}_{-d}}$.

Remark 8 (Structured DNN) *The constructed DNN classifier family $\mathcal{F}_n$ has special structures and is sparsely connected as illustrated in Figure 3. Such a structural requirement is not uncommon in nonparametric studies of deep learning where almost all DNN estimators are constructed with special structures (Schmidt-Hieber, 2020; Bauer et al., 2019; Imaizumi and Fukumizu, 2018). The idea of partitioned classifiers is also common outside deep learning, e.g., decision trees (Quinlan, 1986), mixture of experts (Shazeer et al., 2017), cascades (Alpaydin and Kaynak, 1998), etc.*

3.4 Global convergence analysis

In this section, we evaluate the convergence rate of the global classifier $\widehat{f}_n \in \mathcal{F}_n$. Through our construction process, we have shown that for any $\boldsymbol{j}_{-d} \in J_M$, (7) satisfies under E_ξ.

13

Intuitively, even though $K(\boldsymbol{x})$ can vary along the decision boundary, the convergence rate will be dominated by the region around $\text{argmin}_{\boldsymbol{x}} K(\boldsymbol{x})$. We have the following theorem on the overall convergence rate.

Theorem 9 *Under the smooth boundary fragments setting (1), assume conditions (M1,M2). Denote $\kappa^- = \inf_{\boldsymbol{x} \in [0,1]^d} K(\boldsymbol{x}), \kappa^+ = \sup_{\boldsymbol{x} \in [0,1]^d} K(\boldsymbol{x})$. Let $\mathcal{F}_n$ be a ReLU DNN family with proper architectures specified in Section 3.3 and size constraint*

$$N_n L_n = O\left(n^{\frac{\kappa^+(d-1)/2}{(\kappa^++2)\beta+(d-1)\kappa^+}} \cdot \log^{d+1}(n) \right).$$

Then, with probability tending to one, the empirical 0-1 loss minimizer within $\mathcal{F}_n$ satisfies

$$\inf_{\widehat{f}_n \in \mathcal{F}_n} \sup_{G^* \in \mathcal{G}_\beta^*} \mathbb{E}(R(\widehat{f}_n) - R(G^*)) = \widetilde{O}\left(n^{-\frac{(\kappa^-+1)\beta}{(\kappa^-+2)\beta+(d-1)\kappa^-}} \right).$$

Notice that the size constraint on $N_n L_n$ only concerns κ^+ while the final convergence rate only depends on κ^-. This is consistent with Remark 6 where we state that the bottleneck for convergence is κ^- while that for approximation is κ^+. The "with high probability" argument comes from the E_ξ, which is an artifact of the proof and may be relaxed.

The convergence rate in Theorem 9 matches the lower bound in Theorem 4 and is statistically optimal up to a logarithmic term. Combining Theorems 5 and 9, we conclude that, when $K(\boldsymbol{x})$ is highly non-constant in the sense that $\kappa^+ \to \infty$, the proposed localized classifier $\widehat{f}_n$ has faster rate of convergence than the regular DNN classifier.

4. Main results – breaking the "curse of dimensionality"

High-dimensional data are often structured. Taking images as an example, nearby pixels are often highly correlated while distant ones are more independent. The support of d-dimensional real images is speculated to be degenerate since not every combination of pixel values is relatable to real-life objects. To understand why deep learning models are not struggling in ultra-high dimensions, it is of great importance to investigate how well they adapt to various structures underlying the data. To this end, many low-dimensional structures have been investigated, e.g., low-dimensional manifold (Schmidt-Hieber, 2019; Hamm and Steinwart, 2020), hierarchical interaction max-pooling model (Kohler et al., 2020), teacher student setting (Hu et al., 2020), etc.

In this section, as a proof of concept, we focus on a particular structural assumption – compositional smoothness structure (Schmidt-Hieber, 2020) – and show that DNN classifiers can adapt to this low-dimensional structure and optimal convergence rate that only depends on the effective dimension is achievable.

4.1 Smooth boundary with compositional structure

Recall the smooth boundary fragment setting (1). Instead of β-smooth, we assume f^* to be compositions of smooth functions. This assumption is first considered in Schmidt-Hieber (2020) for the regression function and here we are borrowing it for classification. Assume f^* is of the form

$$f^* = h_l \circ h_{q-1} \circ \ldots \circ h_1 \circ h_0 \tag{10}$$

where each $h_i : [a_i, b_i]^{d_i} \to [a_{i+1}, b_{i+1}]^{d_{i+1}}$ and denote the components of h_i by $\{h_{ij}\}_{j=1}^{d_{i+1}}$. Let t_i be the maximal number of variables h_{ij}'s depend on. Thus, each h_{ij} is a t_i-variate function. We further assume that each function h_{ij} shares the same Hölder smoothness β_i. Define

$$\bar{\beta}_i := \beta_i \prod_{j=i+1}^{l} (\beta_j \wedge 1), \qquad i^* = \operatorname*{argmax}_{i=0,1,\cdots,l} \; n^{-\frac{2\bar{\beta}_i}{2\bar{\beta}_i + t_i}}.$$

The overall effective smoothness and effective dimension of f^* in (10) can be described by $\beta^* = \bar{\beta}_{i^*}$ and $d^* = t_{i^*}$, respectively. In this setting, $n^{-2\beta^*/(2\beta^*+d^*)}$ is proven to be the best possible square loss estimation rate from regression (Schmidt-Hieber, 2020). Let $\mathcal{C}(d^*, \beta^*)$ be the corresponding classifier class. The compositional smoothness is a generalization of the regular smoothness. If $f^*(\boldsymbol{x})$ is a $(d-1)$-dimensional β-smooth function as in (1), then $l = 0, \beta^* = \beta, d^* = d - 1$.

4.2 Optimal rate of convergence

We first establish the following convergence rate lower bound under the proposed setting.

Theorem 10 *Assume condition (M1) with noise exponent $\kappa = \inf_{\boldsymbol{x} \in \partial G^*} K(\boldsymbol{x})$ and the compositional smoothness structure (10) on the boundary fragment assumption (1). The 0-1 loss excess risk has the following lower bound*

$$\inf_{\widehat{f}_n} \sup_{C^* \in \mathcal{C}(d^*, \beta^*)} \mathbb{E}[\mathcal{E}(\widehat{f}_n, C^*)] \gtrsim \left(\frac{1}{n}\right)^{\frac{\beta^*(\kappa+1)}{\beta^*(\kappa+2)+d^*\kappa}}.$$

The convergence rate lower bound adapts to the compositional smoothness structure and only depends on β^* and d^*. Next, we evaluate how well DNNs can recover the decision boundary.

To handle the compositional smoothness, we utilize the corresponding ReLU DNN family in Schmidt-Hieber (2020) as our building block and go through the same construction process as in Section 3.3. To distinguish from the regular case in Section 3, we add $*$ to the notations of the network family. Locally, similar to the requirement in Theorem 5, we modify $\widetilde{\mathcal{F}}_{n,\boldsymbol{j}_{-d}}$ according to Schmidt-Hieber (2020) to $\widetilde{\mathcal{F}}^*_{n,\boldsymbol{j}_{-d}}$ with $\widetilde{L}^*_n \asymp \log(n)$,

$$\widetilde{N}^*_n \asymp n^{\frac{\kappa^+(\kappa^-+1)d^*}{(\kappa^-+2)(\kappa^++1)\beta^*+d^*\kappa^+(\kappa^-+1)}}, \quad \widetilde{S}^*_n \asymp n^{\frac{\kappa^+(\kappa^-+1)d^*}{(\kappa^-+2)(\kappa^++1)\beta^*+d^*\kappa^+(\kappa^-+1)}} \cdot \log(n),$$

which is sparsely connected. Similarly, we extend $\widetilde{\mathcal{F}}^*_{n,\boldsymbol{j}_{-d}}$ to $\widetilde{\mathcal{F}}^{+*}_{n,\boldsymbol{j}_{-d}}$ and stack them together to get $\mathcal{F}^*_n$ for global estimation as in Section 3.3. Similar to Theorem 9, the DNN family $\mathcal{F}^*_n$ in Theorem 11 is also constructed with structures and sparsely-connected. The size requirement is slightly different comparing to that in Theorem 9, due to a different approximation scheme.

Theorem 11 *Under the compositional smoothness setting (10), assume condition (M1,M2) and denote $\kappa^- = \inf_{\boldsymbol{x} \in [0,1]^d} K(\boldsymbol{x})$, $\kappa^+ = \sup_{\boldsymbol{x} \in [0,1]^d} K(\boldsymbol{x})$. Let $\mathcal{F}^*_n$ be a ReLU DNN family with proper architectures and size constraint $L^*_n \asymp \log(n)$,*

$$N^*_n \asymp n^{\frac{\kappa^+ d^*}{(\kappa^++2)\beta^*+d^*\kappa^+}} \log^{d-1}(n), \quad S^*_n \asymp n^{\frac{\kappa^+ d^*}{(\kappa^++2)\beta^*+d^*\kappa^+}} \log^d(n).$$

15

Then, with probability $\xrightarrow{n\to\infty} 1$, the empirical 0-1 loss minimizer within $\mathcal{F}_n$ satisfies

$$\inf_{\widehat{f}_n\in\mathcal{F}_n^*} \sup_{C^*\in\mathcal{C}(d^*,\beta^*)} \mathbb{E}(R(\widehat{f}_n)-R(C^*)) = \widetilde{O}\left(n^{-\frac{(\kappa^-+1)\beta^*}{(\kappa^-+2)\beta^*+\kappa^-d^*}}\right).$$

To further illustrate the power of our Theorem 11, we consider a special case where the $f^*(\boldsymbol{x})$ is a $(d-1)$-dimensional additive function, i.e.,

$$f^*(\boldsymbol{x}_{-d}) = \sum_{i<d} f_i^*(x_i) = h_1 \circ h_0, \tag{11}$$

where $h_0(x_1,\cdots,x_{d-1}) = (g_1(x_1),\cdots,g_{d-1}(x_{d-1}))$ and $h_1(x_1,\cdots,x_{d-1}) = x_1+\cdots+x_{d-1}$. In this case, $l=1, \boldsymbol{d}=(d-1,d-1), \boldsymbol{t}=(1,d-1)$. Assume each $g_i(\boldsymbol{x})$ has the same smoothness β. Then $\boldsymbol{\beta}=(\beta,\infty)$ and we have the following corollary for the convergence rate.

Corollary 12 *Under the same setting as in Theorem 11, further assume the additive structure (11) on the boundary function. Then the excess risk convergence rate becomes* $\widetilde{O}\left(n^{-\frac{(\kappa+1)\beta}{(\kappa+2)\beta+\kappa}}\right).$

The compositional smoothness assumption (10) may not be the best choice for describing natural data. But as a proof of concept, we have shown that DNN classifiers can adapt to this specific low-dimensional assumption while achieving the optimal convergence rate and avoid the "curse of dimensionality". In our localized analysis, various other low-dimensional structural assumptions on the decision boundary can be utilized to break the "curse of dimensionality", e.g., low-dimensional manifold support (Schmidt-Hieber, 2019), low intrinsic dimension (Hamm and Steinwart, 2020), etc. We conjecture that as long as DNNs can adapt to these assumptions in the regression setting and achieve optimal convergence rates, such adaptions can be potentially extended to the smooth boundary fragment classification setting.

5. Numerical experiments

In this section, we corroborate our theory with numerical experiments on 2-dimensional synthetic data. Let the marginal distribution of $\boldsymbol{x}$ be uniform on $[0,1]^2$, i.e., $p(\boldsymbol{x})+q(\boldsymbol{x})\equiv 2$. Consider the optimal decision boundary given by $x_2=f^*(x_1)$, where $f^*(x_1) = \cos(6\pi x_1)/4+1/2$. Denote $\delta(\boldsymbol{x}) = \frac{4}{3}(x_2-f^*(x_1))$, which ranges from -1 to 1, representing the signed distance to the decision boundary along x_2. Specify the conditional probability as

$$2\eta(\boldsymbol{x})-1 = \text{sign}(\delta(\boldsymbol{x}))\cdot\delta(\boldsymbol{x})^{K(\boldsymbol{x})}.$$

It's easy to verify that $\eta(\boldsymbol{x})$ is well-defined and (M1) holds with $K(\boldsymbol{x})$. Due to the symmetry of f^* around $1/2$, we have $\mathbb{P}(\eta(\boldsymbol{x})>1/2)=1/2$. Notice that given $\eta(\boldsymbol{x})$, we can write $p(\boldsymbol{x})=2\eta(\boldsymbol{x})$ and $q(\boldsymbol{x})=2-2\eta(\boldsymbol{x})$. To reflect the inconsistent separation, we choose $K(\boldsymbol{x})$ to be piecewise linear, where (M2) also holds with $\alpha=1$. Specifically, let $K(x_1)=1/k$ for
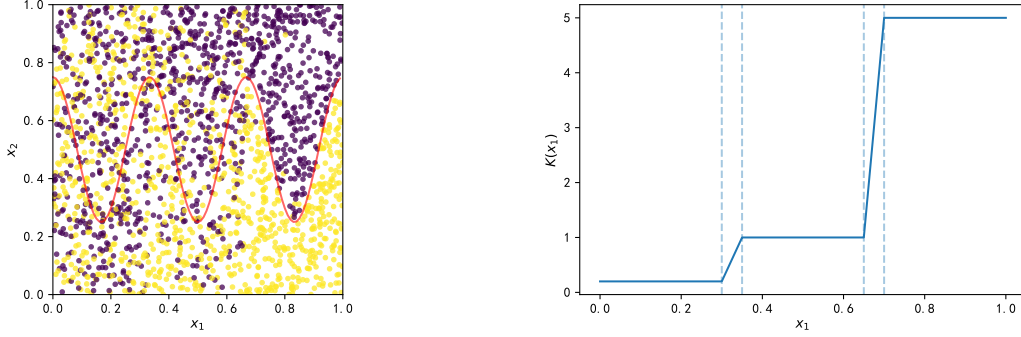
Figure 4: (a) 2000 data points in $[0,1]^2$ with $k = 100$. The colors represent different classes and the red curve is the optimal decision boundary. From left to right, the separation measured by $K(\boldsymbol{x})$ increases. (b) Illustration of the designed $K(\boldsymbol{x})$ in the synthetic data when $k = 5$.
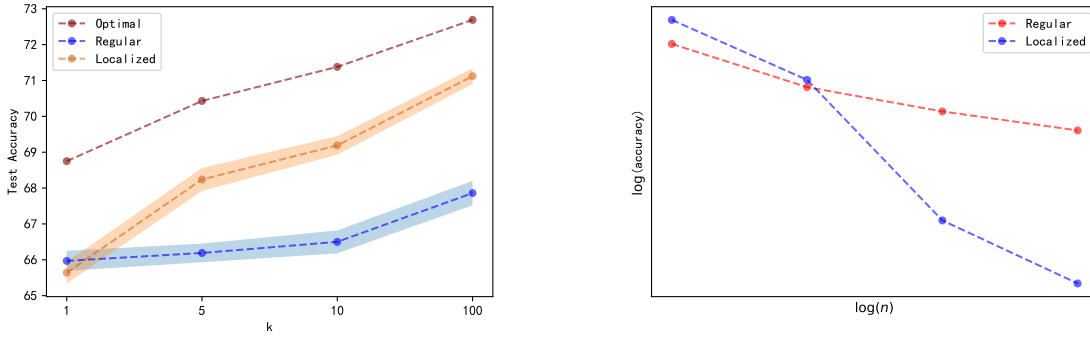

Figure 5: (a) The test accuracy curves for different classifiers vs. $k = 1, 5, 10, 100$. As k gets larger, the separation between classes are stronger but the consistency along the boundary is weaker. The bands represent 2 times the standard deviation from 10 replications. The optimal classifier refers to the Bayes classifier using $\eta(\boldsymbol{x})$. (b) The log-log curve of the empirical excess risk vs. sample size when $k = 10$. The deeper the slope, the faster the convergence rate.

$x_1 \in [0, 0.3]$, $K(x_1) = 1$ for $x_1 \in [0.35, 0.65]$, and $K(x_1) = k$ for $x_1 \in [0.7, 1]$. When $k = 1$, (N) and (N$^+$) both hold with $\kappa = 1$ and the separation is consistent. The larger the k, the more inconsistent the separation. We choose k to vary in $\{1, 5, 10, 100\}$. Figure 4(b) draws $K(\boldsymbol{x})$ when $k = 5$ and Figure 4(a) illustrates the data distribution with 2000 samples when $k = 100$.

On the synthetic data, we train two classifiers: the regular DNN classifier ($\widetilde{f}_n$) and the localized DNN classifier ($\widehat{f}_n$). To be more specific, we choose $\widetilde{f}_n$ to be a 3-layer ReLU

17

network with width 250 and $\widehat{f}_n$ to be the composition of $M = 5$ local ReLU classifiers, each with depth 3 and width 100. The total number of weights for $\widetilde{f}_n$ is slightly larger than that of $\widehat{f}_n$. As a surrogate to the hard-to-optimize 0-1 loss, we choose cross-entropy as the training loss, whose effectiveness in recovering the 0-1 loss minimizer has been investigated (Kohler and Langer, 2020). For each setup, we run 10 replications and monitor the test accuracy on unseen test data. The results are presented in Figure 5(a), where we can see that when $k = 1$, both regular and localized classifiers have comparable performances, but as k gets larger, the gap widens and the localized classifier significantly outperforms. This is consistent with our theoretical findings. The superiority in convergence rate is better reflected in Figure 5(b), where we plot the log-log curve between the empirical excess risk and sample size when $k = 5$. The slope for localized classifiers is significantly steeper, indicating faster convergence rate. More experiment details can be found in Section 7.7.

6. Discussion

This work aims to close the gap between DNN's empirical success and its rate suboptimality in recovering smooth decision boundaries with Tsybakov's noise condition (N). Through a finer-grained analysis of the convergence rate, we uncover a potential source of the suboptimality, i.e., inconsistency of the boundary separation. Correspondingly, we propose a new DNN classifier using a divide-and-conquer technique and a novel separation condition that is a localized version of the classical (N), under which optimal convergence rates are established for DNN classifiers with proper architectures. The established statistical optimality can adapt to low-dimensional structures underlying the data and in turn, may explain the absence of the "curse of dimensionality" in practice.

However, there are many limitations of this work. First, the nonparametric framework analyzes properties of the empirical risk minimizer, e.g., convergence rate, sample complexity, etc., but doesn't provide guidance to practical optimization and how to find them. Second, the proposed localized separation condition, like Tsybakov's noise condition, is of theoretical interest and cannot be verified for real data such as natural images. Similarly, this work only considers the empirical 0-1 loss minimizers, which is also out of theoretical interest rather than practical relevance since it's the most natural and fundamental classification loss. Given that the objective of interest is the 0-1 loss excess risk, the close relationship between the empirical 0-1 loss and the expected 0-1 loss makes our theoretical analysis more straightforward. For future work, it is intriguing to study how DNN classifiers perform under popular surrogate losses such as hinge loss or cross-entropy. Our numerical experiments conducted using cross-entropy indicate similar results may also hold for empirical surrogate loss minimizers. The technical challenge in extending our analysis to surrogate losses lies in additional analysis on the closeness between the minimizers of empirical 0-1 loss and surrogate losses. Our theory can be further strengthened if we could relax the structural requirement for the DNN family to be more general, e.g, fully-connected, or extend the analysis to popular networks such as convolutional neural networks (Krizhevsky et al., 2012), residual neural networks (He et al., 2016), and Transformers (Vaswani et al., 2017).

We believe that the proposed localized separation condition and the corresponding localized analysis are general and potentially transferable to other classification methods, as

long as they can take advantage of the divide-and-conquer strategy. It is worth mentioning similar ideas have been explored in retrieval-based classification models. For instance, Basu et al. (2023) demonstrated that decomposing learning into local sub-tasks allows for simpler parametric components while maintaining strong overall performance, providing theoretical underpinnings for the empirical success of retrieval augmentation in modern machine learning. We hope more investigations can be inspired along this line.

7. Technical details

This paper studies classification under the smooth boundary fragment setting (1) with Tsybakov's noise condition (Mammen and Tsybakov, 1999). First, we provide a proof for Lemma 3, which characterizes the relationship between the classical (N) and the proposed intermediate (N$^+$).

Proof of Lemma 3

Proof Recall the definition of $K(\boldsymbol{x})$ and condition (M1) that there exists $\epsilon_0 > 0$ and $0 < C_{\epsilon_0} < \infty$ such that for all $\boldsymbol{x} \in \partial G^*$ and any $0 < t < \epsilon_0$,

$$\frac{1}{C_{\epsilon_0}} \leq \frac{m_{\boldsymbol{x}}(t)}{|t|^{1/K(\boldsymbol{x})}} \leq C_{\epsilon_0}.$$

The key idea is to treat $\boldsymbol{x}_{-d}$ and x_d separately. For any $0 < t < \epsilon_0^{1/\kappa^+}/C_{\epsilon_0}$, we can write

$$
\begin{aligned}
\mathbb{Q}(\boldsymbol{x} \in [0,1]^d : |p(\boldsymbol{x}) - q(\boldsymbol{x})| \leq t) &= \int_{[0,1]^{d-1}} \int_0^1 \mathbb{I}\{|p(\boldsymbol{x}) - q(\boldsymbol{x})| \leq t\} dx_d d\boldsymbol{x}_{-d} \\
&\leq \int_{[0,1]^{d-1}} \int_{-1}^1 \mathbb{I}\{m_{\boldsymbol{x}_{-d}}(u) \leq t\} du d\boldsymbol{x}_{-d} \\
&\leq 2 \int_{[0,1]^{d-1}} \int_0^1 \mathbb{I}\{\frac{u^{1/K(\boldsymbol{x})}}{C_{\epsilon_0}} \leq t\} du d\boldsymbol{x}_{-d} \\
&\leq 2 \int_{[0,1]^{d-1}} \int_0^1 \mathbb{I}\{u^{1/\kappa^-} \leq tC_{\epsilon_0}\} du d\boldsymbol{x}_{-d} \\
&= 2 \int_{[0,1]^{d-1}} (C_{\epsilon_0} t)^{\kappa^-} d\boldsymbol{x}_{-d} \\
&= 2 C_{\epsilon_0} t^{\kappa^-}.
\end{aligned}
$$

The first equality is by definition. The second inequality is due to condition (M1), which indicates that locally, $m_{\boldsymbol{x}} \geq \epsilon_0^{1/K(x)}/C_{\epsilon_0} \geq \epsilon_0^{1/\kappa^+}/C_{\epsilon_0}$. Now consider (N$^+$) and a set $G \subset [0,1]^d$ containing part of the decision boundary. Notice that only lower bound is concerned so if suffices to prove the inequality for a subset $\subset G$. Without loss of generality, let the

subset be $[0, \epsilon_G]^d$. The analysis is similar to that in $[0, 1]$ and for t small enough we have

$$
\begin{aligned}
\mathbb{Q}(\boldsymbol{x} \in G : |p(\boldsymbol{x}) - q(\boldsymbol{x})| \leq t) &= \int_{[0,\epsilon_G]^{d-1}} \int_{-\epsilon_G}^{\epsilon_G} \mathbb{I}\{m_{\boldsymbol{x}_{-d}}(u) \leq t\} du d\boldsymbol{x}_{-d} \\
&\geq \int_{[0,\epsilon_G]^{d-1}} \int_{-\epsilon_G}^{\epsilon_G} \mathbb{I}\{C_{\epsilon_0} u^{1/K(\boldsymbol{x})} \leq t\} du d\boldsymbol{x}_{-d} \\
&\geq \int_{[0,\epsilon_G]^{d-1}} \int_{-\epsilon_G}^{\epsilon_G} \mathbb{I}\{u^{1/\kappa^+} \leq t/C_{\epsilon_0}\} du d\boldsymbol{x}_{-d} \\
&\geq C_G t^{\kappa^+},
\end{aligned}
$$

where C_G is some constant depending on the set G and C_ϵ. ∎

Next, we prove some useful lemmas utilizing the proposed localized separation condition (M1,M2). Let $G_f := \{\boldsymbol{x} \in \mathcal{X} : f(\boldsymbol{x}_d) - x_d \geq 0\}$. Then we have the following lemma characterizing the relationship between $d_\triangle$ and $d_{p,q}$.

Lemma 13 *Under assumption (M1), further assume on some $D \subset \mathcal{X}$, $0 < \kappa^- \leq K(\boldsymbol{x})$ for all $\boldsymbol{x} \in D$. For any set $G = G_f \subset D$ satisfying $\|f - f^*\|_\infty \leq \epsilon_0$, the following inequality holds*

$$
d_\triangle(G, G^*)^{\frac{\kappa^- + 1}{\kappa^-}} \lesssim d_{p,q}(G, G^*).
$$

Proof Let $\delta(\boldsymbol{x}_{-d}) := |f(\boldsymbol{x}_{-d}) - f^*(\boldsymbol{x}_{-d})| \leq \epsilon_0$. Consider $G \triangle G^*$ in dimension x_d and $\boldsymbol{x}_{-d}$ separately and write $G \triangle G^* = ((G \triangle G^*)_{-d}, (G \triangle G^*)_d)$. Then

$$
\begin{aligned}
d_\triangle(G, G^*) &= \int_{(G \triangle G^*)_{-d}} \int_{(G \triangle G^*)_d} dx_d d\boldsymbol{x}_{-d} \\
&= \int_{(G \triangle G^*)_{-d}} \delta(\boldsymbol{x}_{-d}) d\boldsymbol{x}_{-d}
\end{aligned}
$$

Applying assumption (M1) and Jensen's inequality yields

$$
\begin{aligned}
d_{p,q}(G, G^*) &= \int_{G \triangle G^*} |p(\boldsymbol{x}) - q(\boldsymbol{x})| d\boldsymbol{x} \\
&= \int_{(G \triangle G^*)_{-d}} \int_0^{\delta(\boldsymbol{x}_{-d})} m_{\boldsymbol{x}}(t) dt d\boldsymbol{x}_{-d} \\
&\geq \int_{(G \triangle G^*)_{-d}} \int_0^{\delta(\boldsymbol{x}_{-d})} \frac{1}{C_{\epsilon_0}} t^{1/\kappa^-} dt d\boldsymbol{x}_{-d} \\
&\geq \frac{1}{C_{\epsilon_0}(1 + 1/\kappa^-)} \int_{(G \triangle G^*)_{-d}} \delta(\boldsymbol{x}_{-d})^{\frac{\kappa^- + 1}{\kappa^-}} d\boldsymbol{x}_{-d} \\
&\geq \frac{1}{C_{\epsilon_0}(1 + 1/\kappa^-)} d_\triangle(G, G^*)^{\frac{\kappa^- + 1}{\kappa^-}}
\end{aligned}
$$

∎

Convergence rate proof road map: As stated before, the excess risk can be decomposed into the stochastic (empirical) error and the approximation error. As DNN classifiers get larger, the variance becomes larger while the bias gets smaller. An optimal trade-off can be achieved by carefully choosing the size of the DNN function space, or equally, the approximation error. The better the characterization of the two types of error, the tighter the convergence upper bound. In the following, we investigate these two aspects separately.

7.1 Approximation error

DNNs are universal approximation (Cybenko, 1989). Approximating a generic d-variate β-Hölder smooth function using a DNN family $\widetilde{\mathcal{F}}_n$ has been well-established in literature. Note that here we use $\widetilde{\mathcal{F}}$ to distinguish from the overall DNN classifier family $\mathcal{F}$. Among others (Yarotsky, 2017; Kohler and Langer, 2021), Lu et al. (2020) show that if $\widetilde{\mathcal{F}}_n$ is large enough with depth $\widetilde{L} = O(L \log(L))$ and width $\widetilde{N} = O(N \log(N))$, the L_∞-approximation error is in the order of $O(N^{-2\beta/d} L^{-2\beta/d})$; Schmidt-Hieber (2020) proves that for any $\epsilon > 0$, there exists a neural network $\widetilde{f}_n \in \widetilde{\mathcal{F}}_n$ with $\widetilde{L}_n = O(\log n)$ layers and $\widetilde{S}_n = O(\epsilon^{-d/\beta} \log n)$ non-zero weights such that $\|\widetilde{f}_n - f^*\|_\infty \leq \epsilon$. The detailed theorems are listed below and they cover all the cases considered in this paper. It's worth emphasizing that approximation of smooth functions is not the focus of this work. We aim to better utilize existing approximation results to sharpen the bias bound of the excess risk.

Lemma 14 *[Theorem 1.1 in (Lu et al., 2020)] Given a d-variate β-Hölder smooth function f_0 with $\beta \in \mathbb{N}^+$, for any $N, L \in \mathbb{N}^+$, there exists a function f implemented by a ReLU DNN with width $C_1(N + 2) \log_2(8N)$ and depth $C_2(L + 2) \log_2(4L) + 2d$ such that*

$$\|f - f_0\|_{L^\infty([0,1]^d)} \leq C_3 \|f\|_{C^s([0,1]^d)} N^{-2s/d} L^{-2s/d},$$

where $C_1 = 17s^{d+1}3^d d$, $C_2 = 18s^2$, and $C_3 = 85(s + 1)^d 8^s$.

Lemma 15 *[Approximation Part of Theorem 1 in Schmidt-Hieber (2020)] Consider the d-variate nonparametric regression model for composite regression function f_0 in the class $\mathcal{H}(l, \boldsymbol{d}, \boldsymbol{t}, \boldsymbol{\beta}, R)$. There exists $\widetilde{f}_n$ in the network class $\mathcal{F}_n^{\mathrm{DNN}}(L_n, N_n, S_n, B_n, F_n)$ with $L_n \lesssim \log_2 n$, $B_n = 1$, $F_n \geq \max(R, 1)$,*

$$N_n \lesssim \max_{i=0,\cdots,q} n^{\frac{t_i}{2\beta_i + t_i}}, \quad S_n \lesssim \max_{i=0,\cdots,q} n^{\frac{t_i}{2\beta_i + t_i}} \log n.$$

such that

$$\|\widehat{f}_n - f_0\|_\infty \lesssim n^{-\beta^*/(2\beta^* + d^*)}.$$

Notice that the approximation error bounds in the above lemmas are almost rate-optimal. Detailed descriptions and proofs can be found in the referenced papers. Other neural network approximation results can also be potentially used in our localized analysis as long as the approximation rate is optimal up to a logarithmic term.

Two types of smooth functions are considered, one is general Hölder smooth function and the other is compositional smooth function. Since the former is a special case of the latter, most of the proofs are done on the more general compositional smooth setting (10).

7.2 Stochastic error

The techniques for controlling the stochastic error are from literature on *empirical process* where covering/bracketing number/entropy are widely used to upper bound the empirical error. More details can be found in van de Geer (2000); Kosorok (2007).

However, there are difficulties when it comes to DNNs. A subtle difference in analyzing neural networks lies in their complexity measurement. To be more specific, for some real-value function space $\mathcal{F}$, let the covering entropy be $H(\delta, \mathcal{F}, L_\infty)$. The typical entropy bound is of the form $A\delta^{-\rho}$ where A, ρ are some constants, e.g., for d-variant β-smooth functions, $\rho = d/\beta$. In comparison, that for DNNs is of the form $A\log(1/\delta)$ where $A = O(N^2 L^2)$, which diverges with network size.

Lemma 16 (Lemma 5 of Schmidt-Hieber (2020); Lemma 3 of Suzuki (2018)) *For any $\delta > 0$, a ReLU network family $\mathcal{F}$ satisfies*

$$\log \mathcal{N}(\delta, \mathcal{F}(L, N, S, B), \|\cdot\|_\infty) \leq 2L(S+1)\log(\delta^{-1}(L+1)(N+1)(B \vee 1)).$$

The difference makes existing results not directly applicable and non-trivial modifications have to be made. Below, we prove some empirical error bounds specifically for DNNs. Following the notations from van de Geer (2000) and Tsybakov (2004), let

$$v_n(h) = \sqrt{n} \int h(\boldsymbol{x}) d(P_n - P),$$

where P denotes the data distribution, i.e. $\boldsymbol{x} \sim P$ and P_n denotes the empirical distribution of $\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n$.

Lemma 17 (Theorem 5.11 in van de Geer (2000)) *For some function space $\mathcal{H}$ with $\sup_{h \in \mathcal{H}} \|h(\boldsymbol{x})\|_\infty \leq K$ and $\sup_{h \in \mathcal{H}} \|h(\boldsymbol{x})\|_{L_2(P)} \leq R$. If $a > 0$ satisfies: (1) $a \leq C_1 \sqrt{n} R^2/K$; (2) $a \leq 8\sqrt{n}R$;*

$$(3) \quad a \geq C_0 \left(\int_{a/64\sqrt{n}}^{R} H_B^{1/2}(u, \mathcal{F}, L_2(P)) du \vee R \right);$$

and (4) $C_0^2 \geq C^2(C_1 + 1)$. Then

$$\mathbb{P}\left(\sup_{h \in \mathcal{H}} \left| \sqrt{n} \int h d(P_n - P) \right| \geq a \right) \leq C \exp\left(-\frac{a^2}{C^2(C_1+1)R^2} \right),$$

where P_n is the empirical counterpart of P.

The next lemma investigates the modulus of continuity of the empirical process. It's similar to Lemma 5.13 in van de Geer (2000) but with a key difference in the entropy assumption (12), where the entropy bound contains n.

Lemma 18 *For a probability measure P, let $\mathcal{H}_n$ be a class of uniformly bounded (by 1) functions h in $L_2(P)$ depending on n. Suppose that the δ-entropy with bracketing satisfies for all $0 < \delta < 1$ small enough, the inequality*

$$H_B(\delta, \mathcal{H}_n, L_2(P)) \leq A_n \log(1/\delta), \tag{12}$$

where $0 < A_n = o(n)$. Let h_{0n} be a fixed element in $\mathcal{H}_n$. Let $\mathcal{H}_n(\delta) = \{h_n \in \mathcal{H}_n : \|h_n - h_{0n}\|_{L_2(P)} \leq \delta\}$. Then there exist constants $D_1 > 0, D_2 > 0$ such that for a sequence of i.i.d. random variables $\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n$ with probability distribution P, it holds that for all T large enough,

$$\mathbb{P}\left(\sup_{h_n \in \mathcal{H}_n(\sqrt{A_n/n})} \left|\int (h_n - h_{0n})d(P_n - P)\right| \geq T\frac{A_n}{n}\right)$$
$$\leq C\exp\left(-\frac{TA_n}{8C^2}\right)$$

and for n large enough,

$$\mathbb{P}\left(\sup_{\substack{h_n \in \mathcal{H}_n; \\ \|h_n - h_{0n}\| > \sqrt{A_n/n}}} \frac{|v_n(h_n) - v_n(h_{0n})|}{A_n^{1/2}\|h_n - h_{n0}\|} > D_1 x\right)$$
$$\leq D_2 e^{-A_n x}$$

for all $x \geq 1$.

Proof The main tool for the proof is Lemma 17. Replace $\mathcal{H}$ with $\mathcal{H}_n(\delta)$ in Lemma 17 and take $K = 4$, $R = \sqrt{2}\delta$ and $a = \frac{1}{2}C_1 A_n^{1/2}\delta$, with $C_1 = 2\sqrt{2}C_0$. Then (1) is satisfied if

$$\delta \geq \sqrt{\frac{A_n}{n}}, \tag{13}$$

under which, (2) and (3) is trivially satisfied when n is large enough. Choosing C_0 sufficiently large will ensure (4). Thus, for all δ satisfying (13), we have

$$\mathbb{P}\left(\sup_{h_n \in \mathcal{H}_n(\delta)} \left|\sqrt{n}\int (h_n - h_{0n})d(P_n - P)\right| \geq \frac{C_1}{2}A_n^{1/2}\delta\right)$$
$$\leq C\exp\left(-\frac{C_1 A_n}{16C^2}\right)$$

Let $B = \min\{b > 1 : 2^{-b} \leq \sqrt{A_n/n}\}$ and apply the peeling device. Then,

$$\mathbb{P}\left(\sup_{\substack{h_n \in \mathcal{H}_n; \\ \|h_n - h_{n0}\| > \sqrt{A_n/n}}} \frac{\left|\sqrt{n}\int (h_n - h_{n0})d(P_n - P)\right|}{A_n^{1/2}\|h_n - h_{n0}\|} \geq \frac{C_1}{2}\right)$$
$$\leq \sum_{b=0}^{B}\mathbb{P}\left(\sup_{h_n \in \mathcal{H}_n(2^{-b})} \left|\sqrt{n}\int (h_n - h_{n0})d(P_n - P)\right| \geq \frac{C_1}{2}A_n^{1/2}(2^{-b})\right)$$
$$\leq \sum_{b=0}^{B} C\exp\left(-\frac{C_1 A_n}{16C^2}\right) \leq 2C(\log n)\exp\left(-\frac{C_1 A_n}{16C^2}\right),$$

if $C_1 A_n$ is sufficiently large. ∎

Now that we have introduced necessary tools to analyze both the approximation error and the stochastic error, we continue to prove the main theorems in this work.

7.3 Localized convergence analysis

Let $\rho = \frac{d-1}{\beta}$ in the regular smooth case and $\rho^* = \frac{d^*}{\beta^*}$ for the compositional smooth case, which denote the complexity measurements for the target function space.

Proof of Theorem 5

Proof For ease of notation, we will write G_f and its defining function f interchangeably. For any $\epsilon > 0$, by construction, we can find $\widetilde{f}_n \in \widetilde{\mathcal{F}}_n$ such that $\|\widetilde{f}_n - f^*\|_\infty \leq \epsilon$. Within $D_{\boldsymbol{j}_{-d}}$, the 0-1 loss can be bounded as

$$
\begin{aligned}
d_{\boldsymbol{j}_{-d}}(\widetilde{f}_n, f^*) &= \int_{D_{\boldsymbol{j}_{-d}} : G_{\widetilde{f}_{n,\boldsymbol{j}_{-d}}} \triangle G_{f^*}} |p(\boldsymbol{x}) - q(\boldsymbol{x})| d\boldsymbol{x} \\
&\leq \int_{D_{\boldsymbol{j}_{-d}}} \int_0^\epsilon m_{\boldsymbol{x}}(t) dt d\boldsymbol{x}_{-d} \\
&\leq C_{\epsilon_0} \int_{D_{\boldsymbol{j}_{-d}}} \int_0^\epsilon t^{1/K(\boldsymbol{x})} dt d\boldsymbol{x}_{-d} \\
&\leq \frac{C_{\epsilon_0}}{M^{d-1}(1 + 1/\kappa^+)} \epsilon^{\frac{\kappa^+ + 1}{\kappa^+}}
\end{aligned}
$$

Since $\widehat{f}_{n,\boldsymbol{j}_{-d}}$ is the empirical risk minimizer within $\widetilde{\mathcal{F}}_n$, we have $R_{n,\boldsymbol{j}_{-d}}(\widehat{f}_{n,\boldsymbol{j}_{-d}}) \leq R_{n,\boldsymbol{j}_{-d}}(\widetilde{f}_n)$. Therefore,

$$
\begin{aligned}
d_{\boldsymbol{j}_{-d}}(\widehat{f}_{n,\boldsymbol{j}_{-d}}, f^*) \leq\, & d_{\boldsymbol{j}_{-d}}(\widetilde{f}_n, f^*) + [R_{n,\boldsymbol{j}_{-d}}(\widetilde{f}_n) - R_{n,\boldsymbol{j}_{-d}}(f^*) - d_{\boldsymbol{j}_{-d}}(\widetilde{f}_n, f^*)] \\
& + [R_{n,\boldsymbol{j}_{-d}}(f^*) - R_{n,\boldsymbol{j}_{-d}}(\widehat{f}_{n,\boldsymbol{j}_{-d}}) + d_{\boldsymbol{j}_{-d}}(\widehat{f}_{n,\boldsymbol{j}_{-d}}, f^*)] \\
:\leq\, & \frac{C_{\epsilon_0}}{M^{d-1}(1 + 1/\kappa^+)} \epsilon^{\frac{\kappa^+ + 1}{\kappa^+}} + I(\widetilde{f}_n, f^*) + I(\widehat{f}_{n,\boldsymbol{j}_{-d}}, f^*).
\end{aligned}
$$

For $I(\widetilde{f}_n, f^*)$, by Lemma 18, we have

$$
\begin{aligned}
I(\widetilde{f}_n, f^*) \leq\, & \sup_{\substack{f \in \widetilde{\mathcal{F}}_n : \|f - f^*\|_1 \\ \leq \sqrt{A_n/n}}} \left| R_{n,\boldsymbol{j}_{-d}}(f) - R_{n,\boldsymbol{j}_{-d}}(f^*) - d_{\boldsymbol{j}_{-d}}(f, f^*) \right| + \\
& \sqrt{\frac{A_n d_\triangle(\widetilde{f}_n, f^*)}{n}} \sup_{\substack{f \in \widetilde{\mathcal{F}}_n : \|f - f^*\|_1 \\ > \sqrt{A_n/n}}} \frac{\sqrt{n} \left| R_{n,\boldsymbol{j}_{-d}}(f) - R_{n,\boldsymbol{j}_{-d}}(f^*) - d_{\boldsymbol{j}_{-d}}(f, f^*) \right|}{\sqrt{A_n d_\triangle(f, f^*)}} \\
=\, & O_{\mathbb{P}}\left(\frac{A_n}{n}\right) + \sqrt{\frac{A_n d_\triangle(\widetilde{f}_n, f^*)}{n}} O_{\mathbb{P}}(1),
\end{aligned}
$$

where A_n is from the assumption (12). Similarly for $I(\widehat{f}_{n,\boldsymbol{j}_{-d}}, f^*)$, we have

$$I(\widehat{f}_{n,\boldsymbol{j}_{-d}}, f^*) = O_{\mathbb{P}}\left(\frac{A_n}{n}\right) + \sqrt{\frac{A_n d_\triangle(\widehat{f}_{n,\boldsymbol{j}_{-d}}, f^*)}{n}}\, O_{\mathbb{P}}(1).$$

By construction, $d_\triangle(\widetilde{f}_n, f^*) \le \epsilon$. Hence

$$d_{\boldsymbol{j}_{-d}}(\widehat{f}_{n,\boldsymbol{j}_{-d}}, f^*) \le \frac{C_{\epsilon_0}}{M^{d-1}(1 + 1/\kappa^+)}\epsilon^{\frac{\kappa^+ + 1}{\kappa^+}} + O_{\mathbb{P}}\left(\frac{A_n}{n}\right) +$$
$$\sqrt{\frac{A_n\left(d_\triangle(\widehat{f}_{n,\boldsymbol{j}_{-d}}, f^*) + \epsilon\right)}{n}}\, O_{\mathbb{P}}(1).$$

The last term dominates the second term if $\epsilon \gg n^{-1/(1+\rho)}$. Let $\epsilon \gtrsim n^{-\frac{1}{1+\rho+2/\kappa^+}}$. Then the first term $\epsilon^{\frac{\kappa^+ + 1}{\kappa^+}}$ dominates the $\sqrt{A_n\epsilon/n}$ term in the last term. Omitting the approximation error, i.e.

$$\epsilon^{\frac{\kappa^+ + 1}{\kappa^+}} \lesssim \sqrt{\frac{A_n}{n}}d_\triangle^{1/2}(\widehat{f}_{n,\boldsymbol{j}_{-d}}, f^*),$$

by Lemma 13 we have

$$d_{\boldsymbol{j}_{-d}}(\widehat{f}_{n,\boldsymbol{j}_{-d}}, f^*) \le \sqrt{\frac{A_n}{n}}d_\triangle^{1/2}(\widehat{f}_{n,\boldsymbol{j}_{-d}}, f^*)\, O_{\mathbb{P}}(1)$$
$$\le \sqrt{\frac{A_n}{n}}d_{\boldsymbol{j}_{-d}}(\widehat{f}_{n,\boldsymbol{j}_{-d}}, f^*)^{\frac{\kappa^-}{2(\kappa^- + 1)}}\, O_{\mathbb{P}}(1),$$

which simplifies to

$$d_{\boldsymbol{j}_{-d}}(\widehat{f}_{n,\boldsymbol{j}_{-d}}, f^*) = O_{\mathbb{P}}\left(\frac{A_n}{n}\right)^{\frac{\kappa^- + 1}{\kappa^- + 2}}.$$

From Lemma 16, we know $A_n = O(\widetilde{N}_n^2 \widetilde{L}_n^2)$. Combining with Lemma 14 and Lemma 15, we know that $A_n = \widetilde{O}(\epsilon^{-\rho})$ in both cases. Balancing the approximation error and the empirical error, choose

$$\epsilon \asymp n^{-\frac{\kappa^+(\kappa^- + 1)}{(\kappa^- + 2)(\kappa^+ + 1) + \rho\kappa^+(\kappa^- + 1)}}. \tag{14}$$

In this case, Lemma 14 requires $\widetilde{N}_n\widetilde{L}_n \asymp \epsilon^{1/2}\log^2(n)$ and $A_n \asymp \epsilon^{-\rho}\log^4(n)$. Hence

$$\mathbb{E}\, d_{\boldsymbol{j}_{-d}}(\widehat{f}_{n,\boldsymbol{j}_{-d}}, f^*) = O\left(\frac{1}{n}\right)^{\frac{\kappa^- + 1}{\kappa^- + 2 + \kappa^+\rho\left(\frac{\kappa^- + 1}{\kappa^+ + 1}\right)}} \cdot \log^4(n).$$
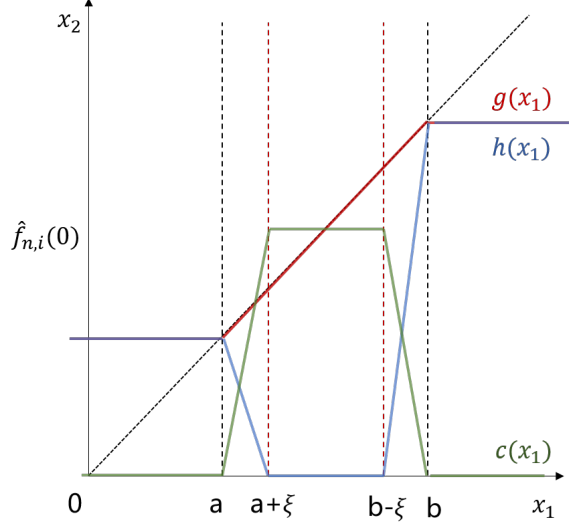
$\blacksquare$

Figure 6: Illustration of the constructed functions g, h, c in $d = 2$ case.

Remark 19 (Size of $\widetilde{\mathcal{F}}_n$) $\widetilde{\mathcal{F}}_n$ *should have the capacity to approximate f^* well such that the L_∞-error within the local region can be bounded by ϵ as chosen in (14). DNN approximating smooth functions have been well-studied and the DNN structure is pretty flexible. For example, following Schmidt-Hieber (2020), the DNN family should be large enough such that $\widetilde{L}_n \asymp \log(n)$ and $\widetilde{N}_n \asymp \epsilon^{-\rho/2}$. Lu et al. (2020) provide a more general approximation result and the DNN family only have to satisfy $\widetilde{N}_n \widetilde{L}_n \asymp \epsilon^{-\rho/2} \log^2 n$.*

Proof of Lemma 1

Proof As illustrated in the proof of Lemma 3, condition (N) and (N$^+$) essentially imply that $K(\boldsymbol{x}) = \kappa$ along the entire decision boundary, which indicates $\kappa^- = \kappa^+$. Following the same proof technique of Theorem 5, we can see that the optimal rate convergence rate can be achieved. ■

7.4 Verification of (P1,P2,P3)

Let's first consider the $d = 2$ case and focus on some region $D_i = \{(x_1, x_2) \in [0, 1]^2 : x_1 \in (a, b)\}$ with $b - a > 2\xi$. Let $f_n \in \widetilde{\mathcal{F}}$ be any DNN. Define three continuous piecewise linear functions

$$g(x_1) = \begin{cases} x_1 & \text{if } a \le x_1 \le b \\ a & \text{if } x_1 < a \\ b & \text{if } x_1 > b \end{cases} , \quad c(x_1) = \begin{cases} f_n(0) & \text{if } a + \xi \le x_1 \le b - \xi \\ 0 & \text{if } x_1 < a \text{ or } x_1 > b \\ \text{linear transition} & \text{else} \end{cases}$$

26

and

$$
h(x_1) = \begin{cases} 0 & \text{if } a + \xi \leq x_1 \leq b - \xi \\ a & \text{if } x_1 < a \\ b & \text{if } x_1 > b \\ \text{linear transition} & \text{else} \end{cases}
$$

Linear transition means linking the end points with a line segment. The constructed piecewise linear functions are illustrated in Figure 6. Let $f_{n,i}^+(x_1) := f_n(g(x_1)) - f_n(h(x_1)) + c(x_1)$. Then, it's easy to verify that

$$
f_{n,i}^+(x_1) = \begin{cases} f_n(x_1) & \text{if } a + \xi \leq x_1 \leq b - \xi \\ 0 & \text{if } x_1 < a \text{ or } x_1 > b \\ \text{piecewise linear} & \text{else} \end{cases}
$$

Therefore, (P1) and (P2) hold and now we evaluate (P3). The constructed g, h, c are all piecewise linear functions with at most 5 pieces. By Theorem 2.2 in Arora et al. (2016), they can all be represented by two-layer ReLU neural networks with width at most 5. $f_{n,i}^+(x_1)$ is constructed by composition and addition of ReLU networks, which correspond to stacking more layers and expanding the width respectively. Easy to see that $f_{n,i}^+(x_1)$ satisfies (P3).

In the $d > 2$ case, we can make similar constructions. Consider some region $D_{\boldsymbol{j}_{-d}}$ and denote $D_{\boldsymbol{j}_{-d}}^\circ := D_{\boldsymbol{j}_{-d}} \backslash D_\xi$. For each of the dimensions $x_1, \ldots, x_{d-1}$, we can define $g_i(x_i), h_i(x_i), c_i(x_i)$ separately as in the $d = 2$ case.

Let $g(\boldsymbol{x}_{-d}) = (g_1(x_1), \ldots, g_{d-1}(x_{d-1}))$, $h(\boldsymbol{x}_{-d}) = (h_1(x_1), \ldots, h_{d-1}(x_{d-1}))$, $c(\boldsymbol{x}_{-d}) = (c_1(x_1), \ldots, c_{d-1}(x_{d-1}))$ and $f_{n,\boldsymbol{j}_{-d}}^+ = (f_n \circ g - f_n \circ h + c)$. Then, it's easy to verify that

$$
f_{n,\boldsymbol{j}_{-d}}^+(\boldsymbol{x}_{-d}) = \begin{cases} f_n(\boldsymbol{x}_{-d}) & \text{if } \boldsymbol{x}_{-d} \in D_{\boldsymbol{j}_{-d}}^\circ \\ 0 & \text{if } \boldsymbol{x}_{-d} \notin D_{\boldsymbol{j}_{-d}} \\ \text{piecewise linear} & \text{else} \end{cases}
$$

Thus, (P1) and (P2) hold. For (P3), notice that $g(\boldsymbol{x}_{-d})$ can be viewed as a ReLU neural network with the same depth as $g_i(x_i)$ but $(d-1)$-times the width.

7.5 Global convergence analysis

The proof of Theorem 11 implies the proof of Theorem 9 as a special case since the latter only considers standard Hölder smoothness while the former considers the more general compositional smoothness. Below we provide the proof of the more general version.

Proof of Theorem 11

Proof Choose $\xi = 1/n^2$, $M = \log n$. Notice that $nM\xi(d-1) \to 0$, i.e., $\mathbb{P}(E_\xi) \to 1$ as $n \to \infty$ for any $d \geq 2$. Without loss of generality, assume $n = \Omega(\epsilon_0^{-(1+\rho)})$. Thus ξ is smaller than ϵ_0. Let

$$
\kappa_{\boldsymbol{j}_{-d}}^- := \min_{\boldsymbol{x} \in D_{\boldsymbol{j}_{-d}}} K(\boldsymbol{x}) \quad \text{and} \quad \kappa_{\boldsymbol{j}_{-d}}^+ := \max_{\boldsymbol{x} \in D_{\boldsymbol{j}_{-d}}} K(\boldsymbol{x}).
$$

Since $R_{n,\boldsymbol{j}_{-d}}(\widehat{f}_n) = R_{n,\boldsymbol{j}_{-d}}(\widehat{f}_{n,\boldsymbol{j}_{-d}}) \leq R_{n,\boldsymbol{j}_{-d}}(\widetilde{f}_n)$ for any $\boldsymbol{j}_{-d} \in J_M$ as in (9), Theorem 5 yields that

$$\sup_{f^* \in \mathcal{F}(d^*,\beta^*)} \mathbb{E}(R_{\boldsymbol{j}_{-d}}(\widehat{f}_n) - R_{\boldsymbol{j}_{-d}}(f^*)) \lesssim \left(\frac{1}{n}\right)^{\frac{\kappa_{\boldsymbol{j}_{-d}}^- + 1}{\kappa_{\boldsymbol{j}_{-d}}^- + 2 + \left(\frac{\kappa_{\boldsymbol{j}_{-d}}^- + 1}{\kappa_{\boldsymbol{j}_{-d}}^+ + 1}\right)\rho\kappa_{\boldsymbol{j}_{-d}}^+}} \cdot \log^4(n).$$

Then, the overall 0-1 loss excess risk can be decomposed as

$$\sup_{f^* \in \mathcal{F}(d^*,\beta^*)} \mathbb{E}(R(\widehat{f}_n) - R(f^*)) \leq \sum_{\boldsymbol{j}_{-d} \in J_M} \sup_{f^* \in \mathcal{F}(d^*,\beta^*)} \mathbb{E}(R_{\boldsymbol{j}_{-d}}(\widehat{f}_n) - R_{\boldsymbol{j}_{-d}}(f^*))$$

$$\lesssim \sum_{\boldsymbol{j}_{-d} \in J_M} \left(\frac{1}{n}\right)^{\frac{\kappa_{\boldsymbol{j}_{-d}}^- + 1}{\kappa_{\boldsymbol{j}_{-d}}^- + 2 + \left(\frac{\kappa_{\boldsymbol{j}_{-d}}^- + 1}{\kappa_{\boldsymbol{j}_{-d}}^+ + 1}\right)\rho\kappa_{\boldsymbol{j}_{-d}}^+}} \cdot \log^4(n).$$

By assumption (M2), we can write for any $\boldsymbol{j}_{-d} \in J_M$ that

$$\left(\frac{1}{n}\right)^{\frac{\kappa_{\boldsymbol{j}_{-d}}^- + 1}{\kappa_{\boldsymbol{j}_{-d}}^- + 2 + \left(\frac{\kappa_{\boldsymbol{j}_{-d}}^- + 1}{\kappa_{\boldsymbol{j}_{-d}}^+ + 1}\right)\rho\kappa_{\boldsymbol{j}_{-d}}^+}} = \left(\frac{1}{n}\right)^{\frac{\kappa_{\boldsymbol{j}_{-d}}^- + 1}{\kappa_{\boldsymbol{j}_{-d}}^- + 2 + \rho\kappa_{\boldsymbol{j}_{-d}}^- + \rho\frac{\kappa_{\boldsymbol{j}_{-d}}^+ - \kappa_{\boldsymbol{j}_{-d}}^-}{\kappa_{\boldsymbol{j}_{-d}}^+ + 1}}}$$

$$\leq \left(\frac{1}{n}\right)^{\frac{\kappa_{\boldsymbol{j}_{-d}}^- + 1}{\kappa_{\boldsymbol{j}_{-d}}^- + 2 + \rho\kappa_{\boldsymbol{j}_{-d}}^- + \rho C_K(\sqrt{d}/M)^\alpha}}$$

$$= \left(\frac{1}{n}\right)^{\frac{\kappa_{\boldsymbol{j}_{-d}}^- + 1}{\kappa_{\boldsymbol{j}_{-d}}^- + 2 + \rho\kappa_{\boldsymbol{j}_{-d}}^-} + \frac{(\kappa_{\boldsymbol{j}_{-d}}^- + 1)\rho C_K(\sqrt{d}/M)^\alpha}{(\kappa_{\boldsymbol{j}_{-d}}^- + 2 + \rho\kappa_{\boldsymbol{j}_{-d}}^- + \rho C_K(\sqrt{d}/M)^\alpha)(\kappa_{\boldsymbol{j}_{-d}}^- + 2 + \rho\kappa_{\boldsymbol{j}_{-d}}^-)}}$$

$$= O\left(\frac{1}{n}\right)^{\frac{\kappa_{\boldsymbol{j}_{-d}}^- + 1}{\kappa_{\boldsymbol{j}_{-d}}^- + 2 + \rho\kappa_{\boldsymbol{j}_{-d}}^-}}.$$

The last equality follows from the fact that $M = \log n$ and $n^{-1/\log n} = O(1)$. Since κ is defined as the overall minimum, under E_ξ, we have

$$\sup_{f^* \in \mathcal{F}(d^*,\beta^*)} \mathbb{E}(R(\widehat{f}_n) - R(f^*)) \lesssim \sum_{\boldsymbol{j}_{-d} \in J_M} \left(\frac{1}{n}\right)^{\frac{\kappa_{\overline{j}}^- + 1}{\kappa_{\overline{j}}^- + 2 + \rho\kappa_{\overline{j}}^-}} \cdot \log^4(n)$$

$$= O\left(n^{-\frac{(\kappa+1)\beta^*}{(\kappa+2)\beta^* + \kappa d^*}}(\log n)^{d+3}\right).$$

■

28
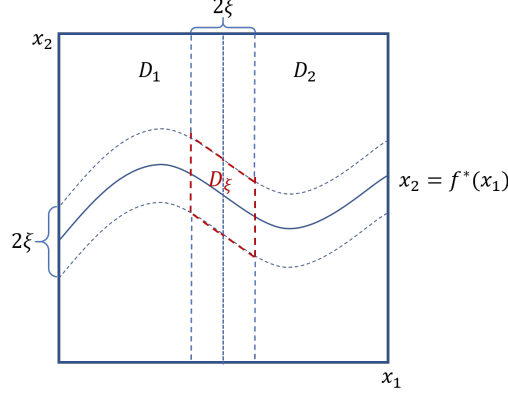
Figure 7: Illustration of region D_ξ in $d = 2, M = 2$ case.

Remark 20 (ξ, ϵ and B_n) E_ξ *is defined for technical purposes to join local regions $D_{\boldsymbol{j}_{-d}}$ so that (7) can hold with high probability. Note that in order to satisfy $nM^{d-1}\xi^{d-1} \to 0$, ξ can be chosen to be arbitrarily small, e.g., much smaller than the required approximation error ϵ. However, this is at the expense of larger DNN derivatives (piecewise linear part in the helper functions in the construction process) within the D_ξ band. See Figure 6 for illustration. If $\xi = o(\epsilon)$, the Lipschitz constants of the constructed c and h in the Section 7.4 will tend to infinity, which means the parameters of c and h will diverge, i.e., the corresponding B_n of $\mathcal{F}_n$ will not be bounded. Although the ReLU network family considered in Schmidt-Hieber (2020) has constrained $B_n = O(1)$, an unbounded B_n is not a problem, and such constraint is not employed in (Lu et al., 2020).*

In Theorem 9, if we want $B_n = O(1)$, we can utilize the approximation results with such constraint, e.g., Lemma 15 (Schmidt-Hieber, 2020), and modify the region D_ξ to be more efficient as shown in Figure 7, where we have $\mathbb{P}(x \in D_\xi) \lesssim M\xi^2 d$. If we further assume $\rho < 1$ and $\kappa^- > \frac{2}{1-\rho}$, we can ensure $\epsilon = O(\xi)$ by choosing $\xi \asymp n^{-\frac{\kappa^-}{\kappa^-+2+\rho\kappa^-}}$. In this case, $nM\xi^2 d \to 0$ and $\mathbb{P}(E_\xi) \to 1$ as $n \to \infty$.

7.6 Lower bound - proof of Theorem 4

The lower bound result comes from estimation of sets in the discriminative analysis (Mammen and Tsybakov, 1999) where two independent samples $\mathcal{X}^+ = \{\boldsymbol{x}_1^+, \cdots, \boldsymbol{x}_n^+\}$ and $\boldsymbol{x}^- = \{\boldsymbol{x}_1^-, \cdots, \boldsymbol{x}_m^-\}$ of $\mathbb{R}^d$-valued i.i.d. observations with unknown densities f or g respectively (w.r.t. a σ-finite measure Q) are given. The goal is to predict whether a new sample $\boldsymbol{x}$ is coming from f or g with a discrimination decision rule defined by a set $G \subset \mathbb{R}^d$ that we attribute $\boldsymbol{x}$ to f if $\boldsymbol{x} \in G$ and to g otherwise. Let the Bayes risk to be

$$R(G) = \frac{1}{2}\left(\int_{G^c} f(\boldsymbol{x})Q(d\boldsymbol{x}) + \int_G g(\boldsymbol{x})Q(d\boldsymbol{x})\right)$$

Denote $G^* = \{\boldsymbol{x} : f(\boldsymbol{x}) \geq g(\boldsymbol{x})\}$ to be the Bayes risk minimizer and consider the distance between two sets G_1, G_2 to be

$$d_{f,g}(G_1, G_2) = \int_{G_1 \triangle G_2} |f(\boldsymbol{x}) - g(\boldsymbol{x})| Q(d\boldsymbol{x}).$$

Let $\widetilde{G}_{m,n}$ be an empirical rule based on observations. The excess risk can be expressed as $R(\widetilde{G}_{m,n}) - R(G^*) = \frac{1}{2} d_{f,g}(\widetilde{G}_{m,n}, G^*)$. In the following, we establish how fast can the excess risk go to zero under the smooth boundary condition.

For positive constants c_1, c_2, η_0, κ and for a σ-finite measure Q, consider densities f, g on $\mathbb{R}^d$ w.r.t. Q and define class $\mathcal{F}$ of paired densities to be

$$\mathcal{F}_\mathcal{G} = \{(f,g) : Q\{\boldsymbol{x} \in \mathcal{X} : |f(\boldsymbol{x}) - g(\boldsymbol{x})| \leq \eta\} \leq c_2 \eta^\kappa \text{ for } 0 \leq \eta \leq \eta_0,$$
$$\{\boldsymbol{x} \in \mathcal{X} : f(\boldsymbol{x}) \geq g(\boldsymbol{x})\} \in \mathcal{G}, f(\boldsymbol{x}), g(\boldsymbol{x}) \leq c_1 \text{ for } x \in \mathcal{X}\}$$

Now let the base measure Q be the Lebesgue measure $\mathbb{Q}$ and recall $d_\triangle(G_1, G_2) = Q(G_1 \triangle G_2)$. The following lemma establishes the connection between $d_\triangle$ and $d_{f,g}$

Lemma 21 (Lemma 2 in (Mammen and Tsybakov, 1999)) *There exists a constant $c(\kappa)$ depending on κ such that for Lebesgue measurable subsets G_1 and G_2 of $\mathcal{X}$ and for $(f,g) \in \mathcal{F}_\mathcal{G}$,*

$$c(\kappa) d_\triangle^{(1+\kappa)/\kappa}(G_1, G_2) \leq d_{f,g}(G_1, G_2) \leq 2c_1 d_\triangle(G_1, G_2)$$

Proof of Theorem 4
Proof Without loss of generality, assume $n \leq m$ so we mainly focus on $\mathcal{X}^+$. Consider the subset of $\mathcal{F}_{\mathcal{G}_h}$ that contains all pairs (f, g_0), where g_0 is fixed and f belongs to a finite class of densities $\mathcal{F}_1$ that will be defined later. Then,

$$\sup_{(f,g) \in \mathcal{F}_{\mathcal{G}_h}} \mathbb{E}_{f,g} d_\triangle(\widetilde{G}_{m,n}, G^*) \geq \sup_{(f,g_0): f \in \mathcal{F}_1} \mathbb{E}_{f,g} d_\triangle(\widetilde{G}_{m,n}, G^*)$$

$$\geq \mathbb{E}_{g_0} \left[\frac{1}{|\mathcal{F}_1|} \sum_{f \in \mathcal{F}_1} \mathbb{E}_f [d_\triangle(\widetilde{G}_{m,n}, G^*) | y_1, \cdots, y_m] \right]$$

where $\mathbb{E}_f$ and $\mathbb{E}_{g_0}$ denotes the expectations w.r.t. the distributions of $(x_1, \cdots, x_n)$ and $(y_1, \cdots, y_m)$ when the underlying densities are f and g_0.

Recall the compositional assumption (10) and let

$$i^* \in \operatorname*{argmax}_{i=0,1,\cdots,q} n^{-\frac{2\bar{\beta}_i}{2\bar{\beta}_i + t_i}} \quad \text{and} \quad \beta^* = \beta_{i^*}$$

Further denote $B = \prod_{l=i^*+1}^{q} (\beta_l \wedge 1)$ and then $\beta^* = \beta^* B$. For simplicity, we give the proof for the case $d^* = t_{i^*} = 1$, that is the effective dimension of the smooth boundaries is 1 instead of $d - 1$. For this case, let $\phi \in \mathcal{C}_1^{\beta^*}(\mathbb{R}, 1)$ be a real-valued function supported on $[-1, 1]$ with $\phi(t) \geq 0$ for all t, $\max \phi(t) = 1$ and $\phi(0) = 1$. For $x = (x_1, \cdots, x_d) \in [0,1]^d$, define

$$g_0(\boldsymbol{x}) = (1 - \eta_0 - b_1) \mathbb{I}_{\{0 < x_2 < \frac{1}{2}\}} + \mathbb{I}_{\{\frac{1}{2} \leq x_2 < \frac{1}{2} + (\tau M^{-\beta^*})^B\}}$$
$$+ (1 + \eta_0 + b_2) \mathbb{I}_{\{\frac{1}{2} + (\tau M^{-\beta^*})^B \leq x_2 \leq 1\}}$$

where $M \geq 2$ is an integer to be specified later and $\tau \in (0,1)$ is a constant. $b_1 = (\tau M^{-\beta^*}/c_2)^{B/\kappa}$ and $b_2 > 0$ is chosen such that g_0 integrates to 1. For $j = 1, 2, \cdots, M$ and $t \in [0,1]$, let

$$\psi_j(t) = \tau M^{-\beta^*} \phi\left(M\left[t - \frac{j-1}{M}\right]\right)$$

Note that ψ_j is only supported on $[\frac{j-1}{M}, \frac{j}{M}]$. For vectors $\omega = (\omega_1, \cdots, \omega_M)$ with elements $w_j \in \{0,1\}$, define

$$b_\omega(t) = \sum_{j=1}^{M} \omega_j \psi_j(t) \tag{15}$$

Now we construct functions in $\mathcal{H}(l, \mathbf{d}, \mathbf{t}, \boldsymbol{\beta}, R)$. For $i < i^*$, let $g_i(\boldsymbol{x}) := (x_1, \cdots, x_{d_i})^\mathsf{T}$. For $i = i^*$ define $g_{i^*,\omega}(\boldsymbol{x}) = (b_\omega(x_1), 0, \cdots, 0)^\mathsf{T}$. For $i > i^*$, set $g_i(\boldsymbol{x}) := (x_1^{\beta_i \wedge 1}, 0, \cdots, 0)^\mathsf{T}$.

$$\widetilde{b}_\omega(\boldsymbol{x}) = g_q \circ \cdots \circ g_{i^*+1} \circ g_{i^*,\omega} \circ g_{i^*-1} \circ \cdots \circ g_0(\boldsymbol{x}) = b_\omega(x_1)^B$$

Notice that $\widetilde{b}_\omega(\boldsymbol{x}) \leq (\tau M^{-\beta^*})^B$. Let $\Omega = \{0,1\}^M$. Define

$$f_\omega(\boldsymbol{x}) = 1 + \left[\frac{\frac{1}{2} + (\tau M^{-\beta^*})^B - x_2}{c_2}\right]^{1/\kappa} \mathbb{I}_{\{\frac{1}{2} \leq x_2 \leq \frac{1}{2} + \widetilde{b}_\omega(\boldsymbol{x})\}} - b_3(\omega)\mathbb{I}_{\{\frac{1}{2} + \widetilde{b}_\omega(\boldsymbol{x}) < x_2 \leq 1\}}$$

where $b_3(\omega) > 0$ is chosen such that $f_\omega(x)$ integrates to 1. Note that both $g_0(\boldsymbol{x})$ and $f_\omega(\boldsymbol{x})$ are d-dimensional densities even though they seem to only depend on x_1 and x_2. Other entries follow independent uniform distribution on $[0,1]$ and don't show on the density formulas.

Set $\mathcal{F}_1 = \{f_\omega : \omega \in \Omega\}$ and we will show that $(f_\omega, g_0) \in \mathcal{F}_{\mathcal{G}_h}$ for all $\omega \in \Omega$. To this end, we need to verify that

(a) $f_\omega(\boldsymbol{x}) \leq c_1$ for $x \in K$;

(b) $\{\boldsymbol{x} \in \mathcal{X} : f_\omega(\boldsymbol{x}) \geq g_0(\boldsymbol{x})\} \in \mathcal{G}_h$;

(c) $Q\{\boldsymbol{x} \in \mathcal{X} : |f_\omega(\boldsymbol{x}) - g_0(\boldsymbol{x})| \leq \eta\} \leq c_2 \eta^\kappa$ for all $0 < \eta < \eta_0$.

For (a), since f_ω integrates to 1,

$$b_3(\omega) \leq \max_{\{\frac{1}{2} \leq x_2 \leq \frac{1}{2} + \widetilde{b}_\omega(\boldsymbol{x})\}} \left[\frac{\frac{1}{2} + (\tau M^{-\beta^*})^B - x_2}{c_2}\right]^{1/\kappa}$$

$$\leq \left[\frac{2\tau^B M^{-\beta^*}}{c_2}\right]^{1/\kappa} = \mathcal{O}(M^{-\beta^*/\kappa})$$

Thus, $f_\omega(\boldsymbol{x}) \leq c_1$ for c_1 and M large enough.
(b) is satisfied since

$$\{\boldsymbol{x} : f_\omega(\boldsymbol{x}) \geq g_0(\boldsymbol{x})\} = \{\boldsymbol{x} : 0 \leq x_2 \leq \frac{1}{2} + \widetilde{b}_\omega(x_1)\}$$

and by construction, $\widetilde{b}_\omega(\boldsymbol{x}) \in \mathcal{H}(l, \mathbf{d}, \mathbf{t}, \boldsymbol{\beta}, R)$ for τ small enough.

(c) follows that

$$Q\{\boldsymbol{x} \in \mathcal{X} : |f_\omega(\boldsymbol{x}) - g_0(\boldsymbol{x})| \leq \eta\}$$

$$\leq Q\left\{\boldsymbol{x} \in \mathcal{X} : \frac{1}{2} \leq x_2 \leq \frac{1}{2} + (\tau M^{-\beta^*})^B, \left[\frac{1/2 + (\tau M^{-\beta^*})^B - x_2}{c_2}\right]^{1/\kappa} \leq \eta\right\}$$

$$\leq Q\left\{\boldsymbol{x} \in \mathcal{X} : \frac{1}{2} + (\tau M^{-\beta^*})^B - c_2\eta^\kappa \leq x_2 \leq \frac{1}{2} + (\tau M^{-\beta^*})^B\right\}$$

$$\leq c_2\eta^\kappa$$

After verifying $(f_\omega, g_0) \in \mathcal{F}_{\mathcal{G}_h}$ for all $\omega \in \Omega$, we now establish how fast can

$$S := \frac{1}{|\mathcal{F}_1|} \sum_{f \in \mathcal{F}_1} \mathbb{E}_f[d_\triangle(\widetilde{G}_{m,n}, G^*)|y_1, \cdots, y_m]$$

go to zero. To this end, we use the Assouad's lemma stated in (Korostelev and Tsybakov, 2012) which is adapted to the estimation of sets.

For $j = 1, \cdots, M$ and for a vector $\omega = (\omega_1, \cdots, \omega_M)$, we write

$$\omega_{j0} = (\omega_1, \cdots, \omega_{j-1}, 0, \omega_{j+1}, \cdots, \omega_M)$$

$$\omega_{j1} = (\omega_1, \cdots, \omega_{j-1}, 1, \omega_{j+1}, \cdots, \omega_M)$$

For $i = 0$ and $i = 1$, let P_{ji} be the probability measure corresponding to the distribution of $x_1, \cdots, x_n$ when the underlying density is $f_{\omega_{ji}}$. Denote the expectation w.r.t. P_{ji} as $\mathbb{E}_{ji}$. Let

$$\mathcal{D}_j = \left\{\boldsymbol{x} \in \mathcal{X} : \frac{1}{2} + \widetilde{b}_{\omega_{j0}}(\boldsymbol{x}) < x_2 \leq \frac{1}{2} + \widetilde{b}_{\omega_{j1}}(\boldsymbol{x})\right\}$$

$$= \left\{\boldsymbol{x} \in \mathcal{X} : b_{\omega_{j0}}(x_1) < \left(x_2 - \frac{1}{2}\right)^{1/B} \leq b_{\omega_{j1}}(x_1)\right\}$$

$$= \left\{\boldsymbol{x} \in \mathcal{X} : b_{\omega_{j0}}(x_1) < \left(x_2 - \frac{1}{2}\right)^{1/B} \leq b_{\omega_{j0}}(x_1) + \psi_j(x_1)\right\}$$

Then

$$S \geq \frac{1}{2} \sum_{j=1}^M Q\{\mathcal{D}_j\} \int \min\{dP_{j1}, dP_{j0}\}$$

$$\geq \frac{1}{2} \sum_{j=1}^M \int_0^1 \psi_j(x_1)^B dx_1 \int \min\{dP_{j1}, dP_{j0}\}$$

$$\geq \frac{1}{2} \sum_{j=1}^M \tau^B M^{-\beta^*} \int \phi(Mt)^B dt \int \min\{dP_{j1}, dP_{j0}\}$$

$$\geq \frac{1}{4} \sum_{j=1}^M \tau^B M^{-\beta^*} \int \phi(Mt)^B dt \left[1 - H^2(P_{10}, P_{11})/2\right]^n$$

where $H(\cdot, \cdot)$ denotes the Hellinger distance.

$$H^2(P_{10}, P_{11}) = \int \left[\sqrt{f_{\omega_{10}}(\boldsymbol{x})} - \sqrt{f_{\omega_{11}}(\boldsymbol{x})} \right]^2 d\boldsymbol{x}$$

$$\leq \int_0^1 \left\{ \int_{\frac{1}{2}}^{\frac{1}{2}+\psi_1(x_1)^B} \left[1 - \sqrt{1 + \left(\frac{\frac{1}{2} + \tau^B M^{-\beta^*} - x_2}{c_2} \right)^{1/\kappa}} \right]^2 dx_2 \right.$$

$$\left. + \int_{\frac{1}{2}}^1 \left[\sqrt{1 - b_3(\omega_{10})} - \sqrt{1 - b_3(\omega_{11})} \right]^2 dx_2 \right\} dx_1$$

$$\leq \int_0^1 \int_{(\tau M^{-\beta^*})^B - \psi_1(x_1)^B}^{(\tau M^{-\beta^*})^B} \left[1 - \sqrt{1 + \left(\frac{v}{c_2} \right)^{1/\kappa}} \right]^2 dv dx_1$$

$$+ |b_3(\omega_{10}) - b_3(\omega_{11})|^2$$

For the first term,

$$\int_0^1 \int_{\tau^B M^{-\beta^*} - \psi_1(x_1)^B}^{\tau^B M^{-\beta^*}} \left[1 - \sqrt{1 + \left(\frac{v}{c_2} \right)^{1/\kappa}} \right]^2 dv dx_1$$

$$\leq \int_0^1 \int_{\tau^B M^{-\beta^*} - \psi_1(x_1)^B}^{\tau^B M^{-\beta^*}} \left(\frac{v}{c_2} \right)^{2/\kappa} dv dx_1$$

$$\leq \frac{\kappa c_2^{-2/\kappa}}{\kappa + 2} \int_0^1 \left(\tau^B M^{-\beta^*} \right)^{1+2/\kappa} - \left(\tau^B M^{-\beta^*} - \psi_1(x_1)^B \right)^{1+2/\kappa} dx_1$$

$$\leq \frac{\kappa c_2^{-2/\kappa}}{\kappa + 2} \left(\tau^B M^{-\beta^*} \right)^{1+2/\kappa} \int \left(1 - (1 - \phi(Mt)^B)^{1+2/\kappa} \right) dt$$

$$= \mathcal{O} \left(M^{-\beta^*(1+2/\kappa)-1} \right)$$

On the other hand,

$$\int_0^1 \int_{1/2}^{1/2+b_\omega(x_1)^B} \left[\frac{\frac{1}{2} + \tau^B M^{-\beta^*} - x_2}{c_2} \right]^{1/\kappa} dx_2 dx_1 = b_3(\omega) \left[\frac{1}{2} - b_\omega(x_1)^B \right]$$

yields

$$b_3(\omega_{11}) = \frac{1}{\frac{1}{2} - b_{\omega_{11}}(x_1)^B} \int_0^1 \int_{1/2}^{1/2+b_{\omega_{11}}(x_1)^B} \left[\frac{\frac{1}{2} + \tau^B M^{-\beta^*} - x_2}{c_2} \right]^{1/\kappa} dx_2 dx_1$$

$$\leq \frac{M c_2^{-1/\kappa}}{\frac{1}{2} - \tau^B M^{-\beta^*}} \int_0^1 \int_{\tau^B M^{-\beta^*}(1-\phi(Mx_1))}^{\tau^B M^{-\beta^*}} u^{1/\kappa} du dx_1$$

$$= \frac{M c_2^{-1/\kappa} \tau^B}{(\frac{1}{2} - \tau^B M^{-\beta^*})(1+1/\kappa)} M^{-\beta^*(1+1/\kappa)} \int (1 - (1 - \phi(Mt)^B)^{1+1/\kappa}) dt$$

$$\leq \frac{c_2^{-1/\kappa} \tau^B}{(\frac{1}{2} - \tau^B M^{-\beta^*})(1+1/\kappa)} M^{-\beta^*(1+1/\kappa)}$$

$$= \mathcal{O}(M^{-\beta^*(1+1/\kappa)})$$

Hence $|b_3(\omega_{11}) - b_3(\omega_{10})| = \mathcal{O}(M^{-\beta^*(1+1/\kappa)-1})$ and we have

$$H^2(P_{10}, P_{11}) = \mathcal{O}\left(M^{-\beta^*(1+2/\kappa)-1} \vee M^{-\beta^*(2+2/\kappa)-2} \right)$$

$$= \mathcal{O}\left(M^{-\beta^*(1+2/\kappa)-1} \right)$$

Now choose M as the smallest integer that is larger or equal to

$$n^{\frac{\kappa}{(2+\kappa)\beta^*+\kappa}}$$

Then we have $H^2(P_{10}, P_{11}) \leq C^* n^{-1} (1 + o(1))$ for some constant C^* depending only on κ, c_2, τ, ϕ and

$$\int \min\{dP_{j1}, dP_{j0}\} \geq \frac{1}{2} \left[1 - \frac{C^*}{2} n^{-1}(1 + o(1)) \right]^n \geq C_1^*$$

for n large enough and C_1^* is another constant. Thus for n large enough,

$$S \geq \frac{1}{2} C_1^* \tau^B M^{-\beta^*} \int \phi(t) dt \geq C_2^* n^{-\frac{\kappa \beta^*}{(2+\kappa)\beta^*+\kappa}}$$

The constant C_2^* only depends on κ, c_2, τ and ϕ.

Combining all the results so far yields that

$$\liminf_{n\to\infty} \inf_{\widetilde{G}_{m,n}} \sup_{(f,g)\in\mathcal{F}_{\mathcal{G}_h}} (n \wedge m)^{\frac{\beta^*\kappa}{\beta^*(\kappa+2)+d^*\kappa}} \mathbb{E}_{f,g}[d_\triangle(\widetilde{G}_{m,n}, G^*)] > 0$$

holds when $d^* = 1$. Using Lemma 21, we have

$$\liminf_{n\to\infty} \inf_{\widetilde{G}_{m,n}} \sup_{(f,g)\in\mathcal{F}_{\mathcal{G}_h}} (n \wedge m)^{\frac{\beta^*(\kappa+1)}{\beta^*(\kappa+2)+d^*\kappa}} \mathbb{E}_{f,g}[d_{f,g}(\widetilde{G}_{m,n}, G^*)] > 0$$

■

Note that under the composition smoothness assumption on the decision boundaries, existing upper bounds are suboptimal. The sub optimality comes from the κ term in the denominator of the rates and the $\log n$ term. Instead of κ as in the lower bound, we have $\kappa + 1$ in the upper bound.

7.7 Numerical experiment details

All experiments are conducted using PyTorch. The ReLU neural networks are coded with the `torch.nn` package, with the default initialization. Sample size is chosen to be 1000 for the experiments in Figure 5(a). The neural network training is done by stochastic gradient descent (initial learning rate=0.1, momentum=0.9, weight decay=0.001). The batch size is chosen to be 100. The total iteration number is 10000, with learning rate decayed by $1/10$ every 2000 steps. To make a fair comparison, we fix the random seed for the data generating process. The randomness comes from network initialization and batch selection. Test accuracy is evaluated by sampling one million test data points. When producing Figure 5(b), we choose $k = 10$ and $n = 200, 800, 3200, 12800$. Each point in the figure is an average from 10 replications.

References

Ethem Alpaydin and Cenk Kaynak. Cascading classifiers. *Kybernetika*, 34(4):369–374, 1998.

Raman Arora, Amitabh Basu, Poorya Mianjy, and Anirbit Mukherjee. Understanding deep neural networks with rectified linear units. *arXiv preprint arXiv:1611.01491*, 2016.

Jean-Yves Audibert, Alexandre B Tsybakov, et al. Fast learning rates for plug-in classifiers. *The Annals of statistics*, 35(2):608–633, 2007.

Peter L Bartlett, Michael I Jordan, and Jon D McAuliffe. Convexity, classification, and risk bounds. *Journal of the American Statistical Association*, 101(473):138–156, 2006.

Soumya Basu, Ankit Singh Rawat, and Manzil Zaheer. A statistical perspective on retrieval-based models. In *International Conference on Machine Learning*, pages 1852–1886. PMLR, 2023.

Benedikt Bauer, Michael Kohler, et al. On deep learning as a remedy for the curse of dimensionality in nonparametric regression. *The Annals of Statistics*, 47(4):2261–2285, 2019.

Thijs Bos and Johannes Schmidt-Hieber. Convergence rates of deep relu networks for multiclass classification. *arXiv preprint arXiv:2108.00969*, 2021.

Olivier Bousquet and André Elisseeff. Stability and generalization. *The Journal of Machine Learning Research*, 2:499–526, 2002.

Yuan Cao and Quanquan Gu. Generalization error bounds of gradient descent for learning overparameterized deep ReLU networks. *arXiv preprint arXiv:1902.01384*, 2019.

George Cybenko. Approximations by superpositions of a sigmoidal function. *Mathematics of Control, Signals and Systems*, 2:183–192, 1989.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

Gintare Karolina Dziugaite and Daniel M Roy. Computing nonvacuous generalization bounds for deep (stochastic) neural networks with many more parameters than training data. *arXiv preprint arXiv:1703.11008*, 2017.

Thomas Hamm and Ingo Steinwart. Adaptive learning rates for support vector machines working on data with low intrinsic dimension. *arXiv preprint arXiv:2003.06202*, 2020.

Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The elements of statistical learning: data mining, inference, and prediction*. Springer Science & Business Media, 2009.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.

Tianyang Hu, Zuofeng Shang, and Guang Cheng. Sharp rate of convergence for deep neural network classifiers under the teacher-student setting. *arXiv preprint arXiv:2001.06892*, 2020.

Tianyang Hu, Jun Wang, Wenjia Wang, and Zhenguo Li. Understanding square loss in training overparametrized neural network classifiers. *arXiv preprint arXiv:2112.03657*, 2021.

Masaaki Imaizumi and Kenji Fukumizu. Deep neural networks learn non-smooth functions effectively. *arXiv preprint arXiv:1802.04474*, 2018.

Yiding Jiang, Behnam Neyshabur, Hossein Mobahi, Dilip Krishnan, and Samy Bengio. Fantastic generalization measures and where to find them. *arXiv preprint arXiv:1912.02178*, 2019.

Yongdai Kim, Ilsang Ohn, and Dongha Kim. Fast convergence rates of deep neural networks for classification. *Neural Networks*, 138:179–197, 2021.

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

Michael Kohler and Sophie Langer. On the rate of convergence of fully connected deep neural network regression estimates. *The Annals of Statistics*, 49(4):2231–2249, 2021.

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

Michael Kohler, Adam Krzyżak, and Sophie Langer. Estimation of a function of low local dimensionality by deep neural networks. *IEEE Transactions on Information Theory*, 2022.

Aleksandr Petrovich Korostelev and Alexandre B Tsybakov. *Minimax theory of image reconstruction*, volume 82. Springer Science & Business Media, 2012.

Michael R Kosorok. *Introduction to empirical processes and semiparametric inference*. Springer Science & Business Media, 2007.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, pages 1097–1105, 2012.

Ruiqi Liu, Zuofeng Shang, and Guang Cheng. On deep instrumental variables estimate. *arXiv preprint arXiv:2004.14954*, 2020.

Ruiqi Liu, Ben Boukai, and Zuofeng Shang. Optimal nonparametric inference via deep neural network. *Journal of Mathematical Analysis and Applications*, 505:125561, 2022.

Jianfeng Lu, Zuowei Shen, Haizhao Yang, and Shijun Zhang. Deep network approximation for smooth functions. *arXiv preprint arXiv:2001.03040*, 2020.

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

Kien Nguyen, Clinton Fookes, Arun Ross, and Sridha Sridharan. Iris recognition with off-the-shelf cnn features: A deep learning perspective. *IEEE Access*, 6:18848–18855, 2017.

J. Ross Quinlan. Induction of decision trees. *Machine learning*, 1:81–106, 1986.

Johannes Schmidt-Hieber. Deep relu network approximation of functions on a manifold. *arXiv preprint arXiv:1908.00695*, 2019.

Johannes Schmidt-Hieber. Nonparametric regression using deep neural networks with relu activation function. *The Annals of Statistics*, 48(4):1875–1897, 2020.

Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. *arXiv preprint arXiv:1701.06538*, 2017.

Ingo Steinwart, Clint Scovel, et al. Fast rates for support vector machines using gaussian kernels. *The Annals of Statistics*, 35(2):575–607, 2007.

Taiji Suzuki. Adaptivity of deep relu network for learning in besov and mixed smooth besov spaces: optimal rate and curse of dimensionality. *arXiv preprint arXiv:1810.08033*, 2018.

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

Sara van de Geer. *Empirical Processes in M-estimation*. Cambridge University Press, 2000.

Vladimir N Vapnik. An overview of statistical learning theory. *IEEE transactions on neural networks*, 10(5):988–999, 1999.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Shuoyang Wang and Zuofeng Shang. Minimax optimal high-dimensional classification using deep neural networks. *Stat*, page e482, 2022.

Shuoyang Wang, Guanqun Cao, and Zuofeng Shang. Deep neural network classifier for multi-dimensional functional data. *arXiv preprint arXiv:2205.08592*, 2022a.

Shuoyang Wang, Zuofeng Shang, Guanqun Cao, and Jun S. Liu. Optimal classification for functional data. *arXiv preprint arXiv:2103.00569*, 2022b.

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

Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization. *arXiv preprint arXiv:1611.03530*, 2016.

Tong Zhang. Convergence of large margin separable linear classification. *Advances in neural information processing systems*, 13, 2000.