Journal of Machine Learning Research 25 (2024) 1-117 Submitted 1/22; Revised 7/23; Published 4/24

Classification with Deep Neural Networks and Logistic Loss

Zihan Zhang ZIHANZHANG19@FUDAN.EDU.CN
Shanghai Center for Mathematical Sciences
Fudan University, Shanghai 200433, China
School of Data Science
City University of Hong Kong, Kowloon, Hong Kong

Lei Shi[*] LEISHI@FUDAN.EDU.CN
School of Mathematical Sciences and Shanghai
Key Laboratory for Contemporary Applied Mathematics
Fudan University, Shanghai 200433, China
Shanghai Artificial Intelligence Laboratory
701 Yunjin Road, Shanghai 200232, China

Ding-Xuan Zhou DINGXUAN.ZHOU@SYDNEY.EDU.AU
School of Mathematics and Statistics
University of Sydney, Sydney NSW 2006, Australia

Editor: Maxim Raginsky

Abstract

Deep neural networks (DNNs) trained with the logistic loss (also known as the cross entropy loss) have made impressive advancements in various binary classification tasks. Despite the considerable success in practice, generalization analysis for binary classification with deep neural networks and the logistic loss remains scarce. The unboundedness of the target function for the logistic loss in binary classification is the main obstacle to deriving satisfactory generalization bounds. In this paper, we aim to fill this gap by developing a novel theoretical analysis and using it to establish tight generalization bounds for training fully connected ReLU DNNs with logistic loss in binary classification. Our generalization analysis is based on an elegant oracle-type inequality which enables us to deal with the boundedness restriction of the target function. Using this oracle-type inequality, we establish generalization bounds for fully connected ReLU DNN classifiers $\hat{f}_n^{\mathbf{FNN}}$ trained by empirical logistic risk minimization with respect to i.i.d. samples of size n, which lead to sharp rates of convergence as $n \to \infty$. In particular, we obtain optimal convergence rates for $\hat{f}_n^{\mathbf{FNN}}$ (up to some logarithmic factor) only requiring the Hölder smoothness of the conditional class probability η of data. Moreover, we consider a compositional assumption that requires η to be the composition of several vector-valued multivariate functions of which each component function is either a maximum value function or a Hölder smooth function only depending on a small number of its input variables. Under this assumption, we can even derive optimal convergence rates for $\hat{f}_n^{\mathbf{FNN}}$ (up to some logarithmic factor) which are independent of the input dimension of data. This result explains why in practice DNN classifiers can overcome the curse of dimensionality and perform well in high-dimensional classification problems. Furthermore, we establish dimension-free rates of convergence under other circumstances such as when the decision boundary is piecewise smooth and the input data are bounded away from it. Besides the novel oracle-type inequality, the sharp

∗. Corresponding author

convergence rates presented in our paper also owe to a tight error bound for approximating the natural logarithm function near zero (where it is unbounded) by ReLU DNNs. In addition, we justify our claims for the optimality of rates by proving corresponding minimax lower bounds. All these results are new in the literature and will deepen our theoretical understanding of classification with deep neural networks.

Keywords: deep learning; deep neural networks; binary classification; logistic loss; generalization analysis

1. Introduction

In this paper, we study the binary classification problem using deep neural networks (DNNs) with the rectified linear unit (ReLU) activation function. Deep learning based on DNNs has recently achieved remarkable success in a wide range of classification tasks including text categorization (Iyyer et al., 2015), image classification (Krizhevsky et al., 2012), and speech recognition (Hinton et al., 2012), which has become a cutting-edge learning method. ReLU is one of the most popular activation functions, as scalable computing and stochastic optimization techniques can facilitate the training of ReLU DNNs (Han et al., 2015; Kingma and Ba, 2015). Given a positive integer d, consider the binary classification problem where we regard $[0,1]^d$ as the input space and $\{-1,1\}$ as the output space representing the two labels of input data. Let P be a Borel probability measure on $[0,1]^d\times\{-1,1\}$, regarded as the data distribution (i.e., the joint distribution of the input and output data). The goal of classification is to learn a real-valued function from a *hypothesis space* $\mathcal{F}$ (i.e., a set of candidate functions) based on the sample of the distribution P. The predictive performance of any (deterministic) real-valued function f which has a Borel measurable restriction to $[0,1]^d$ (i.e., the domain of f contains $[0,1]^d$, and $[0,1]^d\ni x\mapsto f(x)\in\mathbb{R}$ is Borel measurable) is measured by the *misclassification error* of f with respect to P, given by

$$\mathcal{R}_P(f):=P\left(\left\{(x,y)\in[0,1]^d\times\{-1,1\}\middle| y\neq\operatorname{sgn}(f(x))\right\}\right), \tag{1.1}$$

or equivalently, the *excess misclassification error*

$$\mathcal{E}_P(f):=\mathcal{R}_P(f)-\inf\left\{\mathcal{R}_P(g)\middle| g:[0,1]^d\to\mathbb{R}\text{ is Borel measurable}\right\}. \tag{1.2}$$

Here $\operatorname{sgn}(\cdot)$ denotes the sign function which is defined as $\operatorname{sgn}(t)=1$ if $t\geq 0$ and $\operatorname{sgn}(t)=-1$ otherwise. The misclassification error $\mathcal{R}_P(f)$ characterizes the probability that the binary classifier $\operatorname{sgn}\circ f$ makes a wrong prediction, where $\circ$ means function composition, and by a *binary classifier* (or *classifier* for short) we mean a $\{-1,1\}$-valued function whose domain contains the input space $[0,1]^d$. Since any real-valued function f with its domain containing $[0,1]^d$ determines a classifier $\operatorname{sgn}\circ f$, we in this paper may call such a function f a *classifier* as well.

Note that the function we learn in a classification problem is based on the sample, meaning that it is not deterministic but a random function. Thus we take the expectation to measure its efficiency using the (excess) misclassification error. More specifically, let $\{(X_i,Y_i)\}_{i=1}^n$ be an independent and identically distributed (i.i.d.) sample of the distribution P and the hypothesis space $\mathcal{F}$ be a set of real-valued functions which have a Borel measurable restriction to $[0,1]^d$. We desire to construct an $\mathcal{F}$-valued statistic $\hat{f}_n$ from the sample

$\{(X_i, Y_i)\}_{i=1}^n$ and the classification performance of $\hat{f}_n$ can be characterized by upper bounds for the expectation of the excess misclassification error $\mathbb{E}\left[\mathcal{E}_P(\hat{f}_n)\right]$. One possible way to produce $\hat{f}_n$ is the *empirical risk minimization* with some *loss function* $\phi : \mathbb{R} \to [0, \infty)$, which is given by

$$\hat{f}_n \in \underset{f \in \mathcal{F}}{\arg\min} \frac{1}{n} \sum_{i=1}^n \phi\left(Y_i f(X_i)\right). \tag{1.3}$$

If $\hat{f}_n$ satisfies (1.3), then we will call $\hat{f}_n$ an *empirical ϕ-risk minimizer* (ERM with respect to ϕ, or ϕ-ERM) over $\mathcal{F}$. For any real-valued function f which has a Borel measurable restriction to $[0,1]^d$, the *ϕ-risk* and *excess ϕ-risk* of f with respect to P, denoted by $\mathcal{R}_P^\phi(f)$ and $\mathcal{E}_P^\phi(f)$ respectively, are defined as

$$\mathcal{R}_P^\phi(f) := \int_{[0,1]^d \times \{-1,1\}} \phi(yf(x)) \mathrm{d}P(x, y) \tag{1.4}$$

and

$$\mathcal{E}_P^\phi(f) := \mathcal{R}_P^\phi(f) - \inf\left\{ \mathcal{R}_P^\phi(g) \middle| g : [0,1]^d \to \mathbb{R} \text{ is Borel measurable} \right\}. \tag{1.5}$$

To derive upper bounds for $\mathbb{E}\left[\mathcal{E}_P(\hat{f}_n)\right]$, we can first establish upper bounds for $\mathbb{E}\left[\mathcal{E}_P^\phi(\hat{f}_n)\right]$, which are typically controlled by two parts, namely the sample error and the approximation error (e.g., cf. Chapter 2 of Cucker and Zhou (2007)). Then we are able to bound $\mathbb{E}\left[\mathcal{E}_P(\hat{f}_n)\right]$ by $\mathbb{E}\left[\mathcal{E}_P^\phi(\hat{f}_n)\right]$ through the so-called *calibration inequality* (also known as *Comparison Theorem*, see, e.g., Theorem 10.5 of Cucker and Zhou (2007) and Theorem 3.22 of Steinwart and Christmann (2008)). In this paper, we will call any upper bound for $\mathbb{E}\left[\mathcal{E}_P(\hat{f}_n)\right]$ or $\mathbb{E}\left[\mathcal{E}_P^\phi(\hat{f}_n)\right]$ a *generalization bound.*

Note that $\lim\limits_{n \to \infty} \frac{1}{n} \sum_{i=1}^n \phi\left(Y_i f(X_i)\right) = \mathcal{R}_P^\phi(f)$ almost surely for all measurable f. Therefore, the empirical ϕ-risk minimizer $\hat{f}_n$ defined in (1.3) can be regarded as an estimation of the so-called *target function* which minimizes the ϕ-risk $\mathcal{R}_P^\phi$ over all Borel measurable functions f. The target function can be defined pointwise. Rigorously, we say a measurable function $f^* : [0,1]^d \to [-\infty, \infty]$ is a target function of the ϕ-risk under the distribution P if for P_X-almost all $x \in [0,1]^d$ the value of f^* at x minimizes $\int_{\{-1,1\}} \phi(yz) \mathrm{d}P(y|x)$ over all $z \in [-\infty, \infty]$, i.e.,

$$f^*(x) \in \underset{z \in [-\infty, \infty]}{\arg\min} \int_{\{-1,1\}} \phi(yz) \mathrm{d}P(y|x) \text{ for } P_X\text{-almost all } x \in [0,1]^d, \tag{1.6}$$

where $\phi(yz) := \overline{\lim\limits_{t \to yz}}\, \phi(t)$ if $z \in \{-\infty, \infty\}$, P_X is the marginal distribution of P on $[0,1]^d$, and $P(\cdot|x)$ is the *regular conditional distribution* of P on $\{-1,1\}$ given x (cf. Lemma A.3.16 in Steinwart and Christmann (2008)). In this paper, we will use $f_{\phi,P}^*$ to denote the target function of the ϕ-risk under P. Note that $f_{\phi,P}^*$ may take values in $\{-\infty, \infty\}$, and $f_{\phi,P}^*$

minimizes $\mathcal{R}_P^\phi$ in the sense that

$$\begin{aligned}\mathcal{R}_P^\phi(f_{\phi,P}^*) &:= \int_{[0,1]^d\times\{-1,1\}} \phi(yf_{\phi,P}^*(x))\mathrm{d}P(x,y)\\ &= \inf\left\{\mathcal{R}_P^\phi(g)\middle|\, g:[0,1]^d\to\mathbb{R} \text{ is Borel measurable}\right\},\end{aligned} \tag{1.7}$$

where $\phi(yf_{\phi,P}^*(x)) := \overline{\lim}_{t\to yf_{\phi,P}^*(x)} \phi(t)$ if $yf_{\phi,P}^*(x)\in\{-\infty,\infty\}$ (cf. Lemma C.1).

In practice, the choice of the loss function ϕ varies, depending on the classification method used. For neural network classification, although other loss functions have been investigated, the *logistic loss* $\phi(t) = \log(1+\mathrm{e}^{-t})$, also known as the *cross entropy loss*, is most commonly used (see, e.g., Janocha and Czarnecki (2016); Hui and Belkin (2021); Hu et al. (2022b)). We now explain why the logistic loss is related to cross entropy. Let $\mathcal{X}$ be an arbitrary nonempty countable set equipped with the sigma algebra consisting of all its subsets. For any two probability measures Q_0 and Q on $\mathcal{X}$, the *cross entropy* of Q relative to Q_0 is defined as $\mathrm{H}(Q_0,Q) := -\sum_{z\in\mathcal{X}} Q_0(\{z\})\cdot\log Q(\{z\})$, where $\log 0 := -\infty$ and $0\cdot(-\infty) := 0$ (cf. (2.112) of Murphy (2012)). One can show that $\mathrm{H}(Q_0,Q)\geq \mathrm{H}(Q_0,Q_0)\geq 0$ and

$$\{Q_0\} = \arg\min_Q \mathrm{H}(Q_0,Q) \text{ if } \mathrm{H}(Q_0,Q_0)<\infty.$$

Therefore, roughly speaking, the cross entropy $\mathrm{H}(Q_0,Q)$ characterizes how close Q is to Q_0. For any $a\in[0,1]$, let $\mathscr{M}_a$ denote the probability measure on $\{-1,1\}$ with $\mathscr{M}_a(\{1\}) = a$ and $\mathscr{M}_a(\{-1\}) = 1-a$. Recall that any real-valued Borel measurable function f defined on the input space $[0,1]^d$ can induce a classifier $\operatorname{sgn}\circ f$. We can interpret the construction of the classifier $\operatorname{sgn}\circ f$ from f as follows. Consider the *logistic function*

$$\bar{l}:\mathbb{R}\to(0,1),\ z\mapsto\frac{1}{1+\mathrm{e}^{-z}}, \tag{1.8}$$

which is strictly increasing. For each $x\in[0,1]^d$, f induces a probability measure $\mathscr{M}_{\bar{l}(f(x))}$ on $\{-1,1\}$ via $\bar{l}$, which we regard as a prediction made by f of the distribution of the output data (i.e., the two labels $+1$ and -1) given the input data x. Observe that the larger $f(x)$ is, the closer the number $\bar{l}(f(x))$ gets to 1, and the more likely the event $\{1\}$ occurs under the distribution $\mathscr{M}_{\bar{l}(f(x))}$. If $\mathscr{M}_{\bar{l}(f(x))}(\{+1\})\geq\mathscr{M}_{\bar{l}(f(x))}(\{-1\})$, then $+1$ is more likely to appear given the input data x and we thereby think of f as classifying the input x as class $+1$. Otherwise, when $\mathscr{M}_{\bar{l}(f(x))}(\{+1\})<\mathscr{M}_{\bar{l}(f(x))}(\{-1\})$, x is classified as -1. In this way, f induces a classifier given by

$$x\mapsto\begin{cases}+1, & \text{if } \mathscr{M}_{\bar{l}(f(x))}(\{1\})\geq\mathscr{M}_{\bar{l}(f(x))}(\{-1\}),\\ -1, & \text{if } \mathscr{M}_{\bar{l}(f(x))}(\{1\})<\mathscr{M}_{\bar{l}(f(x))}(\{-1\}).\end{cases} \tag{1.9}$$

Indeed, the classifier in (1.9) is exactly $\operatorname{sgn}\circ f$. Thus we can also measure the predictive performance of f in terms of $\mathscr{M}_{\bar{l}(f(\cdot))}$ (instead of $\operatorname{sgn}\circ f$). To this end, one natural way is to compute the average "extent" of how close $\mathscr{M}_{\bar{l}(f(x))}$ is to the true conditional distribution of the output given the input x. If we use the cross entropy to characterize this "extent", then its average, which measures the classification performance of f, will be

$\int_{[0,1]^d} \mathrm{H}(\mathscr{Y}_x, \mathscr{M}_{\bar{l}(f(x))})\mathrm{d}\mathscr{X}(x)$, where $\mathscr{X}$ is the distribution of the input data, and $\mathscr{Y}_x$ is the conditional distribution of the output data given the input x. However, one can show that this quantity is just the logistic risk of f. Indeed,

$$\begin{aligned}
&\int_{[0,1]^d} \mathrm{H}(\mathscr{Y}_x, \mathscr{M}_{\bar{l}(f(x))})\mathrm{d}\mathscr{X}(x)\\
&= \int_{[0,1]^d} \Big(-\mathscr{Y}_x(\{1\}) \cdot \log(\mathscr{M}_{\bar{l}(f(x))}(\{1\})) - \mathscr{Y}_x(\{-1\}) \log(\mathscr{M}_{\bar{l}(f(x))}(\{-1\}))\Big)\,\mathrm{d}\mathscr{X}(x)\\
&= \int_{[0,1]^d} \left(-\mathscr{Y}_x(\{1\}) \cdot \log(\bar{l}(f(x))) - \mathscr{Y}_x(\{-1\}) \log(1-\bar{l}(f(x)))\right)\mathrm{d}\mathscr{X}(x)\\
&= \int_{[0,1]^d} \Big(\mathscr{Y}_x(\{1\}) \cdot \log(1+\mathrm{e}^{-f(x)}) + \mathscr{Y}_x(\{-1\}) \log(1+\mathrm{e}^{f(x)})\Big)\,\mathrm{d}\mathscr{X}(x)\\
&= \int_{[0,1]^d} \left(\mathscr{Y}_x(\{1\}) \cdot \phi(f(x)) + \mathscr{Y}_x(\{-1\})\phi(-f(x))\right)\mathrm{d}\mathscr{X}(x)\\
&= \int_{[0,1]^d}\int_{\{-1,1\}} \phi(yf(x))\mathrm{d}\mathscr{Y}_x(y)\mathrm{d}\mathscr{X}(x) = \int_{[0,1]^d\times\{-1,1\}} \phi(yf(x))\mathrm{d}P(x,y) = \mathcal{R}_P^\phi(f),
\end{aligned}$$

where ϕ is the logistic loss and P is the joint distribution of the input and output data, i.e., $\mathrm{d}P(x,y) = \mathrm{d}\mathscr{Y}_x(y)\mathrm{d}\mathscr{X}(x)$. Therefore, the average cross entropy of the distribution $\mathscr{M}_{\bar{l}(f(x))}$ induced by f to the true conditional distribution of the output data given the input data x is equal to the logistic risk of f with respect to the joint distribution of the input and output data, which explains why the logistic loss is also called the cross entropy loss. Compared with the misclassification error $\mathcal{R}_P(f)$ which measures the performance of the classifier $f(x)$ in correctly generating the class label $\mathrm{sgn}(f(x))$ that equals the most probable class label of the input data x (i.e., the label $y_x \in \{-1,+1\}$ such that $\mathscr{Y}_x(\{y_x\}) \geq \mathscr{Y}_x(\{-y_x\})$), the logistic risk $\mathcal{R}_P^\phi(f)$ measures how close the induced distribution $\mathscr{M}_{\bar{l}(f(x))}$ is to the true conditional distribution $\mathscr{Y}_x$. Consequently, in comparison with the (excess) misclassification error, the (excess) logistic risk is also a reasonable quantity for characterizing the performance of classifiers but from a different angle. When classifying with the logistic loss, we are essentially learning the conditional distribution $\mathscr{Y}_x$ through the cross entropy and the logistic function $\bar{l}$. Moreover, for any classifier $\hat{f}_n : [0,1]^d \to \mathbb{R}$ trained with logistic loss, the composite function $\bar{l} \circ \hat{f}_n(x) = \mathscr{M}_{\bar{l}\circ\hat{f}_n(x)}(\{1\})$ yields an estimation of the *conditional class probability function* $\eta(x) := P(\{1\}\,|x) = \mathscr{Y}_x(\{1\})$. Therefore, classifiers trained with logistic loss essentially capture more information about the exact value of the conditional class probability function $\eta(x)$ than we actually need to minimize the misclassification error $\mathcal{R}_P(\cdot)$, since the knowledge of the sign of $2\eta(x)-1$ is already sufficient for minimizing $\mathcal{R}_P(\cdot)$ (see (2.49)). In addition, we point out that the excess logistic risk $\mathcal{E}_P^\phi(f)$ is actually the average *Kullback-Leibler divergence* (*KL divergence*) from $\mathscr{M}_{\bar{l}(f(x))}$ to $\mathscr{Y}_x$. Here for any two probability measures Q_0 and Q on some countable set $\mathcal{X}$, the KL divergence from Q to Q_0 is defined as $\mathrm{KL}(Q_0||Q) := \sum_{z\in\mathcal{X}} Q_0(\{z\}) \cdot \log \frac{Q_0(\{z\})}{Q(\{z\})}$, where $Q_0(\{z\}) \cdot \log \frac{Q_0(\{z\})}{Q(\{z\})} := 0$ if $Q_0(\{z\}) = 0$ and $Q_0(\{z\}) \cdot \log \frac{Q_0(\{z\})}{Q(\{z\})} := \infty$ if $Q_0(\{z\}) > 0 = Q(\{z\})$ (cf. (2.111) of Murphy (2012) or Definition 2.5 of Tsybakov (2009)).

In this work, we focus on the generalization analysis of binary classification with empirical risk minimization over ReLU DNNs. That is, the classifiers under consideration are produced by algorithm (1.3) in which the hypothesis space $\mathcal{F}$ is generated by deep ReLU networks. Based on recent studies in complexity and approximation theory of DNNs (e.g., Bartlett et al. (2019); Petersen and Voigtlaender (2018); Yarotsky (2017)), several researchers have derived generalization bounds for ϕ-ERMs over DNNs in binary classification problems (Farrell et al., 2021; Kim et al., 2021; Shen et al., 2021). However, to the best of our knowledge, the existing literature fails to establish satisfactory generalization analysis if the target function $f^*_{\phi,P}$ is unbounded. In particular, take ϕ to be the logistic loss, i.e., $\phi(t) = \log(1 + \mathrm{e}^{-t})$. The target function is then explicitly given by $f^*_{\phi,P} \xlongequal{P_X\text{-a.s.}} \log \frac{\eta}{1-\eta}$ with $\eta(x) := P(\{1\}\,|x)$ ($x \in [0,1]^d$) being the conditional class probability function of P (cf. Lemma C.2), where recall that $P(\cdot|x)$ denotes the conditional probability of P on $\{-1,1\}$ given x. Hence $f^*_{\phi,P}$ is unbounded if η can be arbitrarily close to 0 or 1, which happens in many practical problems (see Section 3 for more details). For instance, we have $\eta(x) = 0$ or $\eta(x) = 1$ for a noise-free distribution P, implying $f^*_{\phi,P}(x) = \infty$ for P_X-almost all $x \in [0,1]^d$, where P_X is the marginal distribution of P on $[0,1]^d$. DNNs trained with the logistic loss perform efficiently in various image recognition applications as the smoothness of the loss function can further simplify the optimization procedure (Goodfellow et al., 2016; Krizhevsky et al., 2012; Simonyan and Zisserman, 2015). However, due to the unboundedness of $f^*_{\phi,P}$, the existing generalization analysis for classification with DNNs and the logistic loss either results in slow rates of convergence (e.g., the logarithmic rate in Shen et al. (2021)) or can only be conducted under very restrictive conditions (e.g., Kim et al. (2021); Farrell et al. (2021)) (cf. the discussions in Section 3). The unboundedness of the target function brings several technical difficulties to the generalization analysis. Indeed, if $f^*_{\phi,P}$ is unbounded, it cannot be approximated uniformly by continuous functions on $[0,1]^d$, which poses extra challenges for bounding the approximation error. Besides, previous sample error estimates based on concentration techniques are no longer valid because these estimates usually require involved random variables to be bounded or to satisfy strong tail conditions (cf. Chapter 2 of Wainwright (2019)). Therefore, in contrast to empirical studies, the previous strategies for generalization analysis could not demonstrate the efficiency of classification with DNNs and the logistic loss.

To fill this gap, in this paper we develop a novel theoretical analysis to establish tight generalization bounds for training DNNs with ReLU activation function and logistic loss in binary classification. Our main contributions are summarized as follows.

- For ϕ being the logistic loss, we establish an oracle-type inequality to bound the excess ϕ-risk without using the explicit form of the target function $f^*_{\phi,P}$. Through constructing a suitable bivariate function $\psi : [0,1]^d \times \{-1,1\} \to \mathbb{R}$, generalization analysis based on this oracle-type inequality can remove the boundedness restriction of the target function. Similar results hold even for the more general case when ϕ is merely Lipschitz continuous (see Theorem 2.1 and related discussions in Section 2.1).

- By using our oracle-type inequality, we establish tight generalization bounds for fully connected ReLU DNN classifiers $\hat{f}^{\mathbf{FNN}}_n$ trained by empirical logistic risk minimization (see (2.14)) and obtain sharp convergence rates in various settings:

    - ◦ We establish optimal convergence rates for the excess logistic risk of $\hat{f}^{\mathbf{FNN}}_n$ only requiring the Hölder smoothness of the conditional probability function η of the data distribution. Specifically, for Hölder-β smooth η, we show that the convergence rates of the excess logistic risk of $\hat{f}_n^{\mathbf{FNN}}$ can achieve $\mathcal{O}(\left(\frac{(\log n)^5}{n}\right)^{\frac{\beta}{\beta+d}})$, which is optimal up to the logarithmic term $(\log n)^{\frac{5\beta}{\beta+d}}$. From this we obtain the convergence rate $\mathcal{O}(\left(\frac{(\log n)^5}{n}\right)^{\frac{\beta}{2\beta+2d}})$ of the excess misclassification error of $\hat{f}_n^{\mathbf{FNN}}$, which is very close to the optimal rate, by using the calibration inequality (see Theorem 2.2). As a by-product, we also derive a new tight error bound for the approximation of the natural logarithm function (which is unbounded near zero) by ReLU DNNs (see Theorem 2.4). This bound plays a key role in establishing the aforementioned optimal rates of convergence.

- We consider a compositional assumption which requires the conditional probability function η to be the composition $h_q \circ h_{q-1} \circ \cdots \circ h_1 \circ h_0$ of several vector-valued multivariate functions h_i, satisfying that each component function of h_i is either a Hölder-β smooth function only depending on (a small number) d_* of its input variables or the maximum value function among some of its input variables. We show that under this compositional assumption the convergence rate of the excess logistic risk of $\hat{f}_n^{\mathbf{FNN}}$ can achieve $\mathcal{O}(\left(\frac{(\log n)^5}{n}\right)^{\frac{\beta\cdot(1\wedge\beta)^q}{d_*+\beta\cdot(1\wedge\beta)^q}})$, which is optimal up to the logarithmic term $(\log n)^{\frac{5\beta\cdot(1\wedge\beta)^q}{d_*+\beta\cdot(1\wedge\beta)^q}}$. We then use the calibration inequality to obtain the convergence rate $\mathcal{O}(\left(\frac{(\log n)^5}{n}\right)^{\frac{\beta\cdot(1\wedge\beta)^q}{2d_*+2\beta\cdot(1\wedge\beta)^q}})$ of the excess misclassification error of $\hat{f}_n^{\mathbf{FNN}}$ (see Theorem 2.3). Note that the derived convergence rates $\mathcal{O}(\left(\frac{(\log n)^5}{n}\right)^{\frac{\beta\cdot(1\wedge\beta)^q}{d_*+\beta\cdot(1\wedge\beta)^q}})$ and $\mathcal{O}(\left(\frac{(\log n)^5}{n}\right)^{\frac{\beta\cdot(1\wedge\beta)^q}{2d_*+2\beta\cdot(1\wedge\beta)^q}})$ are independent of the input dimension d, thereby circumventing the well-known curse of dimensionality. It can be shown that the above compositional assumption is likely to be satisfied in practice (see comments before Theorem 2.3). Thus this result helps to explain the huge success of DNNs in practical classification problems, especially high-dimensional ones.
- We derive convergence rates of the excess misclassification error of $\hat{f}_n^{\mathbf{FNN}}$ under the piecewise smooth decision boundary condition combining with the noise and margin conditions (see Theorem 2.5). As a special case of this result, we show that when the input data are bounded away from the decision boundary almost surely, the derived rates can also be dimension-free.

- We demonstrate the optimality of the convergence rates stated above by presenting corresponding minimax lower bounds (see Theorem 2.6 and Corollary 2.1).

The rest of this paper is organized as follows. In the remainder of this section, we first introduce some conventions and notations that will be used in this paper. Then we describe the mathematical modeling of fully connected ReLU neural networks which defines the hypothesis spaces in our setting. At the end of this section, we provide a symbol glossary for the convenience of readers. In Section 2, we present our main results in this paper, including the oracle-type inequality, several generalization bounds for classifiers obtained from empirical logistic risk minimization over fully connected ReLU DNNs, and two minimax lower bounds. Section 3 provides discussions and comparisons with related works and Section 4 concludes the paper. In Appendix A and Appendix B, we present covering number bounds and some approximation bounds for the space of fully connected ReLU DNNs respectively. Finally, in Appendix C, we give detailed proofs of results in the main body of this paper.

1.1 Conventions and Notations

Throughout this paper, we follow the conventions that $0^0 := 1$, $1^\infty := 1$, $\frac{z}{0} := \infty =: \infty^c$, $\log(\infty) := \infty$, $\log 0 := -\infty$, $0 \cdot w := 0 =: w \cdot 0$ and $\frac{a}{\infty} := 0 =: b^\infty$ for any $a \in \mathbb{R}, b \in [0,1), c \in (0,\infty)$, $z \in [0,\infty]$, $w \in [-\infty,\infty]$ where we denote by log the natural logarithm function (i.e. the base-e logarithm function). The terminology "measurable" means "Borel measurable" unless otherwise specified. Any Borel subset of some Euclidean space $\mathbb{R}^m$ is equipped with the Borel sigma algebra by default. Let $\mathcal{G}$ be an arbitrary measurable space and n be a positive integer. We call any sequence of $\mathcal{G}$-valued random variables $\{Z_i\}_{i=1}^n$ a *sample* in $\mathcal{G}$ of size n. Furthermore, for any measurable space $\mathcal{F}$ and any sample $\{Z_i\}_{i=1}^n$ in $\mathcal{G}$, an $\mathcal{F}$-valued statistic on $\mathcal{G}^n$ from the sample $\{Z_i\}_{i=1}^n$ is a random variable $\hat{\theta}$ together with a measurable map $\mathcal{T} : \mathcal{G}^n \to \mathcal{F}$ such that $\hat{\theta} = \mathcal{T}(Z_1, \ldots, Z_n)$, where $\mathcal{T}$ is called *the map associated with* the statistic $\hat{\theta}$. Let $\hat{\theta}$ be an arbitrary $\mathcal{F}$-valued statistic from some sample $\{Z_i\}_{i=1}^n$ and $\mathcal{T}$ is the map associated with $\hat{\theta}$. Then for any measurable space $\mathcal{D}$ and any measurable map $\mathcal{T}_0 : \mathcal{F} \to \mathcal{D}$, $\mathcal{T}_0(\hat{\theta}) = \mathcal{T}_0(\mathcal{T}(Z_1, \ldots, Z_n))$ is a $\mathcal{D}$-valued statistic from the sample $\{Z_i\}_{i=1}^n$, and $\mathcal{T}_0 \circ \mathcal{T}$ is the map associated with $\mathcal{T}_0(\hat{\theta})$.

Next we will introduce some notations used in this paper. We denote by $\mathbb{N}$ the set of all positive integers $\{1, 2, 3, 4, \ldots\}$. For $d \in \mathbb{N}$, we use $\mathcal{F}_d$ to denote the set of all Borel measurable functions from $[0,1]^d$ to $(-\infty,\infty)$, and use $\mathcal{H}_0^d$ to denote the set of all Borel probability measures on $[0,1]^d \times \{-1,1\}$. For any set A, the indicator function of A is given by

$$\mathbb{1}_A(x) := \begin{cases} 0, & \text{if } x \notin A, \\ 1, & \text{if } x \in A, \end{cases} \tag{1.10}$$

and the number of elements of A is denoted by $\#(A)$. For any finite dimensional vector v and any positive integer l less than or equal to the dimension of v, we denote by $(v)_l$ the l-th component of v. More generally, for any nonempty subset $I = \{i_1, i_2, \ldots, i_m\}$ of $\mathbb{N}$ with $1 \le i_1 < i_2 < \cdots < i_m \le$ the dimension of v, we denote $(v)_I := \big((v)_{i_1}, (v)_{i_2}, \ldots, (v)_{i_m}\big)$, which is a $\#(I)$-dimensional vector. For any function f, we use $\mathbf{dom}(f)$ to denote the domain of f, and use $\mathbf{ran}(f)$ to denote the range of f, that is, $\mathbf{ran}(f) := \big\{f(x)\big| x \in \mathbf{dom}(f)\big\}$. If f is a $[-\infty,\infty]^m$-valued function for some $m \in \mathbb{N}$ with $\mathbf{dom}(f)$ containing a nonempty set Ω, then the uniform norm of f on Ω is given by

$$\|f\|_\Omega := \sup\left\{ \left|\big(f(x)\big)_i\right| \,\middle|\, x \in \Omega,\, i \in \{1,2,\ldots,m\}\right\}. \tag{1.11}$$

For integer $m \ge 2$ and real numbers $a_1, \cdots, a_m$, define $a_1 \vee a_2 \vee \cdots \vee a_m = \max\{a_1, a_2, \cdots a_m\}$ and $a_1 \wedge a_2 \wedge \cdots \wedge a_m = \min\{a_1, a_2, \cdots a_m\}$. Given a real matrix $\boldsymbol{A} = (a_{i,j})_{i=1,\ldots,m, j=1,\ldots,l}$

and $t \in [0, \infty]$, the ℓ^t-norm of $\boldsymbol{A}$ is defined by

$$\|\boldsymbol{A}\|_t := \begin{cases} \sum_{i=1}^{m}\sum_{j=1}^{l} \mathbb{1}_{(0,\infty)}(|a_{i,j}|), & \text{if } t = 0, \\ \left|\sum_{i=1}^{m}\sum_{j=1}^{l} |a_{i,j}|^t\right|^{1/t}, & \text{if } 0 < t < \infty, \\ \sup\left\{|a_{i,j}|\,\middle|\, i \in \{1, \cdots, m\}, j \in \{1, \cdots, l\}\right\}, & \text{if } t = \infty. \end{cases} \tag{1.12}$$

Note that a vector is exactly a matrix with only one column or one row. Consequently, (1.12) with $l = 1$ or $m = 1$ actually defines the ℓ^t-norm of a real vector $\boldsymbol{A}$. Let $\mathcal{G}$ be a measurable space, $\{Z_i\}_{i=1}^n$ be a sample in $\mathcal{G}$ of size n, $\mathcal{P}_n$ be a probability measure on $\mathcal{G}^n$, and $\hat{\theta}$ be a $[-\infty, \infty]$-valued statistic on $\mathcal{G}^n$ from the sample $\{Z_i\}_{i=1}^n$. Then we denote

$$\boldsymbol{E}_{\mathcal{P}_n}[\hat{\theta}] := \int \mathcal{T} \mathrm{d}\mathcal{P}_n \tag{1.13}$$

provided that the integral $\int \mathcal{T} \mathrm{d}\mathcal{P}_n$ exists, where $\mathcal{T}$ is the map associated with $\hat{\theta}$. Therefore,

$$\boldsymbol{E}_{\mathcal{P}_n}[\hat{\theta}] = \mathbb{E}\left[\mathcal{T}(Z_1, \ldots, Z_n)\right] = \mathbb{E}[\hat{\theta}]$$

if the joint distribution of $(Z_1, \ldots, Z_n)$ is exactly $\mathcal{P}_n$. Let P be a Borel probability measure on $[0,1]^d \times \{-1, 1\}$ and $x \in [0,1]^d$. We use $P(\cdot|x)$ to denote the regular conditional distribution of P on $\{-1, 1\}$ given x, and P_X to denote the marginal distribution of P on $[0,1]^d$. For short, we will call the function $[0,1]^d \ni x \mapsto P(\{1\}|x) \in [0,1]$ the *conditional probability function* (instead of the *conditional class probability function*) of P. For any probability measure $\mathscr{Q}$ defined on some measurable space $(\Omega, \mathcal{F})$ and any $n \in \mathbb{N}$, we use $\mathscr{Q}^{\otimes n}$ to denote the product measure $\underbrace{\mathscr{Q} \times \mathscr{Q} \times \cdots \mathscr{Q}}_{n}$ defined on the product measurable space $(\underbrace{\Omega \times \Omega \times \cdots \Omega}_{n}, \underbrace{\mathcal{F} \otimes \mathcal{F} \otimes \cdots \mathcal{F}}_{n})$.

1.2 Spaces of Fully Connected Neural Networks

In this paper, we restrict ourselves to neural networks with the ReLU activation function. Consequently, hereinafter, for simplicity, we sometimes omit the word "ReLU" and the terminology "neural networks" will always refer to "ReLU neural networks".

The ReLU function is given by $\sigma : \mathbb{R} \to [0, \infty),\ t \mapsto \max\{t, 0\}$. For any vector $v \in \mathbb{R}^m$ with m being some positive integer, the v-shifted ReLU function is defined as $\sigma_v : \mathbb{R}^m \to [0, \infty)^m,\ x \mapsto \sigma(x - v)$, where the function σ is applied componentwise.

Neural networks considered in this paper can be expressed as a family of real-valued functions which take the form

$$f : \mathbb{R}^d \to \mathbb{R}, \quad x \mapsto \boldsymbol{W}_L \sigma_{v_L} \boldsymbol{W}_{L-1} \sigma_{v_{L-1}} \cdots \boldsymbol{W}_1 \sigma_{v_1} \boldsymbol{W}_0 x, \tag{1.14}$$

where the depth L denotes the number of hidden layers, m_k is the width of k-th layer, $\boldsymbol{W}_k$ is an $\mathrm{m}_{k+1} \times \mathrm{m}_k$ weight matrix with $\mathrm{m}_0 = d$ and $\mathrm{m}_{L+1} = 1$, and the shift vector $v_k \in \mathbb{R}^{\mathrm{m}_k}$

is called a bias. The architecture of a neural network is parameterized by weight matrices $\{\boldsymbol{W}_k\}_{k=0}^L$ and biases $\{v_k\}_{k=1}^L$, which will be estimated from data. Throughout the paper, whenever we talk about a neural network, we will explicitly associate it with a function f of the form (1.14) generated by $\{\boldsymbol{W}_k\}_{k=0}^L$ and $\{v_k\}_{k=1}^L$.

The space of fully connected neural networks is characterized by their depth and width, as well as the number of nonzero parameters in weight matrices and bias vectors. In addition, the complexity of this space is also determined by the $\|\cdot\|_\infty$-bounds of neural network parameters and $\|\cdot\|_{[0,1]^d}$-bounds of associated functions in form (1.14). Concretely, let $(G,N)\in[0,\infty)^2$ and $(S,B,F)\in[0,\infty]^3$, the space of fully connected neural networks is defined as

$$\mathcal{F}_d^{\mathbf{FNN}}(G,N,S,B,F):=\left\{f:\mathbb{R}^d\to\mathbb{R}\;\middle|\;\begin{array}{l}f\text{ is defined in (1.14) satisfying that}\\ L\leq G,\ \mathrm{m}_1\vee\mathrm{m}_2\vee\cdots\vee\mathrm{m}_L\leq N,\\ \left(\sum_{k=0}^L\|\boldsymbol{W}_k\|_0\right)+\left(\sum_{k=1}^L\|v_k\|_0\right)\leq S,\\ \sup\limits_{k=0,1,\cdots,L}\|\boldsymbol{W}_k\|_\infty\vee\sup\limits_{k=1,\cdots,L}\|v_k\|_\infty\leq B,\\ \text{and }\|f\|_{[0,1]^d}\leq F\end{array}\right\}. \quad (1.15)$$

In this definition, the freedom in choosing the position of nonzero entries of $\boldsymbol{W}_k$ reflects the fully connected nature between consecutive layers of the neural network f. It should be noticed that B and F in the definition (1.15) above can be ∞, meaning that there is no restriction on the upper bounds of $\|\boldsymbol{W}_k\|_\infty$ and $\|v_k\|_\infty$, or $\|f\|_{[0,1]^d}$. The parameter S in (1.15) can also be ∞, leading to a structure without sparsity. The space $\mathcal{F}_d^{\mathbf{FNN}}(G,N,S,B,F)$ incorporates all the essential features of fully connected neural network architectures and has been adopted to study the generalization properties of fully connected neural network models in regression and classification (Kim et al., 2021; Schmidt-Hieber, 2020).

1.3 Glossary

At the end of this section, we provide a glossary of frequently used symbols in this paper for the convenience of readers.

Symbol	**Meaning**	**Definition**
$\mathbb{Z}$	The set of integers.	
$\mathbb{N}$	The set of positive integers.	
$\mathbb{R}$	The set of real numbers.	
$\vee$	Taking the maximum, e.g., $a_1\vee a_2\vee a_3\vee a_4$ is equal to the maximum of $a_1,\ldots a_4$.	
$\wedge$	Taking the minimum, e.g., $a_1\wedge a_2\wedge a_3\wedge a_4$ is equal to the minimum of $a_1,\ldots a_4$.	
$\circ$	Function composition, e.g., for $f:\mathbb{R}\to\mathbb{R}$ and $g:\mathbb{R}\to\mathbb{R}$, $g\circ f$ denotes the map $\mathbb{R}\ni x\mapsto g(f(x))\in\mathbb{R}$.	
$\mathbf{dom}(f)$	The domain of a function f.	Below Eq. (1.10)

$\mathbf{ran}(f)$	The range of a function f.	Below Eq. (1.10)
$\boldsymbol{A}^\top$	The transpose of a matrix $\boldsymbol{A}$.	
$\#(A)$	The number of elements of a set A.	
$\lfloor\;\rfloor$	The floor function, which is defined as $\lfloor x \rfloor := \sup\{z \in \mathbb{Z} \mid z \leq x\}$.	
$\lceil\;\rceil$	The ceiling function, which is defined as $\lceil x \rceil := \inf\{z \in \mathbb{Z} \mid z \geq x\}$.	
$\mathbb{1}_A$	The indicator function of a set A.	Eq. (1.10)
$(v)_l$	The l-th component of a vector v.	Below Eq. (1.10)
$(v)_I$	The $\#(I)$-dimensional vector whose components are exactly $\{(v)_i\}_{i\in I}$.	Below Eq. (1.10)
$\|\cdot\|_\Omega$	The uniform norm on a set Ω.	Eq. (1.11)
$\|\cdot\|_t$	The ℓ^t-norm.	Eq. (1.12)
$\|\cdot\|_{\mathcal{C}^{k,\lambda}(\Omega)}$	The Hölder norm.	Eq. (2.12)
sgn	The sign function.	Below Eq. (1.2)
σ	The ReLU function, that is, $\mathbb{R} \ni t \mapsto \max\{0,t\} \in [0,\infty)$.	Above Eq. (1.14)
σ_v	The v-shifted ReLU function.	Above Eq. (1.14)
$\mathscr{M}_a$	The probability measure on $\{-1,1\}$ with $\mathscr{M}_a(\{1\}) = a$.	Above Eq. (1.8)
P_X	The marginal distribution of P on $[0,1]^d$.	Below Eq. (1.6)
$P(\cdot\|x)$	The regular conditional distribution of P on $\{-1,1\}$ given $x \in [0,1]^d$.	Below Eq. (1.6)
$P_{\eta,\mathscr{Q}}$	The probability on $[0,1]^d \times \{-1,1\}$ of which the marginal distribution on $[0,1]^d$ is $\mathscr{Q}$ and the conditional probability function is η.	Eq. (2.57)
P_η	The probability on $[0,1]^d \times \{-1,1\}$ of which the marginal distribution on $[0,1]^d$ is the Lebesgue measure and the conditional probability function is η.	Below Eq. (2.57)
$\boldsymbol{E}_{\mathcal{P}_n}[\hat{\theta}]$	The expectation of a statistic $\hat{\theta}$ when the joint distribution of the sample on which $\hat{\theta}$ depends is $\mathcal{P}_n$.	Eq. (1.13)
$\mathscr{Q}^{\otimes n}$	The product measure $\underbrace{\mathscr{Q} \times \mathscr{Q} \times \cdots \mathscr{Q}}_{n}$.	Below Eq. (1.13)
$\mathcal{R}_P(f)$	The misclassification error of f with respect to P.	Eq. (1.1)
$\mathcal{E}_P(f)$	The excess misclassification error of f with respect to P.	Eq. (1.2)
$\mathcal{R}_P^\phi(f)$	The ϕ-risk of f with respect to P.	Eq. (1.4)
$\mathcal{E}_P^\phi(f)$	The excess ϕ-risk of f with respect to P.	Eq. (1.5)
$f^*_{\phi,P}$	The target function of the ϕ-risk under some distribution P.	Eq. (1.6)
$\mathcal{N}(\mathcal{F},\gamma)$	The covering number of a class of real-valued functions $\mathcal{F}$ with radius γ in the uniform norm.	Eq. (2.1)

$\mathcal{B}_r^{\beta}(\Omega)$	The closed ball of radius r centered at the origin in the Hölder space of order β on Ω.	Eq. (2.13)
$\mathcal{G}_d^{\mathbf{M}}(d_\star)$	The set of all functions from $[0,1]^d$ to $\mathbb{R}$ which compute the maximum value of up to $d_\star$ components of their input vectors.	Eq. (2.27)
$\mathcal{G}_d^{\mathbf{H}}(d_*,\beta,r)$	The set of all functions in $\mathcal{B}_r^{\beta}([0,1]^d)$ whose output values depend on exactly d_* components of their input vectors.	Eq. (2.28)
$\mathcal{G}_\infty^{\mathbf{M}}(d_\star)$	$\mathcal{G}_\infty^{\mathbf{M}}(d_\star) := \bigcup_{d=1}^{\infty} \mathcal{G}_d^{\mathbf{M}}(d_\star)$	Above Eq. (2.30)
$\mathcal{G}_\infty^{\mathbf{H}}(d_*,\beta,r)$	$\mathcal{G}_\infty^{\mathbf{H}}(d_*,\beta,r) := \bigcup_{d=1}^{\infty} \mathcal{G}_d^{\mathbf{H}}(d_*,\beta,r)$	Above Eq. (2.30)
$\mathcal{G}_d^{\mathbf{CH}}(\cdots)$	$\mathcal{G}_d^{\mathbf{CH}}(q,K,d_*,\beta,r)$ consists of compositional functions $h_q \circ \cdots \circ h_0$ satisfying that each component function of h_i belongs to $\mathcal{G}_\infty^{\mathbf{H}}(d_*,\beta,r)$.	Eq. (2.31)
$\mathcal{G}_d^{\mathbf{CHOM}}(\cdots)$	$\mathcal{G}_d^{\mathbf{CHOM}}(q,K,d_\star,d_*,\beta,r)$ consists of compositional functions $h_q \circ \cdots \circ h_0$ satisfying that each component function of h_i belongs to $\mathcal{G}_\infty^{\mathbf{H}}(d_*,\beta,r) \cup \mathcal{G}_\infty^{\mathbf{M}}(d_\star)$.	Eq. (2.32)
$\mathcal{C}^{d,\beta,r,I,\Theta}$	The set of binary classifiers $\mathtt{C} : [0,1]^d \to \{-1,+1\}$ such that $\left\{x \in [0,1]^d \middle\vert\, \mathtt{C}(x) = +1\right\}$ is the union of some disjoint closed regions with piecewise Hölder smooth boundary.	Eq. (2.46)
$\Delta_{\mathtt{C}}(x)$	The distance from some point $x \in [0,1]^d$ to the decision boundary of some classifier $\mathtt{C} \in \mathcal{C}^{d,\beta,r,I,\Theta}$.	Eq. (2.48)
$\mathcal{F}_d$	The set of all Borel measurable functions from $[0,1]^d$ to $(-\infty,\infty)$.	Above Eq. (1.10)
$\mathcal{F}_d^{\mathbf{FNN}}(\cdots)$	The class of ReLU neural networks defined on $\mathbb{R}^d$.	Eq. (1.15)
$\mathcal{H}_0^d$	The set of all Borel probability measures on $[0,1]^d \times \{-1,1\}$.	Above Eq. (1.10)
$\mathcal{H}_1^{d,\beta,r}$	The set of all probability measures $P \in \mathcal{H}_0^d$ whose conditional probability function coincides with some function in $\mathcal{B}_r^{\beta}\left([0,1]^d\right)$ P_X-a.s..	Eq. (2.15)
$\mathcal{H}_{2,s_1,c_1,t_1}^{d,\beta,r}$	The set of all probability measures P in $\mathcal{H}_1^{d,\beta,r}$ satisfying the noise condition (2.24).	Eq. (2.26)
$\mathcal{H}_{3,A}^{d,\beta,r}$	The set of all probability measures $P \in \mathcal{H}_0^d$ whose marginal distribution on $[0,1]^d$ is the Lebesgue measure and whose conditional probability function is in $\mathcal{B}_r^{\beta}\left([0,1]^d\right)$ and bounded away from $\frac{1}{2}$ almost surely.	Eq. (2.58)
$\mathcal{H}_{4,q,K,d_\star,d_*}^{d,\beta,r}$	The set of all probability measures $P \in \mathcal{H}_0^d$ whose conditional probability function coincides with some function in $\mathcal{G}_d^{\mathbf{CHOM}}(q,K,d_\star,d_*,\beta,r)$ P_X-a.s..	Eq. (2.34)

$\mathcal{H}^{d,\beta,r}_{5,A,q,K,d_*}$	The set of all probability measures $P \in \mathcal{H}^d_0$ whose marginal distribution on $[0,1]^d$ is the Lebesgue measure and whose conditional probability function is in $\mathcal{G}^{\mathbf{CH}}_d(q,K,d_*,\beta,r)$ and bounded away from $\frac{1}{2}$ almost surely.	Eq. (2.58)
$\mathcal{H}^{d,\beta,r,I,\Theta,s_1,s_2}_{6,t_1,c_1,t_2,c_2}$	The set of all probability measures $P \in \mathcal{H}^d_0$ which satisfy the piecewise smooth decision boundary condition (2.50), the noise condition (2.24) and the margin condition (2.51) for some $\mathtt{C} \in \mathcal{C}^{d,\beta,r,I,\Theta}$.	Eq. (2.52)
$\mathcal{H}^{d,\beta}_7$	The set of all probability measures $P \in \mathcal{H}^d_0$ such that the target function of the logistic risk under P belongs to $\mathcal{B}^\beta_1\left([0,1]^d\right)$.	Above Eq. (3.4)
$\hat{f}^{\mathbf{FNN}}_n$	The DNN estimator obtained from empirical logistic risk minimization over the space of fully connected ReLU DNNs.	Eq. (2.14)

Table 1: Glossary of frequently used symbols in this paper

2. Main Results

In this section, we give our main results, consisting of upper bounds presented in Subsection 2.1 and lower bounds presented in Subsection 2.2.

2.1 Main Upper Bounds

In this subsection, we state our main results about upper bounds for the (excess) logistic risk or (excess) misclassification error of empirical logistic risk minimizers. The first result, given in Theorem 2.1, is an oracle-type inequality which provides upper bounds for the logistic risk of empirical logistic risk minimizers. Oracle-type inequalities have been extensively studied in the literature of nonparametric statistics (see Johnstone (1998) and references therein). As one of the main contributions in this paper, this inequality deserves special attention in its own right, allowing us to establish a novel strategy for generalization analysis. Before we state Theorem 2.1, we introduce some notations. For any pseudometric space $(\mathcal{F},\rho)$ (cf. Section 10.5 of Anthony and Bartlett (2009)) and $\gamma \in (0,\infty)$, the covering number of $(\mathcal{F},\rho)$ with radius γ is defined as

$$\mathcal{N}\left(\left(\mathcal{F},\rho\right),\gamma\right) := \inf\left\{\#\left(\mathcal{A}\right)\middle|\begin{array}{l}\mathcal{A}\subset\mathcal{F}\text{, and for any } f\in\mathcal{F} \text{ there}\\ \text{exists } g\in\mathcal{A} \text{ such that } \rho(f,g)\leq\gamma\end{array}\right\},$$

where we recall that $\#\left(\mathcal{A}\right)$ denotes the number of elements of the set $\mathcal{A}$. When the pseudometric ρ on $\mathcal{F}$ is clear and no confusion arises, we write $\mathcal{N}(\mathcal{F},\gamma)$ instead of $\mathcal{N}\left((\mathcal{F},\rho),\gamma\right)$ for simplicity. In particular, if $\mathcal{F}$ consists of real-valued functions which are bounded on $[0,1]^d$, we will use $\mathcal{N}(\mathcal{F},\gamma)$ to denote

$$\mathcal{N}\left(\left(\mathcal{F},\rho:(f,g)\mapsto\sup_{x\in[0,1]^d}\left|f(x)-g(x)\right|\right),\gamma\right) \tag{2.1}$$

unless otherwise specified. Recall that the ϕ-risk of a measurable function $f:[0,1]^d\to\mathbb{R}$ with respect to a distribution P on $[0,1]^d\times\{-1,1\}$ is denoted by $\mathcal{R}_P^\phi(f)$ and defined in (1.4).

Theorem 2.1 *Let $\{(X_i,Y_i)\}_{i=1}^n$ be an i.i.d. sample of a probability distribution P on $[0,1]^d\times\{-1,1\}$, $\mathcal{F}$ be a nonempty class of uniformly bounded real-valued functions defined on $[0,1]^d$, and $\hat{f}_n$ be an ERM with respect to the logistic loss $\phi(t)=\log(1+\mathrm{e}^{-t})$ over $\mathcal{F}$, i.e.,*

$$\hat{f}_n\in\mathop{\arg\min}_{f\in\mathcal{F}}\frac{1}{n}\sum_{i=1}^n\phi\left(Y_if(X_i)\right). \tag{2.2}$$

If there exists a measurable function $\psi:[0,1]^d\times\{-1,1\}\to\mathbb{R}$ and a constant triple $(M,\Gamma,\gamma)\in(0,\infty)^3$ such that

$$\int_{[0,1]^d\times\{-1,1\}}\psi(x,y)\mathrm{d}P(x,y)\leq\inf_{f\in\mathcal{F}}\int_{[0,1]^d\times\{-1,1\}}\phi(yf(x))\mathrm{d}P(x,y), \tag{2.3}$$

$$\sup\left\{\phi(t)\,\middle|\,|t|\leq\sup_{f\in\mathcal{F}}\|f\|_{[0,1]^d}\right\}\vee\sup\left\{|\psi(x,y)|\,\middle|\,(x,y)\in[0,1]^d\times\{-1,1\}\right\}\leq M, \tag{2.4}$$

$$\begin{aligned}&\int_{[0,1]^d\times\{-1,1\}}(\phi(yf(x))-\psi\left(x,y\right))^2\mathrm{d}P(x,y)\\&\leq\Gamma\cdot\int_{[0,1]^d\times\{-1,1\}}\left(\phi(yf(x))-\psi(x,y)\right)\mathrm{d}P(x,y),\ \forall\ f\in\mathcal{F},\end{aligned} \tag{2.5}$$

and

$$W:=\max\left\{3,\ \mathcal{N}\left(\mathcal{F},\gamma\right)\right\}<\infty.$$

Then for any $\varepsilon\in(0,1)$, there holds

$$\begin{aligned}&\mathbb{E}\left[\mathcal{R}_P^\phi\left(\hat{f}_n\right)-\int_{[0,1]^d\times\{-1,1\}}\psi(x,y)\mathrm{d}P(x,y)\right]\\&\leq80\cdot\frac{(1+\varepsilon)^2}{\varepsilon}\cdot\frac{\Gamma\log W}{n}+(20+20\varepsilon)\cdot\frac{M\log W}{n}+(20+20\varepsilon)\cdot\sqrt{\gamma}\cdot\sqrt{\frac{\Gamma\log W}{n}}\\&\quad+4\gamma+(1+\varepsilon)\cdot\inf_{f\in\mathcal{F}}\left(\mathcal{R}_P^\phi(f)-\int_{[0,1]^d\times\{-1,1\}}\psi(x,y)\mathrm{d}P(x,y)\right).\end{aligned} \tag{2.6}$$

According to its proof in Appendix C.2, Theorem 2.1 remains true when the logistic loss is replaced by any nonnegative function ϕ satisfying

$$\left|\phi(t)-\phi(t')\right|\leq\left|t-t'\right|,\ \forall\ t,t'\in\left[-\sup_{f\in\mathcal{F}}\|f\|_{[0,1]^d},\sup_{f\in\mathcal{F}}\|f\|_{[0,1]^d}\right].$$

Then by rescaling, Theorem 2.1 can be further generalized to the case when ϕ is any nonnegative locally Lipschitz continuous loss function such as the exponential loss or the

LUM (large-margin unified machine) loss (cf. Liu et al. (2011)). Generalization analysis for classification with these loss functions based on oracle-type inequalities similar to Theorem 2.1 has been studied in our coming work Zhang et al. (2024).

Let us give some comments on conditions (2.3) and (2.5) of Theorem 2.1. To our knowledge, these two conditions are introduced for the first time in this paper, and will play pivotal roles in our estimates. Let ϕ be the logistic loss and P be a probability measure on $[0,1]^d \times \{-1,1\}$. Recall that $f^*_{\phi,P}$ denotes the target function of the logistic risk. If

$$\int_{[0,1]^d\times\{-1,1\}} \psi(x,y)\mathrm{d}P(x,y) = \inf\left\{ \mathcal{R}^{\phi}_P(f) \,\middle|\, f:[0,1]^d\to\mathbb{R} \text{ is measurable}\right\}, \tag{2.7}$$

then condition (2.3) is satisfied and the left hand side of (2.6) is exactly $\mathbb{E}\left[\mathcal{E}^{\phi}_P\left(\hat{f}_n\right)\right]$. Therefore, Theorem 2.1 can be used to establish excess ϕ-risk bounds for the ϕ-ERM $\hat{f}_n$. In particular, one can take $\psi(x,y)$ to be $\phi(yf^*_{\phi,P}(x))$ to ensure the equality (2.7) (recalling (1.7)). It should be pointed out that if $\psi(x,y)=\phi(yf^*_{\phi,P}(x))$, inequality (2.5) is of the same form as the following inequality with $\tau=1$, which asserts that there exist $\tau\in[0,1]$ and $\Gamma>0$ such that

$$\int_{[0,1]^d\times\{-1,1\}} \Big(\phi(yf(x)) - \phi\left(yf^*_{\phi}(x)\right)\Big)^2 \mathrm{d}P(x,y) \le \Gamma\cdot\Big(\mathcal{E}^{\phi}_P(f)\Big)^{\tau},\ \forall\, f\in\mathcal{F}. \tag{2.8}$$

This inequality appears naturally when bounding the sample error by using concentration inequalities, which is of great importance in previous generalization analysis for binary classification (cf. condition (A4) in Kim et al. (2021) and Definition 10.15 in Cucker and Zhou (2007)). In Farrell et al. (2021), the authors actually prove that if the target function $f^*_{\phi,P}$ is bounded and the functions in $\mathcal{F}$ are uniformly bounded by some $F>0$, the inequality (2.5) holds with $\psi(x,y)=\phi(yf^*_{\phi,P}(x))$ and

$$\Gamma = \frac{2}{\inf\left\{\phi''(t)\,\middle|\, t\in\mathbb{R},\ |t|\le\max\left\{F, \left\|f^*_{\phi,P}\right\|_{[0,1]^d}\right\}\right\}}.$$

Here $\phi''(t)$ denotes the second order derivative of $\phi(t)=\log(1+\mathrm{e}^{-t})$ which is given by $\phi''(t)=\frac{\mathrm{e}^t}{(1+\mathrm{e}^t)^2}$. The boundedness of $f^*_{\phi,P}$ is a key ingredient leading to the main results in Farrell et al. (2021) (see Section 3 for more details). However, $f^*_{\phi,P}$ is explicitly given by $\log\frac{\eta}{1-\eta}$ with $\eta(x)=P(\{1\}\,|x)$, which tends to infinity when η approaches to 0 or 1. In some cases, the uniformly boundedness assumption on $f^*_{\phi,P}$ is too restrictive. When $f^*_{\phi,P}$ is unbounded, i.e., $\|f^*_{\phi,P}\|_{[0,1]^d}=\infty$, condition (2.5) will not be satisfied by simply taking $\psi(x,y)=\phi(yf^*_{\phi,P}(x))$. Since in this case we have $\inf_{t\in(-\infty,+\infty)}\phi''(t)=0$, one cannot find a finite constant Γ to guarantee the validity of (2.5), i.e., the inequality (2.8) cannot hold for $\tau=1$, which means the previous strategy for generalization analysis in Farrell et al. (2021) fails to work. In Theorem 2.1, the requirement for $\psi(x,y)$ is much more flexible, we don't require $\psi(x,y)$ to be $\phi(yf^*_{\phi,P}(x))$ or even to satisfy (2.7). In this paper, by recurring to Theorem 2.1, we carefully construct ψ to avoid using $f^*_{\phi,P}$ directly in the following estimates. Based on this strategy, under some mild regularity conditions on η, we can develop a more general analysis to demonstrate the performance of neural network classifiers trained with the logistic loss regardless of the unboundedness of $f^*_{\phi,P}$. The derived generalization bounds and rates of convergence are stated in Theorem 2.2, Theorem 2.3, and Theorem 2.5, which are new in the literature and constitute the main contributions of this paper. It is worth noticing that in Theorem 2.2 and Theorem 2.3, we use Theorem 2.1 to obtain optimal rates of convergence (up to some logarithmic factor), which demonstrates the tightness and power of the inequality (2.6) in Theorem 2.1. To obtain these optimal rates from Theorem 2.1, a delicate construction of ψ which allows small constants M and Γ in (2.4) and (2.5) is necessary. One frequently used form of ψ in this paper is

$$\begin{aligned}\psi :&[0,1]^d \times \{-1,1\} \to \mathbb{R},\\ &(x,y) \mapsto \begin{cases} \phi\left(y \log \dfrac{\eta(x)}{1-\eta(x)}\right), & \eta(x) \in [\delta_1, 1-\delta_1],\\ 0, & \eta(x) \in \{0,1\},\\ \eta(x)\log\dfrac{1}{\eta(x)} + (1-\eta(x))\log\dfrac{1}{1-\eta(x)}, & \eta(x) \in (0,\delta_1)\cup(1-\delta_1,1),\end{cases}\end{aligned} \tag{2.9}$$

which can be regarded as a truncated version of $\phi(yf^*_{\phi,P}(x)) = \phi\left(y\log\frac{\eta(x)}{1-\eta(x)}\right)$, where δ_1 is some suitable constant in $(0,1/2]$. However, in Theorem 2.5 we use a different form of ψ, which will be specified later.

The proof of Theorem 2.1 is based on the following error decomposition

$$\mathbb{E}\left[\mathcal{R}^\phi_P\left(\hat{f}_n\right) - \Psi\right] \leq \mathrm{T}_{\varepsilon,\psi,n} + (1+\varepsilon)\cdot \inf_{g\in\mathcal{F}}\left(\mathcal{R}^\phi_P(g) - \Psi\right),\ \forall\, \varepsilon\in[0,1), \tag{2.10}$$

where $\mathrm{T}_{\varepsilon,\psi,n} := \mathbb{E}\left[\mathcal{R}^\phi_P\left(\hat{f}_n\right) - \Psi - (1+\varepsilon)\cdot\frac{1}{n}\sum_{i=1}^n\left(\phi\left(Y_i\hat{f}_n(X_i)\right) - \psi(X_i,Y_i)\right)\right]$ and $\Psi = \int_{[0,1]^d\times\{-1,1\}}\psi(x,y)\mathrm{d}P(x,y)$ (see (C.13)). Although (2.10) is true for $\varepsilon = 0$, it's better to take $\varepsilon > 0$ in (2.10) to obtain sharp rates of convergence. This is because bounding the term $\mathrm{T}_{\varepsilon,\psi,n}$ with $\varepsilon\in(0,1)$ is easier than bounding $\mathrm{T}_{0,\psi,n}$. To see this, note that for $\varepsilon\in(0,1)$ we have

$$\mathrm{T}_{\varepsilon,\psi,n} = (1+\varepsilon)\cdot\mathrm{T}_{0,\psi,n} - \varepsilon\cdot\mathbb{E}\left[\mathcal{R}^\phi_P\left(\hat{f}_n\right) - \Psi\right] \leq (1+\varepsilon)\cdot\mathrm{T}_{0,\psi,n},$$

meaning that we can always establish tighter upper bounds for $\mathrm{T}_{\varepsilon,\psi,n}$ than for $\mathrm{T}_{0,\psi,n}$ (up to the constant factor $1+\varepsilon<2$). Indeed, $\varepsilon>0$ is necessary in establishing Theorem 2.1, as indicated in its proof in Appendix C.2. We also point out that, setting $\varepsilon=0$ and $\psi\equiv 0$ (hence $\Psi=0$) in (2.10), and subtracting $\inf\left\{\mathcal{R}^\phi_P(g)\,\middle|\, g:[0,1]^d\to\mathbb{R} \text{ measurable}\right\}$ from both sides, we will obtain a simpler error decomposition

$$\begin{aligned}\mathbb{E}\left[\mathcal{E}^\phi_P\left(\hat{f}_n\right)\right] &\leq \mathbb{E}\left[\mathcal{R}^\phi_P\left(\hat{f}_n\right) - \frac{1}{n}\sum_{i=1}^n\left(\phi\left(Y_i\hat{f}_n(X_i)\right)\right)\right] + \inf_{g\in\mathcal{F}}\mathcal{E}^\phi_P(g)\\ &\leq \mathbb{E}\left[\sup_{g\in\mathcal{F}}\left|\mathcal{R}^\phi_P(g) - \frac{1}{n}\sum_{i=1}^n\left(\phi\left(Y_ig(X_i)\right)\right)\right|\right] + \inf_{g\in\mathcal{F}}\mathcal{E}^\phi_P(g),\end{aligned} \tag{2.11}$$

which is frequently used in the literature (see e.g., Lemma 2 in Kohler and Langer (2020) and the proof of Proposition 4.1 in Mohri et al. (2018)). Note that (2.11) does not require the explicit form of $f^*_{\phi,P}$, which means that we can also use this error decomposition to establish rates of convergence for $\mathbb{E}\left[\mathcal{E}^{\phi}_P(\hat{f}_n)\right]$ regardless of the unboundedness of $f^*_{\phi,P}$. However, in comparison with Theorem 2.1, using (2.11) may result in slow rates of convergence because of the absence of the positive parameter ε and a carefully constructed function ψ.

We now state Theorem 2.2 which establishes generalization bounds for empirical logistic risk minimizers over DNNs. In order to present this result, we need the definition of Hölder spaces (Evans, 2010). The Hölder space $\mathcal{C}^{k,\lambda}(\Omega)$, where $\Omega\subset\mathbb{R}^d$ is a closed domain, $k\in\mathbb{N}\cup\{0\}$ and $\lambda\in(0,1]$, consists of all those functions from Ω to $\mathbb{R}$ which have continuous derivatives up to order k and whose k-th partial derivatives are Hölder-λ continuous on Ω. Here we say a function $g:\Omega\to\mathbb{R}$ is Hölder-λ continuous on Ω, if

$$|g|_{\mathcal{C}^{0,\lambda}(\Omega)}:=\sup_{\Omega\ni x\neq z\in\Omega}\frac{|g(x)-g(z)|}{\|x-z\|_2^{\lambda}}<\infty.$$

Then the Hölder spaces $\mathcal{C}^{k,\lambda}(\Omega)$ can be assigned the norm

$$\|f\|_{\mathcal{C}^{k,\lambda}(\Omega)}:=\max_{\|\boldsymbol{m}\|_1\leq k}\|\mathrm{D}^{\boldsymbol{m}}f\|_{\Omega}+\max_{\|\boldsymbol{m}\|_1=k}|\mathrm{D}^{\boldsymbol{m}}f|_{\mathcal{C}^{0,\lambda}(\Omega)}, \tag{2.12}$$

where $\boldsymbol{m}=(m_1,\cdots,m_d)\in(\mathbb{N}\cup\{0\})^d$ ranges over multi-indices (hence $\|\boldsymbol{m}\|_1=\sum_{i=1}^d m_i$) and $\mathrm{D}^{\boldsymbol{m}}f(x_1,\ldots,x_d)=\frac{\partial^{m_1}}{\partial x_1^{m_1}}\cdots\frac{\partial^{m_d}}{\partial x_d^{m_d}}f(x_1,\ldots,x_d)$. Given $\beta\in(0,\infty)$, we say a function $f:\Omega\to\mathbb{R}$ is Hölder-β smooth if $f\in\mathcal{C}^{k,\lambda}(\Omega)$ with $k=\lceil\beta\rceil-1$ and $\lambda=\beta-\lceil\beta\rceil+1$, where $\lceil\beta\rceil$ denotes the smallest integer larger than or equal to β. For any $\beta\in(0,\infty)$ and any $r\in(0,\infty)$, let

$$\mathcal{B}^{\beta}_r(\Omega):=\left\{f:\Omega\to\mathbb{R}\,\middle|\,\begin{matrix}f\in\mathcal{C}^{k,\lambda}(\Omega)\text{ and }\|f\|_{\mathcal{C}^{k,\lambda}(\Omega)}\leq r\text{ for}\\ k=-1+\lceil\beta\rceil\text{ and }\lambda=\beta-\lceil\beta\rceil+1\end{matrix}\right\} \tag{2.13}$$

denote the closed ball of radius r centered at the origin in the Hölder space of order β on Ω. Recall that the space $\mathcal{F}^{\mathbf{FNN}}_d(G,N,S,B,F)$ generated by fully connected neural networks is given in (1.15), which is parameterized by the depth and width of neural networks (bounded by G and N), the number of nonzero entries in weight matrices and bias vectors (bounded by S), and the upper bounds of neural network parameters and associated functions of form (1.14) (denoted by B and F). In the following theorem, we show that to ensure the rate of convergence as the sample size n becomes large, all these parameters should be taken within certain ranges scaling with n. For two positive sequences $\{\lambda_n\}_{n\geq1}$ and $\{\nu_n\}_{n\geq1}$, we say $\lambda_n\lesssim\nu_n$ holds if there exist $n_0\in\mathbb{N}$ and a positive constant c independent of n such that $\lambda_n\leq c\nu_n,\forall\, n\geq n_0$. In addition, we write $\lambda_n\asymp\nu_n$ if and only if $\lambda_n\lesssim\nu_n$ and $\nu_n\lesssim\lambda_n$. Recall that the excess misclassification error of $f:\mathbb{R}^d\to\mathbb{R}$ with respect to some distribution P on $[0,1]^d\times\{-1,1\}$ is defined as

$$\mathcal{E}_P(f)=\mathcal{R}_P(f)-\inf\left\{\mathcal{R}_P(g)\,\middle|\,g:[0,1]^d\to\mathbb{R}\text{ is Borel measurable}\right\},$$

where $\mathcal{R}_P(f)$ denotes the misclassification error of f given by

$$\mathcal{R}_P(f)=P\left(\left\{(x,y)\in[0,1]^d\times\{-1,1\}\,\middle|\,y\neq\operatorname{sgn}(f(x))\right\}\right).$$

Theorem 2.2 *Let $d \in \mathbb{N}$, $(\beta, r) \in (0,\infty)^2$, $n \in \mathbb{N}$, $\nu \in [0,\infty)$, $\{(X_i, Y_i)\}_{i=1}^n$ be an i.i.d. sample in $[0,1]^d \times \{-1,1\}$ and $\hat{f}_n^{\mathbf{FNN}}$ be an ERM with respect to the logistic loss $\phi(t) = \log\left(1+\mathrm{e}^{-t}\right)$ over $\mathcal{F}_d^{\mathbf{FNN}}(G,N,S,B,F)$, i.e.,*

$$\hat{f}_n^{\mathbf{FNN}} \in \underset{f\in\mathcal{F}_d^{\mathbf{FNN}}(G,N,S,B,F)}{\arg\min} \frac{1}{n}\sum_{i=1}^n \phi\left(Y_i f(X_i)\right). \tag{2.14}$$

Define

$$\mathcal{H}_1^{d,\beta,r} := \left\{ P \in \mathcal{H}_0^d \middle| \begin{matrix} P_X(\left\{z\in[0,1]^d \middle| P(\{1\}|z) = \hat{\eta}(z)\right\}) = 1 \\ \text{for some } \hat{\eta} \in \mathcal{B}_r^\beta\left([0,1]^d\right) \end{matrix} \right\}. \tag{2.15}$$

Then there exists a constant $\mathrm{c} \in (0,\infty)$ only depending on (d,β,r), such that the estimator $\hat{f}_n^{\mathbf{FNN}}$ defined by (2.14) with

$$\begin{aligned} &\mathrm{c}\log n \le G \lesssim \log n,\ N \asymp \left(\frac{(\log n)^5}{n}\right)^{\frac{-d}{d+\beta}},\ S \asymp \left(\frac{(\log n)^5}{n}\right)^{\frac{-d}{d+\beta}} \cdot \log n, \\ &1 \le B \lesssim n^\nu,\ \text{and}\ \frac{\beta}{d+\beta}\cdot\log n \le F \lesssim \log n \end{aligned} \tag{2.16}$$

satisfies

$$\sup_{P\in\mathcal{H}_1^{d,\beta,r}} \boldsymbol{E}_{P^{\otimes n}}\left[\mathcal{E}_P^\phi\left(\hat{f}_n^{\mathbf{FNN}}\right)\right] \lesssim \left(\frac{(\log n)^5}{n}\right)^{\frac{\beta}{\beta+d}} \tag{2.17}$$

and

$$\sup_{P\in\mathcal{H}_1^{d,\beta,r}} \boldsymbol{E}_{P^{\otimes n}}\left[\mathcal{E}_P\left(\hat{f}_n^{\mathbf{FNN}}\right)\right] \lesssim \left(\frac{(\log n)^5}{n}\right)^{\frac{\beta}{2\beta+2d}}. \tag{2.18}$$

Theorem 2.2 will be proved in Appendix C.4. As far as we know, for classification with neural networks and the logistic loss ϕ, generalization bounds presented in (2.17) and (2.18) establish fastest rates of convergence among the existing literature under the Hölder smoothness condition on the conditional probability function η of the data distribution P. Note that to obtain such generalization bounds in (2.17) and (2.18) we do not require any assumption on the marginal distribution P_X of the distribution P. For example, we dot not require that P_X is absolutely continuous with respect to the Lebesgue measure. The rate $\mathcal{O}(\left(\frac{(\log n)^5}{n}\right)^{\frac{\beta}{\beta+d}})$ in (2.17) for the convergence of excess ϕ-risk is indeed optimal (up to some logarithmic factor) in the minimax sense (see Corollary 2.1 and comments therein). However, the rate $\mathcal{O}(\left(\frac{(\log n)^5}{n}\right)^{\frac{\beta}{2\beta+2d}})$ in (2.18) for the convergence of excess misclassification error is not optimal. According to Theorem 4.1, Theorem 4.2, Theorem 4.3 and their proofs in Audibert and Tsybakov (2007), there holds

$$\inf_{\hat{f}_n} \sup_{P\in\mathcal{H}_1^{d,\beta,r}} \boldsymbol{E}_{P^{\otimes n}}\left[\mathcal{E}_P(\hat{f}_n)\right] \asymp n^{-\frac{\beta}{2\beta+d}}, \tag{2.19}$$

where the infimum is taken over all $\mathcal{F}_d$-valued statistics from the sample $\{(X_i, Y_i)\}_{i=1}^n$. Therefore, the rate $\mathcal{O}(\left(\frac{(\log n)^5}{n}\right)^{\frac{\beta}{2\beta+2d}})$ in (2.18) does not match the minimax optimal rate

$\mathcal{O}((\frac{1}{n})^{\frac{\beta}{2\beta+d}})$. Despite suboptimality, the rate $\mathcal{O}(\left(\frac{(\log n)^5}{n}\right)^{\frac{\beta}{2\beta+2d}})$ in (2.18) is fairly close to the optimal rate $\mathcal{O}((\frac{1}{n})^{\frac{\beta}{2\beta+d}})$, especially when $\beta >> d$ because the exponents satisfy

$$\lim_{\beta\to+\infty}\frac{\beta}{2\beta+2d}=\frac{1}{2}=\lim_{\beta\to+\infty}\frac{\beta}{2\beta+d}.$$

In our proof of Theorem 2.2, the rate $\mathcal{O}(\left(\frac{(\log n)^5}{n}\right)^{\frac{\beta}{2\beta+2d}})$ in (2.18) is derived directly from the rate $\left(\frac{(\log n)^5}{n}\right)^{\frac{\beta}{\beta+d}}$ in (2.17) via the so-called calibration inequality which takes the form

$$\mathcal{E}_P(f)\leq c\cdot\left(\mathcal{E}_P^{\phi}(f)\right)^{\frac{1}{2}} \text{ for any } f\in\mathcal{F}_d \text{ and any } P\in\mathcal{H}_0^d \tag{2.20}$$

with c being a constant independent of P and f (see (C.98)). Indeed, it follows from Theorem 8.29 of Steinwart and Christmann (2008) that

$$\mathcal{E}_P\left(f\right)\leq 2\sqrt{2}\cdot\left(\mathcal{E}_P^{\phi}\left(f\right)\right)^{\frac{1}{2}} \text{ for any } f\in\mathcal{F}_d \text{ and any } P\in\mathcal{H}_0^d. \tag{2.21}$$

In other words, (2.20) holds when $c=2\sqrt{2}$. Interestingly, we can use Theorem 2.2 to obtain that the inequality (2.20) is optimal in the sense that the exponent $\frac{1}{2}$ cannot be replaced by a larger one. Specifically, by using (2.17) of our Theorem 2.2 together with (2.19), we can prove that $\frac{1}{2}$ is the largest number s such that there holds

$$\mathcal{E}_P(f)\leq c\cdot\left(\mathcal{E}_P^{\phi}(f)\right)^{s} \text{ for any } f\in\mathcal{F}_d \text{ and any } P\in\mathcal{H}_0^d \tag{2.22}$$

for some constant c independent of P or f. We now demonstrate this by contradiction. Fix $d\in\mathbb{N}$. Suppose there exists an $s\in(1/2,\infty)$ and a $c\in(0,\infty)$ such that (2.22) holds. Since

$$\lim_{\beta\to+\infty}\frac{(\frac{2}{3}\wedge s)\cdot\beta}{d+\beta}=\frac{2}{3}\wedge s>1/2=\lim_{\beta\to+\infty}\frac{\beta}{2\beta+d},$$

we can choose β large enough such that $\frac{(\frac{2}{3}\wedge s)\cdot\beta}{d+\beta}>\frac{\beta}{2\beta+d}$. Besides, it follows from $\mathcal{E}_P(f)\leq 1$ and (2.22) that

$$\mathcal{E}_P(f)\leq\left|\mathcal{E}_P(f)\right|^{\frac{\frac{2}{3}\wedge s}{s}}\leq\left|c\cdot\left(\mathcal{E}_P^{\phi}(f)\right)^{s}\right|^{\frac{\frac{2}{3}\wedge s}{s}}\leq(1+c)\cdot\left(\mathcal{E}_P^{\phi}(f)\right)^{(\frac{2}{3}\wedge s)} \tag{2.23}$$

for any $f\in\mathcal{F}_d$ and any $P\in\mathcal{H}_0^d$. Let $r=3$ and $\hat{f}_n^{\mathbf{FNN}}$ be the estimator in Theorem 2.2. Then it follows from (2.17), (2.19), (2.23) and Hölder's inequality that

$$\begin{aligned}
n^{-\frac{\beta}{2\beta+d}} &\asymp \inf_{\hat{f}_n}\sup_{P\in\mathcal{H}_1^{d,\beta,r}}\boldsymbol{E}_{P^{\otimes n}}\left[\mathcal{E}_P(\hat{f}_n)\right]\leq\sup_{P\in\mathcal{H}_1^{d,\beta,r}}\boldsymbol{E}_{P^{\otimes n}}\left[\mathcal{E}_P\left(\hat{f}_n^{\mathbf{FNN}}\right)\right]\\
&\leq\sup_{P\in\mathcal{H}_1^{d,\beta,r}}\boldsymbol{E}_{P^{\otimes n}}\left[(1+c)\cdot\left(\mathcal{E}_P^{\phi}(\hat{f}_n^{\mathbf{FNN}})\right)^{(\frac{2}{3}\wedge s)}\right]
\end{aligned}$$

$$
\begin{aligned}
&\leq (1+c)\cdot \sup_{P\in\mathcal{H}_1^{d,\beta,r}} \left(\boldsymbol{E}_{P^{\otimes n}}\left[\mathcal{E}_P^{\phi}(\hat{f}_n^{\mathbf{FNN}})\right]\right)^{(\frac{2}{3}\wedge s)} \\
&\leq (1+c)\cdot \left(\sup_{P\in\mathcal{H}_1^{d,\beta,r}} \boldsymbol{E}_{P^{\otimes n}}\left[\mathcal{E}_P^{\phi}(\hat{f}_n^{\mathbf{FNN}})\right]\right)^{(\frac{2}{3}\wedge s)} \\
&\lesssim \left(\left(\frac{(\log n)^5}{n}\right)^{\frac{\beta}{\beta+d}}\right)^{(\frac{2}{3}\wedge s)} = \left(\frac{(\log n)^5}{n}\right)^{\frac{(\frac{2}{3}\wedge s)\cdot\beta}{\beta+d}}.
\end{aligned}
$$

Hence $n^{-\frac{\beta}{2\beta+d}} \lesssim \left(\frac{(\log n)^5}{n}\right)^{\frac{(\frac{2}{3}\wedge s)\cdot\beta}{\beta+d}}$, which contradicts the fact that $\frac{(\frac{2}{3}\wedge s)\cdot\beta}{d+\beta} > \frac{\beta}{2\beta+d}$. This proves the desired result. Due to the optimality of (2.20) and the minimax lower bound $\mathcal{O}(n^{-\frac{\beta}{d+\beta}})$ for rates of convergence of the excess ϕ-risk stated in Corollary 2.1, we deduce that rates of convergence of the excess misclassification error obtained directly from those of the excess ϕ-risk and the calibration inequality which takes the form of (2.22) can never be faster than $\mathcal{O}(n^{-\frac{\beta}{2d+2\beta}})$. Therefore, the convergence rate $\mathcal{O}(\left(\frac{(\log n)^5}{n}\right)^{\frac{\beta}{2\beta+2d}})$ of the excess misclassification error in (2.18) is the fastest one (up to the logarithmic term $(\log n)^{\frac{5\beta}{2\beta+2d}}$) among all those that are derived directly from the convergence rates of the excess ϕ-risk and the calibration inequality of the form (2.22), which justifies the tightness of (2.18).

It should be pointed out that the rate $\mathcal{O}(\left(\frac{(\log n)^5}{n}\right)^{\frac{\beta}{2\beta+2d}})$ in (2.18) can be further improved if we assume the following noise condition (cf. Tsybakov (2004)) on P: there exist $c_1 > 0$, $t_1 > 0$ and $s_1 \in [0,\infty]$ such that

$$
P_X\left(\left\{x\in[0,1]^d \middle| \left|2\cdot P(\{1\}|x)-1\right| \leq t\right\}\right) \leq c_1 t^{s_1}, \quad \forall\, 0 < t \leq t_1. \tag{2.24}
$$

This condition measures the size of high-noisy points and reflects the noise level through the exponent $s_1 \in [0,\infty]$. Obviously, every distribution satisfies condition (2.24) with $s_1 = 0$ and $c_1 = 1$, whereas $s_1 = \infty$ implies that we have a low amount of noise in labeling x, i.e., the conditional probability function $P(\{1\}|x)$ is bounded away from $1/2$ for P_X-almost all $x \in [0,1]^d$. Under the noise condition (2.24), the calibration inequality for logistic loss ϕ can be refined as

$$
\mathcal{E}_P(f) \leq \bar{c}\cdot\left(\mathcal{E}_P^{\phi}(f)\right)^{\frac{s_1+1}{s_1+2}} \text{ for all } f\in\mathcal{F}_d, \tag{2.25}
$$

where $\bar{c} \in (0,\infty)$ is a constant only depending on (s_1, c_1, t_1), and $\frac{s_1+1}{s_1+2} := 1$ if $s_1 = \infty$ (cf. Theorem 8.29 in Steinwart and Christmann (2008) and Theorem 1.1 in Xiang (2011)). Combining this calibration inequality (2.25) and (2.17), we can obtain an improved generalization bound given by

$$
\sup_{P\in\mathcal{H}_{2,s_1,c_1,t_1}^{d,\beta,r}} \boldsymbol{E}_{P^{\otimes n}}\left[\mathcal{E}_P\left(\hat{f}_n^{\mathbf{FNN}}\right)\right] \lesssim \left(\frac{(\log n)^5}{n}\right)^{\frac{(s_1+1)\beta}{(s_1+2)(\beta+d)}},
$$

where

$$\mathcal{H}^{d,\beta,r}_{2,s_1,c_1,t_1} := \left\{ P \in \mathcal{H}^{d,\beta,r}_1 \middle| P \text{ satisfies (2.24)} \right\}. \tag{2.26}$$

One can refer to Section 3 for more discussions about comparisons between Theorem 2.2 and other related results.

In our Theorem 2.2, the rates $\left(\frac{(\log n)^5}{n}\right)^{\frac{\beta}{\beta+d}}$ and $\left(\frac{(\log n)^5}{n}\right)^{\frac{\beta}{2\beta+2d}}$ become slow when the dimension d is large. This phenomenon, known as the curse of dimensionality, arises in our Theorem 2.2 because our assumption on the data distribution P is very mild and general. Except for the Hölder smoothness condition on the conditional probability function η of P, we do not require any other assumptions in our Theorem 2.2. The curse of dimensionality cannot be circumvented under such general assumption on P, as shown in Corollary 2.1 and (2.19). Therefore, to overcome the curse of dimensionality, we need other assumptions. In our next theorem, we assume that η is the composition of several multivariate vector-valued functions $h_q \circ \cdots \circ h_1 \circ h_0$ such that each component function of h_i is either a Hölder smooth function whose output values only depend on a small number of its input variables, or the function computing the maximum value of some of its input variables (see (2.32) and (2.34)). Under this assumption, the curse of dimensionality is circumvented because each component function of h_i is either essentially defined on a low-dimensional space or a very simple maximum value function. Our hierarchical composition assumption on the conditional probability function is convincing and likely to be met in practice because many phenomena in natural sciences can be "described well by processes that take place at a sequence of increasing scales and are local at each scale, in the sense that they can be described well by neighbor-to-neighbor interactions" (Appendix 2 of Poggio et al. (2016)). Similar compositional assumptions have been adopted in many works such as Schmidt-Hieber (2020); Kohler and Langer (2020); Kohler et al. (2022). One may refer to Poggio et al. (2015, 2016, 2017); Kohler et al. (2022) for more discussions about the reasonableness of such compositional assumptions.

In our compositional assumption mentioned above, we allow the component function of h_i to be the maximum value function, which is not Hölder-β smooth when $\beta > 1$. The maximum value function is incorporated because taking the maximum value is an important operation to pass key information from lower scale levels to higher ones, which appears naturally in the compositional structure of the conditional probability function η in practical classification problems. To see this, let us consider the following example. Suppose the classification problem is to determine whether an input image contains a cat. We assume the data is perfectly classified, in the sense that the conditional probability function η is equal to zero or one almost surely. It should be noted that the assumption "$\eta = 0$ or 1 almost surely" does not conflict with the continuity of η because the support of the distribution of the input data may be unconnected. This classification task can be done by human beings through considering each subpart of the input image and determining whether each subpart contains a cat. Mathematically, let $\mathcal{V}$ be a family of subset of $\{1, 2, \ldots, d\}$ which consists of all the index sets of those (considered) subparts of the input image $x \in [0, 1]^d$. $\mathcal{V}$ should satisfy

$$\bigcup_{J \in \mathcal{V}} J = \{1, 2, \ldots, d\}$$

because the union of all the subparts should cover the input image itself. For each $J \in \mathcal{V}$, let

$$\eta_J((x)_J) = \begin{cases} 1, & \text{if the subpart } (x)_J \text{ of the input image } x \text{ contains a cat,} \\ 0, & \text{if the subpart } (x)_J \text{ of the input image } x \text{ doesn't contains a cat.} \end{cases}$$

Then we will have $\eta(x) = \max_{J \in \mathcal{V}} \{\eta_J((x)_J)\}$ a.s. because

$$\begin{aligned} &\eta(x) = 1 \xLeftrightarrow{\text{a.s.}} x \text{ contains a cat} \Leftrightarrow \text{at least one of the subpart } (x)_J \text{ contains a cat} \\ &\Leftrightarrow \eta_J((x)_J) = 1 \text{ for at least one } J \in \mathcal{V} \Leftrightarrow \max_{J \in \mathcal{V}} \{\eta_J((x)_J)\} = 1. \end{aligned}$$

Hence the maximum value function emerges naturally in the expression of η.

We now give the specific mathematical definition of our compositional model. For any $(d, d_\star, d_*, \beta, r) \in \mathbb{N} \times \mathbb{N} \times \mathbb{N} \times (0, \infty) \times (0, \infty)$, define

$$\mathcal{G}_d^{\mathbf{M}}(d_\star) := \left\{ f : [0,1]^d \to \mathbb{R} \,\middle|\, \begin{array}{l} \exists\, I \subset \{1, 2, \dots, d\} \text{ such that } 1 \le \#(I) \le d_\star \\ \text{and } f(x) = \max\{(x)_i \,|i \in I\}, \,\forall\, x \in [0,1]^d \end{array} \right\}, \tag{2.27}$$

and

$$\begin{aligned} &\mathcal{G}_d^{\mathbf{H}}(d_*, \beta, r) \\ &:= \left\{ f : [0,1]^d \to \mathbb{R} \,\middle|\, \begin{array}{l} \exists\, I \subset \{1, 2, \dots, d\} \text{ and } g \in \mathcal{B}_r^\beta\left([0,1]^{d_*}\right) \text{ such that} \\ \#(I) = d_* \text{ and } f(x) = g\left((x)_I\right) \text{ for all } x \in [0,1]^d \end{array} \right\}. \end{aligned} \tag{2.28}$$

Thus $\mathcal{G}_d^{\mathbf{M}}(d_\star)$ consists of all functions from $[0,1]^d$ to $\mathbb{R}$ which compute the maximum value of at most $d_\star$ components of their input vectors, and $\mathcal{G}_d^{\mathbf{H}}(d_*, \beta, r)$ consists of all functions from $[0,1]^d$ to $\mathbb{R}$ which only depend on d_* components of the input vector and are Hölder-β smooth with corresponding Hölder-β norm less than or equal to r. Obviously,

$$\mathcal{G}_d^{\mathbf{H}}(d_*, \beta, r) = \varnothing, \ \forall\, (d, d_*, \beta, r) \in \mathbb{N} \times \mathbb{N} \times (0, \infty) \times (0, \infty) \text{ with } d < d_*. \tag{2.29}$$

Next, for any $(d_\star, d_*, \beta, r) \in \mathbb{N} \times \mathbb{N} \times (0, \infty) \times (0, \infty)$, define $\mathcal{G}_\infty^{\mathbf{H}}(d_*, \beta, r) := \bigcup_{d=1}^\infty \mathcal{G}_d^{\mathbf{H}}(d_*, \beta, r)$ and $\mathcal{G}_\infty^{\mathbf{M}}(d_\star) := \bigcup_{d=1}^\infty \mathcal{G}_d^{\mathbf{M}}(d_\star)$. Finally, for any $q \in \mathbb{N} \cup \{0\}$, any $(\beta, r) \in (0, \infty)^2$ and any $(d, d_\star, d_*, K) \in \mathbb{N}^4$ with

$$d_* \le \min\left\{d, K + \mathbb{1}_{\{0\}}(q) \cdot (d - K)\right\}, \tag{2.30}$$

define

$$\mathcal{G}_d^{\mathbf{CH}}(q,K,d_*,\beta,r) := \left\{ h_q\circ\cdots\circ h_1\circ h_0 \;\middle|\; \begin{array}{l} h_0,h_1,\ldots,h_{q-1},h_q \text{ are functions satisfying the following conditions:} \\ \text{(i) } \mathbf{dom}(h_i)=[0,1]^K \text{ for } 0<i\le q \text{ and } \mathbf{dom}(h_0)=[0,1]^d; \\ \text{(ii) } \mathbf{ran}(h_i)\subset[0,1]^K \text{ for } 0\le i<q \text{ and } \mathbf{ran}(h_q)\subset\mathbb{R}; \\ \text{(iii) } h_q\in\mathcal{G}_\infty^{\mathbf{H}}(d_*,\beta,r); \\ \text{(iv) For } 0\le i<q \text{ and } 1\le j\le K, \text{ the } j\text{-th coordinate function of } h_i \text{ given by } \mathbf{dom}(h_i)\ni x\mapsto (h_i(x))_j\in\mathbb{R} \text{ belongs to } \mathcal{G}_\infty^{\mathbf{H}}(d_*,\beta,r) \end{array} \right\} \tag{2.31}$$

and

$$\mathcal{G}_d^{\mathbf{CHOM}}(q,K,d_\star,d_*,\beta,r) := \left\{ h_q\circ\cdots\circ h_1\circ h_0 \;\middle|\; \begin{array}{l} h_0,h_1,\ldots,h_{q-1},h_q \text{ are functions satisfying the following conditions:} \\ \text{(i) } \mathbf{dom}(h_i)=[0,1]^K \text{ for } 0<i\le q \text{ and } \mathbf{dom}(h_0)=[0,1]^d; \\ \text{(ii) } \mathbf{ran}(h_i)\subset[0,1]^K \text{ for } 0\le i<q \text{ and } \mathbf{ran}(h_q)\subset\mathbb{R}; \\ \text{(iii) } h_q\in\mathcal{G}_\infty^{\mathbf{H}}(d_*,\beta,r)\cup\mathcal{G}_\infty^{\mathbf{M}}(d_\star); \\ \text{(iv) For } 0\le i<q \text{ and } 1\le j\le K, \text{ the } j\text{-th coordinate function of } h_i \text{ given by } \mathbf{dom}(h_i)\ni x\mapsto (h_i(x))_j\in\mathbb{R} \text{ belongs to } \mathcal{G}_\infty^{\mathbf{H}}(d_*,\beta,r)\cup\mathcal{G}_\infty^{\mathbf{M}}(d_\star) \end{array} \right\}. \tag{2.32}$$

Obviously, we always have that $\mathcal{G}_d^{\mathbf{CH}}(q,K,d_*,\beta,r)\subset\mathcal{G}_d^{\mathbf{CHOM}}(q,K,d_\star,d_*,\beta,r)$. The condition (2.30), which is equivalent to

$$d_*\le\begin{cases} d, & \text{if } q=0,\\ d\wedge K, & \text{if } q>0,\end{cases}$$

is required in the above definitions because it follows from (2.29) that

$$\mathcal{G}_d^{\mathbf{CH}}(q,K,d_*,\beta,r)=\varnothing \text{ if } d_*>\min\left\{d,K+\mathbb{1}_{\{0\}}(q)\cdot(d-K)\right\}.$$

Thus we impose the condition (2.30) simply to avoid the trivial empty set. The space $\mathcal{G}_d^{\mathbf{CH}}(q,K,d_*,\beta,r)$ consists of composite functions $h_q\circ\cdots h_1\circ h_0$ satisfying that each component function of h_i only depends on d_* components of its input vector and is Hölder-β smooth with corresponding Hölder-β norm less than or equal to r. For example, the function

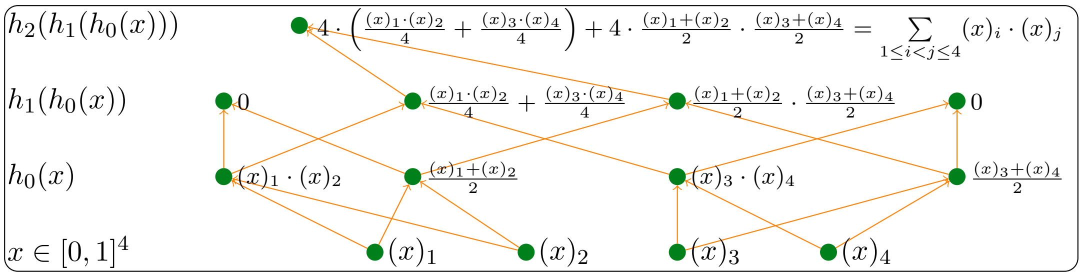

Figure 2.1: An illustration of the function $[0,1]^4 \ni x \mapsto \sum_{1\le i<j\le 4} (x)_i \cdot (x)_j \in \mathbb{R}$, which belongs to $\mathcal{G}_4^{\mathbf{CH}}(2,4,2,2,8)$.

$[0,1]^4 \ni x \mapsto \sum_{1\le i<j\le 4} (x)_i \cdot (x)_j \in \mathbb{R}$ belongs to $\mathcal{G}_4^{\mathbf{CH}}(2,4,2,2,8)$ (cf. Figure 2.1). The definition of $\mathcal{G}_d^{\mathbf{CHOM}}(q,K,d_\star,d_*,\beta,r)$ is similar to that of $\mathcal{G}_d^{\mathbf{CH}}(q,K,d_*,\beta,r)$. The only difference is that, in comparison to $\mathcal{G}_d^{\mathbf{CH}}(q,K,d_*,\beta,r)$, we in the definition of $\mathcal{G}_d^{\mathbf{CHOM}}(q,K,d_\star,d_*,\beta,r)$ additionally allow the component function of h_i to be the function which computes the maximum value of at most $d_\star$ components of its input vector. For example, the function $[0,1]^4 \ni x \mapsto \max_{1\le i<j\le 4} (x)_i \cdot (x)_j \in \mathbb{R}$ belongs to $\mathcal{G}_4^{\mathbf{CHOM}}(2,6,3,2,2,2)$ (cf. Figure 2.2). From the above description of the spaces $\mathcal{G}_d^{\mathbf{CH}}(q,K,d_*,\beta,r)$ and $\mathcal{G}_d^{\mathbf{CHOM}}(q,K,d_\star,d_*,\beta,r)$, we see that the condition (2.30) is very natural because it merely requires the essential input dimension d_* of the Hölder-β smooth component function of h_i to be less than or equal to its actual input dimension, which is d (if $i=0$) or K (if $i>0$). At last, we point out that the space $\mathcal{G}_d^{\mathbf{CH}}(q,K,d_*,\beta,r)$ reduces to the Hölder ball $\mathcal{B}_r^\beta([0,1]^d)$ when $q=0$ and $d_*=d$. Indeed, we have that

$$\begin{aligned}&\mathcal{B}_r^\beta([0,1]^d) = \mathcal{G}_d^{\mathbf{H}}(d,\beta,r) = \mathcal{G}_d^{\mathbf{CH}}(0,K,d,\beta,r)\\ &\subset \mathcal{G}_d^{\mathbf{CHOM}}(0,K,d_\star,d,\beta,r),\ \forall\, K\in\mathbb{N},\ d\in\mathbb{N},\ d_\star\in\mathbb{N},\ \beta\in(0,\infty), r\in(0,\infty).\end{aligned} \tag{2.33}$$

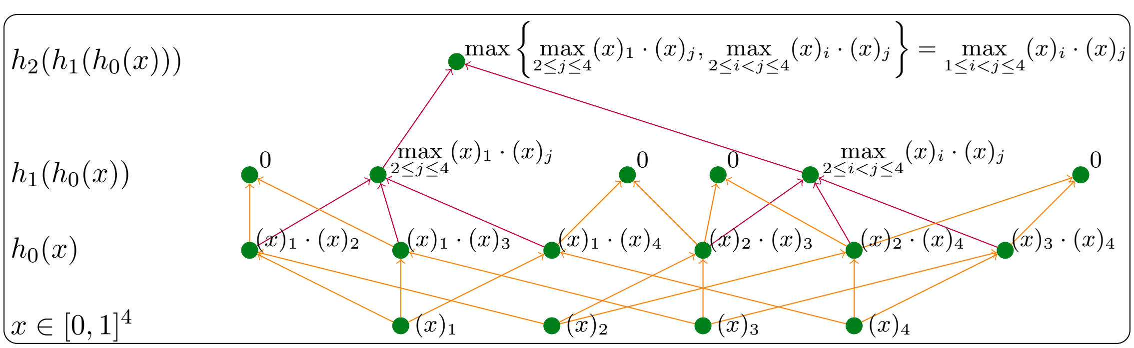

Figure 2.2: An illustration of the function $[0,1]^4 \ni x \mapsto \max_{1\le i<j\le 4} (x)_i \cdot (x)_j \in \mathbb{R}$, which belongs to $\mathcal{G}_4^{\mathbf{CHOM}}(2,6,3,2,2,2)$.

Now we are in a position to state our Theorem 2.3, where we establish sharp convergence rates, which are free from the input dimension d, for fully connected DNN classifiers trained with the logistic loss under the assumption that the conditional probability function η of the data distribution belongs to $\mathcal{G}_d^{\mathbf{CHOM}}(q, K, d_\star, d_*, \beta, r)$. In particular, it can be shown the convergence rate of the excess logistic risk stated in (2.36) in Theorem 2.3 is optimal (up to some logarithmic term). Since $\mathcal{G}_d^{\mathbf{CH}}(q, K, d_*, \beta, r) \subset \mathcal{G}_d^{\mathbf{CHOM}}(q, K, d_\star, d_*, \beta, r)$, the same convergences rates as in Theorem 2.3 can also be achieved under the slightly narrower assumption that η belongs to $\mathcal{G}_d^{\mathbf{CH}}(q, K, d_*, \beta, r)$. The results of Theorem 2.3 break the curse of dimensionality and help explain why deep neural networks perform well, especially in high-dimensional problems.

Theorem 2.3 *Let $q \in \mathbb{N} \cup \{0\}$, $(d, d_\star, d_*, K) \in \mathbb{N}^4$ with $d_* \le \min\left\{d, K + \mathbb{1}_{\{0\}}(q) \cdot (d-K)\right\}$, $(\beta, r) \in (0,\infty)^2$, $n \in \mathbb{N}$, $\nu \in [0,\infty)$, $\{(X_i, Y_i)\}_{i=1}^n$ be an i.i.d. sample in $[0,1]^d \times \{-1,1\}$ and $\hat{f}_n^{\mathbf{FNN}}$ be an ERM with respect to the logistic loss $\phi(t) = \log\left(1 + \mathrm{e}^{-t}\right)$ over the space $\mathcal{F}_d^{\mathbf{FNN}}(G, N, S, B, F)$, which is given by (2.14). Define*

$$\mathcal{H}^{d,\beta,r}_{4,q,K,d_\star,d_*} := \left\{ P \in \mathcal{H}_0^d \,\middle|\, \begin{matrix} P_X(\{ z \in [0,1]^d \,|\, P(\{1\}\,|z) = \hat{\eta}(z)\}) = 1 \\ \text{for some } \hat{\eta} \in \mathcal{G}_d^{\mathbf{CHOM}}(q, K, d_\star, d_*, \beta, r) \end{matrix} \right\}. \tag{2.34}$$

Then there exists a constant $\mathrm{c} \in (0,\infty)$ only depending on $(d_\star, d_*, \beta, r, q)$, such that the estimator $\hat{f}_n^{\mathbf{FNN}}$ defined by (2.14) with

$$\begin{aligned} &\mathrm{c}\log n \le G \lesssim \log n,\ N \asymp \left(\frac{(\log n)^5}{n}\right)^{\frac{-d_*}{d_*+\beta\cdot(1\wedge\beta)^q}},\ S \asymp \left(\frac{(\log n)^5}{n}\right)^{\frac{-d_*}{d_*+\beta\cdot(1\wedge\beta)^q}} \cdot \log n, \\ &1 \le B \lesssim n^\nu,\ \text{and}\ \frac{\beta\cdot(1\wedge\beta)^q}{d_*+\beta\cdot(1\wedge\beta)^q}\cdot\log n \le F \lesssim \log n \end{aligned} \tag{2.35}$$

satisfies

$$\sup_{P \in \mathcal{H}^{d,\beta,r}_{4,q,K,d_\star,d_*}} \boldsymbol{E}_{P^{\otimes n}}\left[\mathcal{E}_P^\phi\left(\hat{f}_n^{\mathbf{FNN}}\right)\right] \lesssim \left(\frac{(\log n)^5}{n}\right)^{\frac{\beta\cdot(1\wedge\beta)^q}{d_*+\beta\cdot(1\wedge\beta)^q}} \tag{2.36}$$

and

$$\sup_{P \in \mathcal{H}^{d,\beta,r}_{4,q,K,d_\star,d_*}} \boldsymbol{E}_{P^{\otimes n}}\left[\mathcal{E}_P\left(\hat{f}_n^{\mathbf{FNN}}\right)\right] \lesssim \left(\frac{(\log n)^5}{n}\right)^{\frac{\beta\cdot(1\wedge\beta)^q}{2d_*+2\beta\cdot(1\wedge\beta)^q}}. \tag{2.37}$$

The proof of Theorem 2.3 is given in Appendix C.4. Note that Theorem 2.3 directly leads to Theorem 2.2 because it follows from (2.33) that

$$\mathcal{H}_1^{d,\beta,r} \subset \mathcal{H}^{d,\beta,r}_{4,q,K,d_\star,d_*} \ \text{ if } q = 0,\ d_* = d \text{ and } d_\star = K = 1.$$

Consequently, Theorem 2.3 can be regarded as a generalization of Theorem 2.2. Note that both the rates $\mathcal{O}(\left(\frac{(\log n)^5}{n}\right)^{\frac{\beta\cdot(1\wedge\beta)^q}{d_*+\beta\cdot(1\wedge\beta)^q}})$ and $\mathcal{O}(\left(\frac{(\log n)^5}{n}\right)^{\frac{\beta\cdot(1\wedge\beta)^q}{2d_*+2\beta\cdot(1\wedge\beta)^q}})$ in (2.36) and (2.37) are independent of the input dimension d, thereby overcoming the curse of dimensionality. Moreover, according to Theorem 2.6 and the comments therein, the rate $\left(\frac{(\log n)^5}{n}\right)^{\frac{\beta\cdot(1\wedge\beta)^q}{d_*+\beta\cdot(1\wedge\beta)^q}}$

in (2.36) for the convergence of the excess logistic risk is even optimal (up to some logarithmic factor). This justifies the sharpness of Theorem 2.3.

Next, we would like to demonstrate the main idea of the proof of Theorem 2.3. The strategy we adopted is to apply Theorem 2.1 with a suitable ψ satisfying (2.7). Let P be an arbitrary probability in $\mathcal{H}^{d,\beta,r}_{4,q,K,d_\star,d_*}$ and denote by η the conditional probability function $P(\{1\}\,|\cdot)$ of P. According to the previous discussions, we cannot simply take $\psi(x,y) = \phi(yf^*_{\phi,P}(x))$ as the target function $f^*_{\phi,P} = \log\frac{\eta}{1-\eta}$ is unbounded. Instead, we define ψ by (2.9) for some carefully selected $\delta_1 \in (0, 1/2]$. For such ψ, we prove

$$\int_{[0,1]^d\times\{-1,1\}} \psi\,(x,y)\mathrm{d}P(x,y) = \inf\left\{ \mathcal{R}^{\phi}_P(f) \middle|\, f : [0,1]^d \to \mathbb{R} \text{ is measurable}\right\} \tag{2.38}$$

in Lemma C.3, and establish a tight inequality of form (2.5) with $\Gamma = \mathcal{O}((\log\frac{1}{\delta_1})^2)$ in Lemma C.10. We then calculate the covering numbers of $\mathcal{F} := \mathcal{F}^{\mathbf{FNN}}_d(G, N, S, B, F)$ by Corollary A.1 and use Lemma C.15 to estimate the approximation error

$$\inf_{f\in\mathcal{F}}\left(\mathcal{R}^{\phi}_P(f) - \int_{[0,1]^d\times\{-1,1\}} \psi(x,y)\mathrm{d}P(x,y)\right)$$

which is essentially $\inf_{f\in\mathcal{F}}\mathcal{E}^{\phi}_P(f)$. Substituting the above estimations into the right hand side of (2.6) and taking supremum over $P \in \mathcal{H}^{d,\beta,r}_{4,q,K,d_\star,d_*}$, we obtain (2.36). We then derive (2.37) from (2.36) through the calibration inequality (2.21).

We would like to point out that the above scheme for obtaining generalization bounds, which is built on our novel oracle-type inequality in Theorem 2.1 with a carefully constructed ψ, is very general. This scheme can be used to establish generalization bounds for classification in other settings, provided that the estimation for the corresponding approximation error is given. For example, one can expect to establish generalization bounds for convolutional neural network (CNN) classification with the logistic loss by using Theorem 2.1 together with recent results about CNN approximation. CNNs perform convolutions instead of matrix multiplications in at least one of their layers (cf. Chapter 9 of Goodfellow et al. (2016)). Approximation properties of various CNN architectures have been intensively studied recently. For instance, 1D CNN approximation is studied in Zhou (2020b,a); Mao et al. (2021); Fang et al. (2020), and 2D CNN approximation is investigated in Kohler et al. (2022); He et al. (2022). With the help of these CNN approximation results and classical concentration techniques, generalization bounds for CNN classification have been established in many works such as Kohler and Langer (2020); Kohler et al. (2022); Shen et al. (2021); Feng et al. (2023). In our coming work Zhang et al. (2024), we will derive generalization bounds for CNN classification with logistic loss on spheres under the Sobolev smooth conditional probability assumption through the novel framework developed in our paper.

In our proof of Theorem 2.2 and Theorem 2.3, a tight error bound for neural network approximation of the logarithm function $\log(\cdot)$ arises as a by-product. Indeed, for a given data distribution P on $[0,1]^d \times \{-1,1\}$, to estimate the approximation error of $\mathcal{F}^{\mathbf{FNN}}_d$, we need to construct neural networks to approximate the target function $f^*_{\phi,P} = \log\frac{\eta}{1-\eta}$, where η denotes the conditional probability function of P. Due to the unboundedness of $f^*_{\phi,P}$, one cannot approximate $f^*_{\phi,P}$ directly. To overcome this difficulty, we consider truncating $f^*_{\phi,P}$ to obtain an efficient approximation. We design neural networks $\tilde{\eta}$ and $\tilde{l}$ to approximate η on $[0,1]^d$ and $\log(\cdot)$ on $[\delta_n, 1-\delta_n]$ respectively, where $\delta_n \in (0,1/4]$ is a carefully selected number which depends on the sample size n and tends to zero as $n\to\infty$. Let Π_{δ_n} denote the clipping function given by $\Pi_{\delta_n}: \mathbb{R}\to[\delta_n,1-\delta_n], t\mapsto \arg\min_{t'\in[\delta_n,1-\delta_n]}|t'-t|$. Then $\tilde{L}: t\mapsto \tilde{l}(\Pi_{\delta_n}(t)) - \tilde{l}(1-\Pi_{\delta_n}(t))$ is a neural network which approximates the function

$$\overline{L}_{\delta_n}: t\mapsto \begin{cases}\log\frac{t}{1-t}, & \text{if } t\in[\delta_n,1-\delta_n],\\ \log\frac{1-\delta_n}{\delta_n}, & \text{if } t>1-\delta_n,\\ \log\frac{\delta_n}{1-\delta_n}, & \text{if } t<\delta_n,\end{cases} \tag{2.39}$$

meaning that the function $\tilde{L}(\tilde{\eta}(x)) = \tilde{l}\left(\Pi_{\delta_n}\left(\tilde{\eta}(x)\right)\right) - \tilde{l}\left(\Pi_{\delta_n}\left(1-\tilde{\eta}(x)\right)\right)$ is a neural network which approximates the truncated $f^*_{\phi,P}$ given by

$$\overline{L}_{\delta_n}\circ\eta: x\mapsto \overline{L}_{\delta_n}(\eta(x)) = \begin{cases} f^*_{\phi,P}(x), & \text{if } \left|f^*_{\phi,P}(x)\right|\le\log\dfrac{1-\delta_n}{\delta_n},\\ \operatorname{sgn}(f^*_{\phi,P}(x))\log\dfrac{1-\delta_n}{\delta_n}, & \text{otherwise.}\end{cases}$$

One can build $\tilde{\eta}$ by applying some existing results on approximation theory of neural networks (see Appendix B). However, the construction of $\tilde{l}$ requires more effort. Since the logarithm function $\log(\cdot)$ is unbounded near 0, which leads to the blow-up of its Hölder norm on $[\delta_n,1-\delta_n]$ when δ_n is becoming small, existing conclusions, e.g., the results in Appendix B, cannot yield satisfactory error bounds for neural network approximation of $\log(\cdot)$ on $[\delta_n,1-\delta_n]$. To see this, let us consider using Theorem B.1 to estimate the approximation error directly. Note that approximating $\log(\cdot)$ on $[\delta_n,1-\delta_n]$ is equivalent to approximating $l_{\delta_n}(t) := \log((1-2\delta_n)t+\delta_n)$ on $[0,1]$. For $\beta_1>0$ with $k=\lceil\beta_1-1\rceil$ and $\lambda=\beta_1-\lceil\beta_1-1\rceil$, denote by $l^{(k)}_{\delta_n}$ the k-th derivative of l_{δ_n}. Then there holds

$$\begin{aligned}
&\|l_{\delta_n}\|_{\mathcal{C}^{k,\lambda}([0,1])} \ge \sup_{0\le t<t'\le 1}\frac{\left|l^{(k)}_{\delta_n}(t)-l^{(k)}_{\delta_n}(t')\right|}{|t-t'|^\lambda}\\
&\ge \frac{\left|l^{(k)}_{\delta_n}(0)-l^{(k)}_{\delta_n}\left(\frac{\delta_n}{1-2\delta_n}\right)\right|}{\left|0-\frac{\delta_n}{1-2\delta_n}\right|^\lambda} \ge \inf_{t\in\left[0,\frac{\delta_n}{1-2\delta_n}\right]}\frac{\left|l^{(k+1)}_{\delta_n}(t)\right|\cdot\left|0-\frac{\delta_n}{1-2\delta_n}\right|}{\left|0-\frac{\delta_n}{1-2\delta_n}\right|^\lambda}\\
&= \inf_{t\in\left[0,\frac{\delta_n}{1-2\delta_n}\right]}\frac{\left|\frac{k!}{((1-2\delta_n)t+\delta_n)^{k+1}}\right|\cdot(1-2\delta_n)^{k+1}\cdot\left|0-\frac{\delta_n}{1-2\delta_n}\right|}{\left|0-\frac{\delta_n}{1-2\delta_n}\right|^\lambda} = \frac{k!}{2^{k+1}}\cdot(1-2\delta_n)^{k+\lambda}\cdot\frac{1}{\delta_n^{k+\lambda}}.
\end{aligned}$$

Hence it follows from $\delta_n\in(0,1/4]$ that

$$\|l_{\delta_n}\|_{\mathcal{C}^{k,\lambda}([0,1])} \ge \frac{\lceil\beta_1-1\rceil!}{2^{\lceil\beta_1\rceil}}\cdot(1-2\delta_n)^{\beta_1}\cdot\frac{1}{\delta_n^{\beta_1}} \ge \frac{\lceil\beta_1-1\rceil!}{4^{\lceil\beta_1\rceil}}\cdot\frac{1}{\delta_n^{\beta_1}} \ge \frac{3}{128}\cdot\frac{1}{\delta_n^{\beta_1}}.$$

By Theorem B.1, for any positive integers m and M' with

$$M' \ge \max\left\{(\beta_1+1), \left(\|l_{\delta_n}\|_{\mathcal{C}^{k,\lambda}([0,1])}\lceil\beta_1\rceil+1\right)\cdot\mathrm{e}\right\} \ge \|l_{\delta_n}\|_{\mathcal{C}^{k,\lambda}([0,1])} \ge \frac{3}{128}\cdot\frac{1}{\delta_n^{\beta_1}}, \tag{2.40}$$

there exists a neural network

$$\tilde{f} \in \mathcal{F}_1^{\mathbf{FNN}}\left(14m(2+\log_2(1\vee\beta_1)), 6\left(1+\lceil\beta_1\rceil\right)M', 987(2+\beta_1)^4M'm, 1, \infty\right) \tag{2.41}$$

such that

$$\begin{aligned}\sup_{x\in[0,1]}\left|l_{\delta_n}(x)-\tilde{f}(x)\right| &\le \|l_{\delta_n}\|_{\mathcal{C}^{k,\lambda}([0,1])}\cdot\lceil\beta_1\rceil\cdot 3^{\beta_1}M'^{-\beta_1}\\ &\quad+\left(1+2\|l_{\delta_n}\|_{\mathcal{C}^{k,\lambda}([0,1])}\cdot\lceil\beta_1\rceil\right)\cdot 6\cdot(2+\beta_1^2)\cdot M'\cdot 2^{-m}.\end{aligned}$$

To make this error less than or equal to a given error threshold ε_n (depending on n), there must hold

$$\varepsilon_n \ge \|l_{\delta_n}\|_{\mathcal{C}^{k,\lambda}([0,1])}\cdot\lceil\beta_1\rceil\cdot 3^{\beta_1}M'^{-\beta_1} \ge \|l_{\delta_n}\|_{\mathcal{C}^{k,\lambda}([0,1])}\cdot M'^{-\beta_1} \ge M'^{-\beta_1}\cdot\frac{3}{128}\cdot\frac{1}{\delta_n^{\beta_1}}.$$

This together with (2.40) gives

$$M' \ge \max\left\{\frac{3}{128}\cdot\frac{1}{\delta_n^{\beta_1}}, \varepsilon_n^{-1/\beta_1}\cdot\left|\frac{3}{128}\right|^{1/\beta_1}\cdot\frac{1}{\delta_n}\right\}. \tag{2.42}$$

Consequently, the width and the number of nonzero parameters of $\tilde{f}$ are greater than or equal to the right hand side of (2.42), which may be too large when δ_n is small (recall that $\delta_n \to 0$ as $n\to\infty$). In this paper, we establish a new sharp error bound for approximating the natural logarithm function $\log(\cdot)$ on $[\delta_n, 1-\delta_n]$, which indicates that one can achieve the same approximation error by using a much smaller network. This refined error bound is given in Theorem 2.4 which is critical in our proof of Theorem 2.2 and also deserves special attention in its own right.

Theorem 2.4 *Given $a\in(0,1/2]$, $b\in(a,1]$, $\alpha\in(0,\infty)$ and $\varepsilon\in(0,1/2]$, there exists*

$$\begin{aligned}\tilde{f}\in\mathcal{F}_1^{\mathbf{FNN}}\Bigg(&A_1\log\frac{1}{\varepsilon}+139\log\frac{1}{a},\ A_2\left|\frac{1}{\varepsilon}\right|^{\frac{1}{\alpha}}\cdot\log\frac{1}{a},\\ &A_3\left|\frac{1}{\varepsilon}\right|^{\frac{1}{\alpha}}\cdot\left|\log\frac{1}{\varepsilon}\right|\cdot\log\frac{1}{a}+65440\left|\log\frac{1}{a}\right|^2, 1, \infty\Bigg)\end{aligned}$$

such that

$$\sup_{z\in[a,b]}\left|\log z-\tilde{f}(z)\right|\le\varepsilon \text{ and } \log a\le\tilde{f}(t)\le\log b,\ \forall\, t\in\mathbb{R},$$

where $(A_1,A_2,A_3)\in(0,\infty)^3$ are constants depending only on α.

In Theorem 2.4, we show that for each fixed $\alpha\in(0,\infty)$ one can construct a neural network to approximate the natural logarithm function $\log(\cdot)$ on $[a,b]$ with error ε, where the depth, width and number of nonzero parameters of this neural network are in the same order of magnitude as $\log\frac{1}{\varepsilon}+\log\frac{1}{a}$, $\left(\frac{1}{\varepsilon}\right)^{\frac{1}{\alpha}}\left(\log\frac{1}{a}\right)$ and $\left(\frac{1}{\varepsilon}\right)^{\frac{1}{\alpha}}\left(\log\frac{1}{\varepsilon}\right)\left(\log\frac{1}{a}\right)+\left(\log\frac{1}{a}\right)^2$ respectively. Recall that in our generalization analysis we need to approximate log on $[\delta_n, 1-\delta_n]$, which is equivalent to approximating $l_{\delta_n}(t) = \log((1-2\delta_n)t+\delta_n)$ on $[0,1]$. Let $\varepsilon_n \in (0,1/2]$ denote the desired accuracy of the approximation of l_{δ_n} on $[0,1]$, which depends on the sample size n and converges to zero as $n \to \infty$. Using Theorem 2.4 with $\alpha = 2\beta_1$, we deduce that for any $\beta_1 > 0$ one can approximate l_{δ_n} on $[0,1]$ with error ε_n by a network of which the width and the number of nonzero parameters are less than $C_{\beta_1}\varepsilon_n^{-\frac{1}{2\beta_1}} |\log \varepsilon_n| \cdot |\log \delta_n|^2$ with some constant $C_{\beta_1} > 0$ (depending only on β_1). The complexity of this neural network is much smaller than that of $\tilde{f}$ defined in (2.41) with (2.42) as $n \to \infty$ since $|\log \delta_n|^2 = \mathrm{o}\left(1/\delta_n\right)$ and $\varepsilon_n^{-\frac{1}{2\beta_1}} |\log \varepsilon_n| = \mathrm{o}\left(\varepsilon_n^{-1/\beta_1}\right)$ as $n \to \infty$. In particular, when

$$\frac{1}{n^{\theta_2}} \lesssim \varepsilon_n \wedge \delta_n \leq \varepsilon_n \vee \delta_n \lesssim \frac{1}{n^{\theta_1}} \text{ for some } \theta_2 \geq \theta_1 > 0 \text{ independent of } n \text{ or } \beta_1, \tag{2.43}$$

which occurs in our generalization analysis (e.g., in our proof of Theorem 2.3, we essentially take $\varepsilon_n = \delta_n \asymp \left(\frac{(\log n)^5}{n}\right)^{\frac{\beta\cdot(1\wedge\beta)^q}{d_*+\beta\cdot(1\wedge\beta)^q}}$, meaning that $n^{\frac{-\beta\cdot(1\wedge\beta)^q}{d_*+\beta\cdot(1\wedge\beta)^q}} \lesssim \varepsilon_n = \delta_n \lesssim n^{\frac{-\beta\cdot(1\wedge\beta)^q}{2d_*+\beta\cdot(1\wedge\beta)^q}}$ (cf. (C.74), (C.78), (C.87) and (C.92)), we will have that the right hand side of (2.42) grows no slower than $n^{\theta_1+\theta_1/\beta_1}$. Hence, in this case, no matter what β_1 is, the width and the number of nonzero parameters of the network $\tilde{f}$, which approximates l_{δ_n} on $[0,1]$ with error ε_n and is obtained by using Theorem B.1 directly (cf. (2.41)), will grow faster than n^{θ_1} as $n \to \infty$. However, it follows from Theorem 2.4 that there exists a network $\overline{f}$ of which the width and the number of nonzero parameters are less than $C_{\beta_1}\varepsilon_n^{-\frac{1}{2\beta_1}} |\log \varepsilon_n| \cdot |\log \delta_n|^2 \lesssim n^{\frac{\theta_2}{2\beta_1}} |\log n|^3$ such that it achieves the same approximation error as that of $\tilde{f}$. By taking β_1 large enough we can make the growth (as $n \to \infty$) of the width and the number of nonzero parameters of $\overline{f}$ slower than n^{θ} for arbitrary $\theta \in (0, \theta_1]$. Therefore, in the usual case when the complexity of $\tilde{\eta}$ is not too small in the sense that the width and the number of nonzero parameters of $\tilde{\eta}$ grow faster than n^{θ_3} as $n \to \infty$ for some $\theta_3 \in (0,\infty)$ independent of n or β_1, we can use Theorem 2.4 with a large enough $\alpha = 2\beta_1$ to construct the desired network $\tilde{l}$ of which the complexity is insignificant in comparison to that of $\tilde{L} \circ \tilde{\eta}$. In other words, the neural network approximation of logarithmic function based on Theorem 2.4 brings little complexity in approximating the target function $f^*_{\phi,P}$. The above discussion demonstrates the tightness of the inequality in Theorem 2.4 and the advantage of Theorem 2.4 over those general results on approximation theory of neural networks such as Theorem B.1.

It is worth mentioning that an alternative way to approximate the function $\overline{L}_{\delta_n}$ defined in (2.39) is by simply using its piecewise linear interpolation. For example, in Kohler and Langer (2020), the authors express the piecewise linear interpolation of $\overline{L}_{\delta_n}$ at equidistant points by a neural network $\tilde{L}$, and construct a CNN $\tilde{\eta}$ to approximate η, leading to an approximation of the truncated target function of the logistic risk $\tilde{L} \circ \tilde{\eta}$. It follows from Proposition 3.2.4 of Atkinson and Han (2009) that

$$h_n^2 \lesssim \left\| \tilde{L} - \overline{L}_{\delta_n} \right\|_{[\delta_n, 1-\delta_n]} \lesssim \frac{h_n^2}{\delta_n^2}, \tag{2.44}$$

where h_n denotes the step size of the interpolation. Therefore, to ensure the error bound ε_n for the approximation of $\overline{L}_{\delta_n}$ by $\tilde{L}$, we must have $h_n \lesssim \sqrt{\varepsilon_n}$, implying that the number of nonzero parameters of $\tilde{L}$ will grow no slower than $\frac{1}{h_n} \gtrsim \frac{1}{\sqrt{\varepsilon_n}}$ as $n \to \infty$. Consequently, in the case (2.43), we will have that the number of nonzero parameters of $\tilde{L}$ will grow no slower than $n^{\theta_1/2}$. Therefore, in contrast to using Theorem 2.4, we cannot make the number of nonzero parameters of the network $\tilde{L}$ obtained from piecewise linear interpolation grow slower than n^{θ} for arbitrarily small $\theta > 0$. As a result, using piecewise linear interpolation to approximate $\overline{L}_{\delta_n}$ may bring extra complexity in establishing the approximation of the target function. However, the advantage of using piecewise linear interpolation is that one can make the depth or width of the network $\tilde{L}$ which expresses the desired interpolation bounded as $n \to \infty$ (cf. Lemma 7 in Kohler and Langer (2020) and its proof therein).

The proof of Theorem 2.4 is in Appendix C.3. The key observation in our proof is the fact that for all $k \in \mathbb{N}$, the following holds true:

$$\log x = \log(2^k \cdot x) - k \log 2, \quad \forall\, x \in (0, \infty). \tag{2.45}$$

Then we can use the values of $\log(\cdot)$ which are taken far away from zero (i.e., $\log(2^k \cdot x)$ in the right hand side of (2.45)) to determine its values taken near zero, while approximating the former is more efficient as the Hölder norm of the natural logarithm function on domains far away from zero can be well controlled.

In the next theorem, we show that if the data distribution has a piecewise smooth decision boundary, then DNN classifiers trained by empirical logistic risk minimization can also achieve dimension-free rates of convergence under the noise condition (2.24) and a margin condition (see (2.51) below). Before stating this result, we need to introduce this margin condition and relevant concepts.

We first define the set of (binary) classifiers which have a piecewise Hölder smooth decision boundary. We will adopt similar notations from Kim et al. (2021) to describe this set. Specifically, let $\beta, r \in (0, \infty)$ and $I, \Theta \in \mathbb{N}$. For $g \in \mathcal{B}_r^{\beta}\left([0,1]^{d-1}\right)$ and $j = 1, 2, \cdots, d$, we define horizon function $\Psi_{g,j} : [0,1]^d \to \{0, 1\}$ as $\Psi_{g,j}(x) := \mathbb{1}_{\{(x)_j \geq g(x_{-j})\}}$, where $x_{-j} := ((x)_1, \cdots, (x)_{j-1}, (x)_{j+1}, \cdots, (x)_d) \in [0,1]^{d-1}$. For each horizon function, the corresponding basis piece $\Lambda_{g,j}$ is defined as $\Lambda_{g,j} := \left\{x \in [0,1]^d \,\middle|\, \Psi_{g,j}(x) = 1\right\}$. Note that $\Lambda_{g,j} = \left\{x \in [0,1]^d \,\middle|\, (x)_j \geq \max\{0, g(x_{-j})\}\right\}$. Thus $\Lambda_{g,j}$ is enclosed by the hypersurface $\mathcal{S}_{g,j} := \left\{x \in [0,1]^d \,\middle|\, (x)_j = \max\{0, g(x_{-j})\}\right\}$ and (part of) the boundary of $[0,1]^d$. We then define the set of pieces which are the intersection of I basis pieces as

$$\mathcal{A}^{d,\beta,r,I} := \left\{ A \,\middle|\, A = \bigcap_{k=1}^{I} \Lambda_{g_k, j_k} \text{ for some } j_k \in \{1, 2, \cdots, d\} \text{ and } g_k \in \mathcal{B}_r^{\beta}\left([0,1]^{d-1}\right) \right\},$$

and define $\mathcal{C}^{d,\beta,r,I,\Theta}$ to be a set of binary classifiers as

$$\begin{aligned} &\mathcal{C}^{d,\beta,r,I,\Theta} \\ &:= \left\{ \mathtt{C}(x) = 2\sum_{i=1}^{\Theta} \mathbb{1}_{A_i}(x) - 1 : [0,1]^d \to \{-1, 1\} \,\middle|\, \begin{array}{l} A_1, A_2, A_3, \cdots, A_\Theta \text{ are} \\ \text{disjoint sets in } \mathcal{A}^{d,\beta,r,I} \end{array} \right\}. \end{aligned} \tag{2.46}$$

Thus $\mathcal{C}^{d,\beta,r,I,\Theta}$ consists of all binary classifiers which are equal to $+1$ on some disjoint sets $A_1, \ldots, A_\Theta$ in $\mathcal{A}^{d,\beta,r,I}$ and -1 otherwise. Let $A_t = \cap_{k=1}^{I} \Lambda_{g_{t,k}, j_{t,k}}$ $(t = 1, 2, \ldots, \Theta)$ be arbitrary disjoint sets in $\mathcal{A}^{d,\beta,r,I}$, where $j_{t,k} \in \{1, 2, \ldots, d\}$ and $g_{t,k} \in \mathcal{B}_r^{\beta}\left([0,1]^{d-1}\right)$. Then $\mathtt{C} : [0,1]^d \to \{-1, 1\}\,, x \mapsto 2\sum_{i=1}^{\Theta} \mathbb{1}_{A_i}(x) - 1$ is a classifier in $\mathcal{C}^{d,\beta,r,I,\Theta}$. Recall that $\Lambda_{g_{t,k}, j_{t,k}}$ is enclosed by $\mathcal{S}_{g_{t,k},j_{t,k}}$ and (part of) the boundary of $[0,1]^d$ for each t, k. Hence for each t, the region A_t is enclosed by hypersurfaces $\mathcal{S}_{g_{t,k},j_{t,k}}$ $(k = 1, \ldots, I)$ and (part of) the boundary of $[0,1]^d$. We say the piecewise Hölder smooth hypersurface

$$D^*_{\mathtt{C}} := \bigcup_{t=1}^{\Theta} \bigcup_{k=1}^{I} \left(\mathcal{S}_{g_{t,k},j_{t,k}} \cap A_t\right) \tag{2.47}$$

is the *decision boundary* of the classifier $\mathtt{C}$ because intuitively, points on different sides of $D^*_{\mathtt{C}}$ are classified into different categories (i.e. $+1$ and -1) by $\mathtt{C}$ (cf. Figure 2.3). Denote by $\Delta_{\mathtt{C}}(x)$ the distance from $x \in [0,1]^d$ to the decision boundary $D^*_{\mathtt{C}}$, i.e.,

$$\Delta_{\mathtt{C}}(x) := \inf\left\{ \left\|x - x'\right\|_2 \middle| x' \in D^*_{\mathtt{C}} \right\}. \tag{2.48}$$

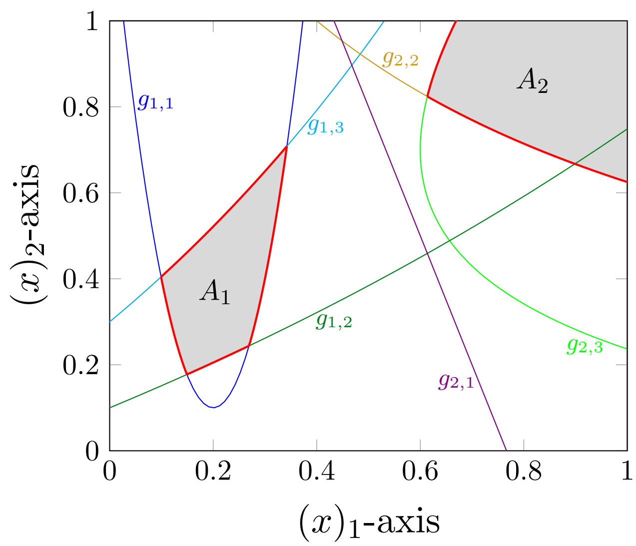

Figure 2.3: Illustration of the sets $A_1, \ldots A_\Theta$ when $d = 2$, $\Theta = 2$, $I = 3$, $j_{2,1} = j_{2,2} = j_{1,1} = j_{1,2} = 2$ and $j_{1,3} = j_{2,3} = 1$. The classifier $\mathtt{C}(x) = 2\sum_{t=1}^{\Theta} \mathbb{1}_{A_t}(x) - 1$ is equal to $+1$ on $A_1 \cup A_2$ and -1 otherwise. The decision boundary $D^*_{\mathtt{C}}$ of $\mathtt{C}$ is marked red.

We then describe the margin condition mentioned above. Let P be a probability measure on $[0,1]^d \times \{-1, 1\}$, which we regard as the joint distribution of the input and output data, and $\eta(\cdot) = P(\{1\}|\cdot)$ is the conditional probability function of P. The corresponding Bayes classifier is the sign of $2\eta - 1$ which minimizes the misclassification error over all measurable functions, i.e.,

$$\mathcal{R}_P(\operatorname{sgn}(2\eta - 1)) = \mathcal{R}_P(2\eta - 1) = \inf\left\{ \mathcal{R}_P(f) \,\middle|\, f : [0,1]^d \to \mathbb{R} \text{ is measurable} \right\}. \tag{2.49}$$

We say the distribution P has a piecewise smooth decision boundary if

$$\exists\, \mathtt{C} \in \mathcal{C}^{d,\beta,r,I,\Theta} \text{ s.t. } \operatorname{sgn}(2\eta - 1) \xlongequal{P_X\text{-a.s.}} \mathtt{C},$$

that is,

$$P_X\left(\left\{x\in[0,1]^d\,\middle|\,\operatorname{sgn}(2\cdot P(\{1\}\,|x)-1)=\mathtt{C}(x)\right\}\right)=1 \tag{2.50}$$

for some $\mathtt{C}\in\mathcal{C}^{d,\beta,r,I,\Theta}$. Suppose $\mathtt{C}\in\mathcal{C}^{d,\beta,r,I,\Theta}$ and (2.50) holds. We call $D^*_{\mathtt{C}}$ the *decision boundary* of P, and for $c_2\in(0,\infty)$, $t_2\in(0,\infty)$, $s_2\in[0,\infty]$, we use the following condition

$$P_X\left(\left\{x\in[0,1]^d\,\middle|\,\Delta_{\mathtt{C}}(x)\le t\right\}\right)\le c_2t^{s_2},\quad\forall\, 0<t\le t_2, \tag{2.51}$$

which we call the *margin condition*, to measure the concentration of the input distribution P_X near the decision boundary $D^*_{\mathtt{C}}$ of P. In particular, when the input data are bounded away from the decision boundary $D^*_{\mathtt{C}}$ of P (P_X-a.s.), (2.51) will hold for $s_2=\infty$.

Now we are ready to give our next main theorem.

Theorem 2.5 *Let $d\in\mathbb{N}\cap[2,\infty)$, $(n,I,\Theta)\in\mathbb{N}^3$, $(\beta,r,t_1,t_2,c_1,c_2)\in(0,\infty)^6$, $(s_1,s_2)\in[0,\infty]^2$, $\{(X_i,Y_i)\}_{i=1}^n$ be a sample in $[0,1]^d\times\{-1,1\}$ and $\hat{f}_n^{\mathbf{FNN}}$ be an ERM with respect to the logistic loss $\phi(t)=\log\left(1+\mathrm{e}^{-t}\right)$ over $\mathcal{F}_d^{\mathbf{FNN}}(G,N,S,B,F)$ which is given by (2.14). Define*

$$\mathcal{H}^{d,\beta,r,I,\Theta,s_1,s_2}_{6,t_1,c_1,t_2,c_2}:=\left\{P\in\mathcal{H}_0^d\,\middle|\,\begin{matrix}(2.24),\ \ (2.50)\ \ \textit{and}\ \ (2.51)\\ \textit{hold for some } \mathtt{C}\in\mathcal{C}^{d,\beta,r,I,\Theta}\end{matrix}\right\}. \tag{2.52}$$

Then the following statements hold true:

(1) For $s_1\in[0,\infty]$ and $s_2=\infty$, the ϕ-ERM $\hat{f}_n^{\mathbf{FNN}}$ with

$$G=G_0\log\frac{1}{t_2\wedge\frac12},\ N=N_0\left(\frac{1}{t_2\wedge\frac12}\right)^{\frac{d-1}{\beta}},\ S=S_0\left(\frac{1}{t_2\wedge\frac12}\right)^{\frac{d-1}{\beta}}\log\left(\frac{1}{t_2\wedge\frac12}\right),$$
$$B=B_0\left(\frac{1}{t_2\wedge\frac12}\right),\ \textit{and}\ \ F\asymp\left(\frac{\log n}{n}\right)^{\frac{1}{s_1+2}}$$

satisfies

$$\sup_{P\in\mathcal{H}^{d,\beta,r,I,\Theta,s_1,s_2}_{6,t_1,c_1,t_2,c_2}}\boldsymbol{E}_{P^{\otimes n}}\left[\mathcal{E}_P\left(\hat{f}_n^{\mathbf{FNN}}\right)\right]\lesssim\left(\frac{\log n}{n}\right)^{\frac{s_1}{s_1+2}}, \tag{2.53}$$

where G_0,N_0,S_0,B_0 are positive constants only depending on d,β,r,I,Θ;

(2) For $s_1=\infty$ and $s_2\in[0,\infty)$, the ϕ-ERM $\hat{f}_n^{\mathbf{FNN}}$ with

$$G\asymp\log n,\ N\asymp\left(\frac{n}{(\log n)^3}\right)^{\frac{d-1}{s_2\beta+d-1}},\ S\asymp\left(\frac{n}{(\log n)^3}\right)^{\frac{d-1}{s_2\beta+d-1}}\log n,$$
$$B\asymp\left(\frac{n}{(\log n)^3}\right)^{\frac{1}{s_2+\frac{d-1}{\beta}}},\ \textit{and}\ \ F=t_1\wedge\frac12$$

satisfies

$$\sup_{P\in\mathcal{H}^{d,\beta,r,I,\Theta,s_1,s_2}_{6,t_1,c_1,t_2,c_2}}\boldsymbol{E}_{P^{\otimes n}}\left[\mathcal{E}_P\left(\hat{f}_n^{\mathbf{FNN}}\right)\right]\lesssim\left(\frac{(\log n)^3}{n}\right)^{\frac{1}{1+\frac{d-1}{\beta s_2}}}; \tag{2.54}$$

(3) For $s_1 \in [0,\infty)$ *and* $s_2 \in [0,\infty)$*, the* ϕ*-ERM* $\hat{f}_n^{\mathbf{FNN}}$ *with*

$$G \asymp \log n,\, N \asymp \left(\frac{n}{(\log n)^3}\right)^{\frac{(d-1)(s_1+1)}{s_2\beta+(s_1+1)(s_2\beta+d-1)}},\, S \asymp \left(\frac{n}{(\log n)^3}\right)^{\frac{(d-1)(s_1+1)}{s_2\beta+(s_1+1)(s_2\beta+d-1)}}\log n,$$

$$B \asymp \left(\frac{n}{(\log n)^3}\right)^{\frac{s_1+1}{s_2+(s_1+1)\left(s_2+\frac{d-1}{\beta}\right)}},\ \textit{and}\ \ F \asymp \left(\frac{(\log n)^3}{n}\right)^{\frac{s_2}{s_2+(s_1+1)\left(s_2+\frac{d-1}{\beta}\right)}}$$

satisfies

$$\sup_{P\in\mathcal{H}^{d,\beta,r,I,\Theta,s_1,s_2}_{6,t_1,c_1,t_2,c_2}} \boldsymbol{E}_{P^{\otimes n}}\left[\mathcal{E}_P\left(\hat{f}_n^{\mathbf{FNN}}\right)\right] \lesssim \left(\frac{(\log n)^3}{n}\right)^{\frac{s_1}{1+(s_1+1)\left(1+\frac{d-1}{\beta s_2}\right)}}. \tag{2.55}$$

It is worth noting that the rate $\mathcal{O}(\left(\frac{\log n}{n}\right)^{\frac{s_1}{s_1+2}})$ established in (2.53) does not depend on the dimension d, and dependency of the rates in (2.54) and (2.55) on the dimension d diminishes as s_2 increases, which demonstrates that the condition (2.51) with $s_2 = \infty$ helps circumvent the curse of dimensionality. In particular, (2.53) will give a fast dimension-free rate of convergence $\mathcal{O}(\frac{\log n}{n})$ if $s_1 = s_2 = \infty$. One may refer to Section 3 for more discussions about the result of Theorem 2.5.

The proof of Theorem 2.5 is in Appendix C.5. Our proof relies on Theorem 2.1 and the fact that the ReLU networks are good at approximating indicator functions of bounded regions with piecewise smooth boundary (Imaizumi and Fukumizu, 2019; Petersen and Voigtlaender, 2018). Let P be an arbitrary probability in $\mathcal{H}^{d,\beta,r,I,\Theta,s_1,s_2}_{6,t_1,c_1,t_2,c_2}$ and denote by η the condition probability function $P(\{1\}|\cdot)$ of P. To apply Theorem 2.1 and make good use of the noise condition (2.24) and the margin condition (2.51), we define another ψ (which is different from that in (2.9)) as

$$\psi : [0,1]^d \times \{-1,1\} \to \mathbb{R}, \quad (x,y) \mapsto \begin{cases} \phi\left(yF_0 \mathrm{sgn}(2\eta(x)-1)\right), & \text{if } |2\eta(x)-1| > \eta_0, \\ \phi\left(y\log\dfrac{\eta(x)}{1-\eta(x)}\right), & \text{if } |2\eta(x)-1| \le \eta_0 \end{cases}$$

for some suitable $\eta_0 \in (0,1)$ and $F_0 \in \left(0, \log\frac{1+\eta_0}{1-\eta_0}\right)$. For such ψ, Lemma C.17 guarantees that inequality (2.3) holds as

$$\int_{[0,1]^d\times\{-1,1\}} \psi(x,y)\mathrm{d}P(x,y) \le \inf\left\{\mathcal{R}_P^\phi(f)\middle| f : [0,1]^d \to \mathbb{R} \text{ is measurable}\right\},$$

and (2.4), (2.5) of Theorem 2.1 are satisfied with $M = \frac{2}{1-\eta_0}$ and $\Gamma = \frac{8}{1-\eta_0^2}$. Moreover, we use the noise condition (2.24) and the margin condition (2.51) to bound the approximation error

$$\inf_{f\in\mathcal{F}_d^{\mathbf{FNN}}(G,N,S,B,F)} \left(\mathcal{R}_P^\phi(f) - \int_{[0,1]^d\times\{-1,1\}} \psi(x,y)\mathrm{d}P(x,y)\right) \tag{2.56}$$

(see (C.111), (C.112), (C.113)). Then, as in the proof of Theorem 2.2, we combine Theorem 2.1 with estimates for the covering number of $\mathcal{F}_d^{\mathbf{FNN}}(G,N,S,B,F)$ and the approximation error (2.56) to obtain an upper bound for $\boldsymbol{E}_{P^{\otimes n}}\left[\mathcal{R}_P^{\phi}\left(\hat{f}_n^{\mathbf{FNN}}\right) - \int \psi \mathrm{d}P\right]$, which, together with the noise condition (2.24), yields an upper bound for $\boldsymbol{E}_{P^{\otimes n}}\left[\mathcal{E}_P(\hat{f}_n^{\mathbf{FNN}})\right]$ (see (C.109)). Finally taking the supremum over all $P \in \mathcal{H}_{6,t_1,c_1,t_2,c_2}^{d,\beta,r,I,\Theta,s_1,s_2}$ gives the desired result. The proof of Theorem 2.5 along with that of Theorem 2.2 and Theorem 2.3 indicates that Theorem 2.1 is very flexible in the sense that it can be used in various settings with different choices of ψ.

2.2 Main Lower Bounds

In this subsection, we will give our main results on lower bounds for convergence rates of the logistic risk, which will justify the optimality of our upper bounds established in the last subsection. To state these results, we need some notations.

Recall that for any $a \in [0,1]$, $\mathscr{M}_a$ denotes the probability measure on $\{-1,1\}$ with $\mathscr{M}_a(\{1\}) = a$ and $\mathscr{M}_a(\{-1\}) = 1-a$. For any measurable $\eta : [0,1]^d \to [0,1]$ and any Borel probability measure $\mathscr{Q}$ on $[0,1]^d$, we denote

$$\begin{aligned} P_{\eta,\mathscr{Q}} : &\Big\{\text{Borel subsets of } [0,1]^d \times \{-1,1\}\Big\} \to [0,1], \\ &S \mapsto \int_{[0,1]^d} \int_{\{-1,1\}} \mathbb{1}_S(x,y) \mathrm{d}\mathscr{M}_{\eta(x)}(y) \mathrm{d}\mathscr{Q}(x). \end{aligned} \tag{2.57}$$

Therefore, $P_{\eta,\mathscr{Q}}$ is the (unique) probability measure on $[0,1]^d \times \{-1,1\}$ of which the marginal distribution on $[0,1]^d$ is $\mathscr{Q}$ and the conditional probability function is η. If $\mathscr{Q}$ is the Lebesgue measure on $[0,1]^d$, we will write P_η for $P_{\eta,\mathscr{Q}}$.

For any $\beta \in (0,\infty)$, $r \in (0,\infty)$, $A \in [0,1)$, $q \in \mathbb{N} \cup \{0\}$, and $(d, d_*, K) \in \mathbb{N}^3$ with $d_* \le \min\left\{d, K + \mathbb{1}_{\{0\}}(q) \cdot (d-K)\right\}$, define

$$\begin{aligned} \mathcal{H}_{3,A}^{d,\beta,r} &:= \left\{ P_\eta \,\middle|\, \begin{matrix} \eta \in \mathcal{B}_r^{\beta}([0,1]^d), \ \mathbf{ran}(\eta) \subset [0,1], \text{ and} \\ \int_{[0,1]^d} \mathbb{1}_{[0,A]}(|2\eta(x)-1|)\mathrm{d}x = 0 \end{matrix} \right\}, \\ \mathcal{H}_{5,A,q,K,d_*}^{d,\beta,r} &:= \left\{ P_\eta \,\middle|\, \begin{matrix} \eta \in \mathcal{G}_d^{\mathbf{CH}}(q,K,d_*,\beta,r), \ \mathbf{ran}(\eta) \subset [0,1], \\ \text{and } \int_{[0,1]^d} \mathbb{1}_{[0,A]}(|2\eta(x)-1|)\mathrm{d}x = 0 \end{matrix} \right\}. \end{aligned} \tag{2.58}$$

Now we can state our Theorem 2.6. Recall that $\mathcal{F}_d$ is the set of all measurable real-valued functions defined on $[0,1]^d$.

Theorem 2.6 *Let ϕ be the logistic loss, $n \in \mathbb{N}$, $\beta \in (0,\infty)$, $r \in (0,\infty)$, $A \in [0,1)$, $q \in \mathbb{N} \cup \{0\}$, and $(d, d_*, K) \in \mathbb{N}^3$ with $d_* \le \min\left\{d, K + \mathbb{1}_{\{0\}}(q) \cdot (d-K)\right\}$. Suppose $\{(X_i,Y_i)\}_{i=1}^n$ is a sample in $[0,1]^d \times \{-1,1\}$ of size n. Then there exists a constant $\mathrm{c}_0 \in (0,\infty)$ only depending on (d_*, β, r, q), such that*

$$\inf_{\hat{f}_n} \sup_{P \in \mathcal{H}_{5,A,q,K,d_*}^{d,\beta,r}} \boldsymbol{E}_{P^{\otimes n}}\left[\mathcal{E}_P^{\phi}(\hat{f}_n)\right] \ge \mathrm{c}_0 n^{-\frac{\beta \cdot (1\wedge\beta)^q}{d_* + \beta\cdot(1\wedge\beta)^q}} \text{ provided that } n > \left|\frac{7}{1-A}\right|^{\frac{d_*+\beta\cdot(1\wedge\beta)^q}{\beta\cdot(1\wedge\beta)^q}},$$

where the infimum is taken over all $\mathcal{F}_d$-valued statistics on $([0,1]^d \times \{-1,1\})^n$ from the sample $\{(X_i,Y_i)\}_{i=1}^n$.

Taking $q = 0$, $K = 1$, and $d_* = d$ in Theorem 2.6, we immediately obtain the following corollary:

Corollary 2.1 *Let ϕ be the logistic loss, $d \in \mathbb{N}$, $\beta \in (0, \infty)$, $r \in (0, \infty)$, $A \in [0, 1)$, and $n \in \mathbb{N}$. Suppose $\{(X_i, Y_i)\}_{i=1}^n$ is a sample in $[0,1]^d \times \{-1, 1\}$ of size n. Then there exists a constant $\mathrm{c}_0 \in (0, \infty)$ only depending on (d, β, r), such that*

$$\inf_{\hat{f}_n} \sup_{P \in \mathcal{H}_{3,A}^{d,\beta,r}} \boldsymbol{E}_{P^{\otimes n}} \left[\mathcal{E}_P^\phi(\hat{f}_n) \right] \geq \mathrm{c}_0 n^{-\frac{\beta}{d+\beta}} \text{ provided that } n > \left| \frac{7}{1-A} \right|^{\frac{d+\beta}{\beta}},$$

where the infimum is taken over all $\mathcal{F}_d$-valued statistics on $([0,1]^d \times \{-1,1\})^n$ from the sample $\{(X_i, Y_i)\}_{i=1}^n$.

Theorem 2.6, together with Corollary 2.1, is proved in Appendix C.6.

Obviously, $\mathcal{H}_{5,A,q,K,d_*}^{d,\beta,r} \subset \mathcal{H}_{4,q,K,d_\star,d_*}^{d,\beta,r}$. Therefore, it follows from Theorem 2.6 that

$$\inf_{\hat{f}_n} \sup_{P \in \mathcal{H}_{4,q,K,d_\star,d_*}^{d,\beta,r}} \boldsymbol{E}_{P^{\otimes n}} \left[\mathcal{E}_P^\phi(\hat{f}_n) \right] \geq \inf_{\hat{f}_n} \sup_{P \in \mathcal{H}_{5,A,q,K,d_*}^{d,\beta,r}} \boldsymbol{E}_{P^{\otimes n}} \left[\mathcal{E}_P^\phi(\hat{f}_n) \right] \gtrsim n^{-\frac{\beta \cdot (1 \wedge \beta)^q}{d_* + \beta \cdot (1 \wedge \beta)^q}}.$$

This justifies that the rate $\mathcal{O}(\left(\frac{(\log n)^5}{n}\right)^{\frac{\beta \cdot (1 \wedge \beta)^q}{d_* + \beta \cdot (1 \wedge \beta)^q}})$ in (2.36) is optimal (up to the logarithmic factor $(\log n)^{\frac{5\beta \cdot (1 \wedge \beta)^q}{d_* + \beta \cdot (1 \wedge \beta)^q}}$). Similarly, it follows from $\mathcal{H}_{3,A}^{d,\beta,r} \subset \mathcal{H}_1^{d,\beta,r}$ and Corollary 2.1 that

$$\inf_{\hat{f}_n} \sup_{P \in \mathcal{H}_1^{d,\beta,r}} \boldsymbol{E}_{P^{\otimes n}} \left[\mathcal{E}_P^\phi(\hat{f}_n) \right] \geq \inf_{\hat{f}_n} \sup_{P \in \mathcal{H}_{3,A}^{d,\beta,r}} \boldsymbol{E}_{P^{\otimes n}} \left[\mathcal{E}_P^\phi(\hat{f}_n) \right] \gtrsim n^{-\frac{\beta}{d+\beta}},$$

which justifies that the rate $\mathcal{O}(\left(\frac{(\log n)^5}{n}\right)^{\frac{\beta}{\beta+d}})$ in (2.17) is optimal (up to the logarithmic factor $(\log n)^{\frac{5\beta}{\beta+d}}$). Moreover, note that any probability P in $\mathcal{H}_{3,A}^{d,\beta,r}$ must satisfy the noise condition (2.24) provided that $s_1 \in [0, \infty]$, $t_1 \in (0, A]$, and $c_1 \in (0, \infty)$. In other words, for any $s_1 \in [0, \infty]$, $t_1 \in (0, A]$, and $c_1 \in (0, \infty)$, there holds $\mathcal{H}_{3,A}^{d,\beta,r} \subset \mathcal{H}_{2,s_1,c_1,t_1}^{d,\beta,r}$, meaning that

$$\begin{aligned} n^{-\frac{\beta}{d+\beta}} &\lesssim \inf_{\hat{f}_n} \sup_{P \in \mathcal{H}_{3,A}^{d,\beta,r}} \boldsymbol{E}_{P^{\otimes n}} \left[\mathcal{E}_P^\phi(\hat{f}_n) \right] \leq \inf_{\hat{f}_n} \sup_{P \in \mathcal{H}_{2,s_1,c_1,t_1}^{d,\beta,r}} \boldsymbol{E}_{P^{\otimes n}} \left[\mathcal{E}_P^\phi(\hat{f}_n) \right] \\ &\leq \inf_{\hat{f}_n} \sup_{P \in \mathcal{H}_1^{d,\beta,r}} \boldsymbol{E}_{P^{\otimes n}} \left[\mathcal{E}_P^\phi(\hat{f}_n) \right] \leq \sup_{P \in \mathcal{H}_1^{d,\beta,r}} \boldsymbol{E}_{P^{\otimes n}} \left[\mathcal{E}_P^\phi\left(\hat{f}_n^{\mathbf{FNN}}\right) \right] \lesssim \left(\frac{(\log n)^5}{n} \right)^{\frac{\beta}{\beta+d}}, \end{aligned}$$

where $\hat{f}_n^{\mathbf{FNN}}$ is the estimator defined in Theorem 2.2. From above inequalities we see that the noise condition (2.24) does little to help improve the convergence rate of the excess ϕ-risk in classification.

The proof of Theorem 2.6 and Corollary 2.1 is based on a general scheme for obtaining lower bounds, which is given in Section 2 of Tsybakov (2009). However, the scheme in Tsybakov (2009) is stated for a class of probabilities $\mathcal{H}$ that takes the form $\mathcal{H} = \{Q_\theta | \theta \in \Theta\}$

with Θ being some pseudometric space. In our setting, we do not have such pseudometric space. Instead, we introduce another quantity

$$\inf_{f\in\mathcal{F}_d}\left|\mathcal{E}^{\phi}_P(f)+\mathcal{E}^{\phi}_Q(f)\right| \tag{2.59}$$

to characterize the difference between any two probability measures P and Q (see (C.126)). Estimating lower bounds for the quantity defined in (2.59) plays a key role in our proof of Theorem 2.6 and Corollary 2.1.

3. Discussions on Related Work

In this section, we compare our results with some existing ones in the literature. We first compare Theorem 2.2 and Theorem 2.5 with related results about binary classification using fully connected DNNs and logistic loss in Kim et al. (2021) and Farrell et al. (2021) respectively. Then we compare our work with Ji et al. (2021), in which the authors carry out generalization analysis for estimators obtained from gradient descent algorithms.

Throughout this section, we will use ϕ to denote the logistic loss (i.e., $\phi(t)=\log(1+\mathrm{e}^{-t})$) and $\{(X_i,Y_i)\}_{i=1}^n$ to denote an i.i.d. sample in $[0,1]^d\times\{-1,1\}$. The symbols d, β, r, I, Θ, t_1, c_1, t_2, c_2 and c will denote arbitrary numbers in $\mathbb{N}$, $(0,\infty)$, $(0,\infty)$, $\mathbb{N}$, $\mathbb{N}$, $(0,\infty)$, $(0,\infty)$, $(0,\infty)$, $(0,\infty)$ and $[0,\infty)$, respectively. The symbol P will always denote some probability measure on $[0,1]^d\times\{-1,1\}$, regarded as the data distribution, and η will denote the corresponding conditional probability function $P(\{1\}\,|\cdot)$ of P.

Recall that $\mathcal{C}^{d,\beta,r,I,\Theta}$, defined in (2.46), is the space consisting of classifiers which are equal to $+1$ on the union of some disjoint regions with piecewise Hölder smooth boundary and -1 otherwise. In Theorem 4.1 of Kim et al. (2021), the authors conduct generalization analysis when the data distribution P satisfies the piecewise smooth decision boundary condition (2.50), the noise condition (2.24), and the margin condition (2.51) with $s_1=s_2=\infty$ for some $\mathtt{C}\in\mathcal{C}^{d,\beta,r,I,\Theta}$. They show that there exist constants G_0,N_0,S_0,B_0,F_0 not depending on the sample size n such that the ϕ-ERM

$$\hat{f}_n^{\mathbf{FNN}}\in\underset{f\in\mathcal{F}_d^{\mathbf{FNN}}(G_0,N_0,S_0,B_0,F_0)}{\arg\min}\frac{1}{n}\sum_{i=1}^n\phi\left(Y_if(X_i)\right)$$

satisfies

$$\sup_{P\in\mathcal{H}^{d,\beta,r,I,\Theta,\infty,\infty}_{6,t_1,c_1,t_2,c_2}}\boldsymbol{E}_{P^{\otimes n}}\left[\mathcal{E}_P\left(\hat{f}_n^{\mathbf{FNN}}\right)\right]\lesssim\frac{(\log n)^{1+\epsilon}}{n} \tag{3.1}$$

for any $\epsilon>0$. Indeed, the noise conditions (2.24) and the margin condition (2.51) with $s_1=s_2=\infty$ are equivalent to the following two conditions: there exist $\eta_0\in(0,1)$ and $\overline{\Delta}>0$ such that

$$P_X\left(\left\{x\in[0,1]^d\middle|\,|2\eta(x)-1|\le\eta_0\right\}\right)=0$$

and

$$P_X\left(\left\{x\in[0,1]^d\mid\Delta_{\mathtt{C}}(x)\le\overline{\Delta}\right\}\right)=0$$

(cf. conditions (N′) and (M′) in Kim et al. (2021)). Under the two conditions above, combining with the assumption $\operatorname{sgn}(2\eta-1)\xlongequal{P_X\text{-a.s.}}\mathtt{C}\in\mathcal{C}^{d,\beta,r,I,\Theta}$, Lemma A.7 of Kim et al.

(2021) asserts that there exists $f_0^* \in \mathcal{F}_d^{\mathbf{FNN}}(G_0, N_0, S_0, B_0, F_0)$ such that

$$f_0^* \in \underset{f \in \mathcal{F}_d^{\mathbf{FNN}}(G_0,N_0,S_0,B_0,F_0)}{\arg\min} \mathcal{R}_P^{\phi}(f)$$

and

$$\mathcal{R}_P(f_0^*) = \mathcal{R}_P(2\eta - 1) = \inf\left\{\mathcal{R}_P(f) \,\middle|\, f : [0,1]^d \to \mathbb{R} \text{ is measurable}\right\}.$$

The excess misclassification error of $f : [0,1]^d \to \mathbb{R}$ is then given by $\mathcal{E}_P(f) = \mathcal{R}_P(f) - \mathcal{R}_P(f_0^*)$. Since f_0^* is bounded by F_0, the authors in Kim et al. (2021) can apply classical concentration techniques developed for bounded random variables (cf. Appendix A.2 of Kim et al. (2021)) to deal with f_0^* (instead of the target function $f_{\phi,P}^*$), leading to the generalization bound (3.1). In this paper, employing Theorem 2.1, we extend Theorem 4.1 of Kim et al. (2021) to much less restrictive cases in which the noise exponent s_1 and the margin exponent s_2 are allowed to be taken from $[0, \infty]$. The derived generalization bounds are presented in Theorem 2.5. In particular, when $s_1 = s_2 = \infty$ (i.e., let $s_1 = \infty$ in statement (1) of Theorem 2.5), we obtain a refined generalization bound under the same conditions as those of Theorem 4.1 in Kim et al. (2021), which asserts that the ϕ-ERM $\hat{f}_n^{\mathbf{FNN}}$ over $\mathcal{F}_d^{\mathbf{FNN}}(G_0, N_0, S_0, B_0, F_0)$ satisfies

$$\sup_{P \in \mathcal{H}_{6,t_1,c_1,t_2,c_2}^{d,\beta,r,I,\Theta,\infty,\infty}} \boldsymbol{E}_{P^{\otimes n}}\left[\mathcal{E}_P\left(\hat{f}_n^{\mathbf{FNN}}\right)\right] \lesssim \frac{\log n}{n}, \tag{3.2}$$

removing the ϵ in their bound (3.1). The above discussion indicates that Theorem 2.1 can lead to sharper estimates in comparison with classical concentration techniques, and can be applied in very general settings. However, we would like to point out that if $s_1 < \infty$ and $s_2 < \infty$, then the convergence rate obtained in Theorem 2.5 (that is, the rate $\mathcal{O}\Big(\left(\frac{(\log n)^3}{n}\right)^{\frac{s_1}{1+(s_1+1)\left(1+\frac{d-1}{\beta s_2}\right)}}\Big)$ in (2.55)) is suboptimal. Indeed, Theorem 3.1 and Theorem 3.4 of Kim et al. (2021) show that DNN classifier $\hat{f}_n^{\mathbf{FNN}}$ trained with empirical hinge risk minimization can achieve a convergence rate

$$\sup_{P \in \mathcal{H}_{6,t_1,c_1,t_2,c_2}^{d,\beta,r,I,\Theta,s_1,s_2}} \boldsymbol{E}_{P^{\otimes n}}\left[\mathcal{E}_P\left(\hat{f}_n^{\mathbf{FNN}}\right)\right] \lesssim \left(\frac{(\log n)^3}{n}\right)^{\frac{s_1+1}{1+(s_1+1)\left(1+\frac{d-1}{\beta\cdot(1\vee s_2)}\right)}}, \tag{3.3}$$

which is strictly faster than the rate $\mathcal{O}\Big(\left(\frac{(\log n)^3}{n}\right)^{\frac{s_1}{1+(s_1+1)\left(1+\frac{d-1}{\beta s_2}\right)}}\Big)$ in (2.55). Moreover, as mentioned below Theorem 3.1 in Kim et al. (2021), even the rate in (3.3) is suboptimal in general. In Hu et al. (2022a), the authors propose a new DNN classifier which are constructed in a divide-and-conquer manner: DNN classifiers are trained with empirical 0-1 risk minimization on each local region and then "aggregated to a global one". Hu et al. (2022a) provides minimax optimal convergence rates for this new DNN classifier under the assumption that the data distribution $P \in H_{6,t_1,1,t_2,1}^{d,\beta,r,1,1,0,0}$ (that is, the decision boundary of P is assumed to be Hölder-β smooth (rather than just piecewise smooth), but the noise condition (2.24) and the margin condition (2.51) are not required) along with a "localized version" of the noise condition (2.24) (see assumptions (M1) and (M2) in Hu et al. (2022a)). It is interesting to further study whether we can apply Theorem 2.1 to establish optimal convergence rates for the new DNN classifiers proposed in Hu et al. (2022a) which are locally trained with some surrogate loss (as we have already pointed out, Theorem 2.1 remains true for any locally Lipschitz continuous loss function ϕ, see the discussion on page 14) such as logistic loss instead of 0-1 loss.

The recent work Farrell et al. (2021) considers estimation and inference using fully connected DNNs and the logistic loss in which their setting can cover both regression and classification. For any probability measure P on $[0,1]^d \times \{-1,1\}$ and any measurable function $f : [0,1]^d \to [-\infty, \infty]$, define $\|f\|_{\mathcal{L}^2_{P_X}} := \left(\int_{[0,1]^d} |f(x)|^2 \,\mathrm{d}P_X(x)\right)^{\frac{1}{2}}$. Recall that $\mathcal{B}^\beta_r(\Omega)$ is defined in (2.13). Let $\mathcal{H}^{d,\beta}_7$ be the set of all probability measures P on $[0,1]^d \times \{-1,1\}$ such that the target function $f^*_{\phi,P}$ belongs to $\mathcal{B}^\beta_1\left([0,1]^d\right)$. In Corollary 1 of Farrell et al. (2021), the authors claimed that if $P \in \mathcal{H}^{d,\beta}_7$ and $\beta \in \mathbb{N}$, then with probability at least $1 - \mathrm{e}^{-\upsilon}$ there holds

$$\left\|\hat{f}^{\mathbf{FNN}}_n - f^*_{\phi,P}\right\|^2_{\mathcal{L}^2_{P_X}} \lesssim n^{-\frac{2\beta}{2\beta+d}} \log^4 n + \frac{\log\log n + \upsilon}{n}, \tag{3.4}$$

where the estimator $\hat{f}^{\mathbf{FNN}}_n \in \mathcal{F}^{\mathbf{FNN}}_d(G, N, S, \infty, F)$ is defined by (2.14) with

$$G \asymp \log n,\ \ N \asymp n^{\frac{d}{d+2\beta}},\ \ S \asymp n^{\frac{d}{d+2\beta}} \log n,\ \text{ and }\ F = 2. \tag{3.5}$$

Note that $f^*_{\phi,P} \in \mathcal{B}^\beta_1\left([0,1]^d\right)$ implies $\|f^*_{\phi,P}\|_\infty \le 1$. From Lemma 8 of Farrell et al. (2021), bounding the quantity $\left\|\hat{f}^{\mathbf{FNN}}_n - f^*_{\phi,P}\right\|^2_{\mathcal{L}^2_{P_X}}$ on the left hand side of (3.4) is equivalent to bounding $\mathcal{E}^\phi_P(\hat{f}^{\mathbf{FNN}}_n)$, since

$$\frac{1}{2(\mathrm{e} + \mathrm{e}^{-1} + 2)} \left\|\hat{f}^{\mathbf{FNN}}_n - f^*_{\phi,P}\right\|^2_{\mathcal{L}^2_{P_X}} \le \mathcal{E}^\phi_P(\hat{f}^{\mathbf{FNN}}_n) \le \frac{1}{4} \left\|\hat{f}^{\mathbf{FNN}}_n - f^*_{\phi,P}\right\|^2_{\mathcal{L}^2_{P_X}}. \tag{3.6}$$

Hence (3.4) actually establishes the same upper bound (up to a constant independent of n and P) for the excess ϕ-risk of $\hat{f}^{\mathbf{FNN}}_n$, leading to upper bounds for the excess misclassification error $\mathcal{E}_P(\hat{f}^{\mathbf{FNN}}_n)$ through the calibration inequality. The authors in Farrell et al. (2021) apply concentration techniques based on (empirical) *Rademacher complexity* (cf. Section A.2 of Farrell et al. (2021) or Bartlett et al. (2005); Koltchinskii (2006)) to derive the bound (3.4), which allows for removing the restriction of uniformly boundedness on the weights and biases in the neural network models, i.e., the hypothesis space generated by neural networks in their analysis can be of the form $\mathcal{F}^{\mathbf{FNN}}_d(G, N, S, \infty, F)$. In our paper, we employ the covering number to measure the complexity of hypothesis space. Due to the lack of compactness, the covering numbers of $\mathcal{F}^{\mathbf{FNN}}_d(G, N, S, \infty, F)$ are in general equal to infinity. Consequently, in our convergence analysis, we require the neural networks to possess bounded weights and biases. The assumption of bounded parameters may lead to additional optimization constraints in the training process. However, it has been found that the weights and biases of a trained neural network are typically around their initial values (cf. Goodfellow et al. (2016)). Thus the boundedness assumption matches what is observed in practice and has been adopted by most of the literature (see, e,g., Kim et al. (2021); Schmidt-Hieber (2020)). In particular, the work Schmidt-Hieber (2020) considers nonparametric regression using neural networks with all parameters bounded by one (i.e., $B = 1$). This assumption can be realized by projecting the parameters of the neural network onto $[-1, 1]$ after each updating. Though the framework developed in this paper would not deliver generalization bounds without restriction of uniformly bounded parameters, we weaken this constraint in Theorem 2.2 by allowing the upper bound B to grow polynomially with the sample size n, which simply requires $1 \leq B \lesssim n^\nu$ for any $\nu > 0$. It is worth mentioning that in our coming work Zhang et al. (2024), we actually establish oracle-type inequalities analogous to Theorem 2.1, with the covering number $\mathcal{N}(\mathcal{F}, \gamma)$ replaced by the supremum of some empirical L_1-covering numbers. These enable us to derive generalization bounds for the empirical ϕ-risk minimizer $\hat{f}_n^{\mathbf{FNN}}$ over $\mathcal{F}_d^{\mathbf{FNN}}(G, N, S, \infty, F)$ because empirical L_1-covering numbers of $\mathcal{F}_d^{\mathbf{FNN}}(G, N, S, \infty, F)$ can be well-controlled, as indicated by Lemma 4 and Lemma 6 of Farrell et al. (2021) (see also Theorem 9.4 of Györfi et al. (2002) and Theorem 7 of Bartlett et al. (2019)). In addition, note that (3.4) can lead to probability bounds (i.e., confidence bounds) for the excess ϕ-risk and misclassification error of $\hat{f}_n^{\mathbf{FNN}}$, while the generalization bounds presented in this paper are only in expectation. Nonetheless, in Zhang et al. (2024), we obtain both probability bounds and expectation bounds for the empirical ϕ-risk minimizer.

As discussed in Section 1, the boundedness assumptions on the target function $f^*_{\phi,P}$ and its derivatives, i.e., $f^*_{\phi,P} \in \mathcal{B}_1^\beta\left([0,1]^d\right)$, are too restrictive. This assumption actually requires that there exists some $\delta \in (0, 1/2)$ such that the conditional class probability $\eta(x) = P(\{1\}|x)$ satisfies $\delta < \eta(x) < 1 - \delta$ for P_X-almost all $x \in [0,1]^d$, which rules out the case when η takes values in 0 or 1 with positive probabilities. However, it is believed that the conditional class probability should be determined by the patterns that make the two classes mutually exclusive, implying that $\eta(x)$ should be closed to either 0 or 1. This is also observed in many benchmark datasets for image recognition. For example, it is reported in Kim et al. (2021), the conditional class probabilities of CIFAR10 data set estimated by neural networks with the logistic loss almost solely concentrate on 0 or 1 and very few are around 0.5 (see Fig.2 in Kim et al. (2021)). Overall, the boundedness restriction on $f^*_{\phi,P}$ is not expected to hold in binary classification as it would exclude the well classified data. We further point out that the techniques used in Farrell et al. (2021) cannot deal with the case when $f^*_{\phi,P}$ is unbounded, or equivalently, when η can take values close to 0 or 1. Indeed, the authors apply approximation theory of neural networks developed in Yarotsky (2017) to construct uniform approximations of $f^*_{\phi,P}$, which requires $f^*_{\phi,P} \in \mathcal{B}_1^\beta\left([0,1]^d\right)$ with $\beta \in \mathbb{N}$. However, if $f^*_{\phi,P}$ is unbounded, uniformly approximating $f^*_{\phi,P}$ by neural networks on $[0,1]^d$ is impossible, which brings the essential difficulty in estimating the approximation error. Besides, the authors use Bernstein's inequality to bound the quantity $\frac{1}{n}\sum_{i=1}^n \Big(\phi(Y_i f_1^*(X_i)) - \phi(Y_i f^*_{\phi,P}(X_i))\Big)$ appearing in the error decomposition for $\left\|\hat{f}_n^{\mathbf{FNN}} - f^*_{\phi,P}\right\|^2_{\mathcal{L}^2_{P_X}}$ (see (A.1) in Farrell et al. (2021)), where $f_1^* \in \arg\min_{f \in \mathcal{F}_d^{\mathbf{FNN}}(G,N,S,\infty,2)} \|f - f^*_{\phi,P}\|_{[0,1]^d}$. We can see that the unboundedness of $f^*_{\phi,P}$ will lead to the unboundedness of the random variable $\Big(\phi(Y f_1^*(X)) - \phi(Y f^*_{\phi,P}(X))\Big)$, which makes Bernstein's inequality invalid to bound its empirical mean by the expectation.

In addition, the boundedness assumption on $f^*_{\phi,P}$ ensures the inequality (3.6) on which the entire framework of convergence estimates in Farrell et al. (2021) is built (cf. Appendix A.1 and A.2 of Farrell et al. (2021)). Without this assumption, most of the theoretical arguments in Farrell et al. (2021) are not feasible. In contrast, we require $\eta \xlongequal{P_X\text{-a.s.}} \hat{\eta}$ for some $\hat{\eta} \in \mathcal{B}^\beta_r\left([0,1]^d\right)$ and $r \in (0,\infty)$ in Theorem 2.2. This Hölder smoothness condition on η is well adopted in the study of binary classifiers (see Audibert and Tsybakov (2007) and references therein). Note that $f^*_{\phi,P} \in \mathcal{B}^\beta_1\left([0,1]^d\right)$ indeed implies $\eta \xlongequal{P_X\text{-a.s.}} \hat{\eta}$ for some $\hat{\eta} \in \mathcal{B}^\beta_r\left([0,1]^d\right)$ and $r \in (0,\infty)$ which only depends on (d,β). Therefore, the setting considered in Theorem 2.2 is more general than that of Farrell et al. (2021). Moreover, the condition $\eta \xlongequal{P_X\text{-a.s.}} \hat{\eta} \in \mathcal{B}^\beta_r\left([0,1]^d\right)$ is more nature, allowing η to take values close to 0 and 1 with positive probabilities. We finally point out that, under the same assumption (i.e., $P \in \mathcal{H}^{d,\beta}_7$), one can use Theorem 2.1 to establish a convergence rate which is slightly improved compared with (3.4). Actually, we can show that there exists a constant $\mathrm{c} \in (0,\infty)$ only depending on (d,β), such that for any $\mu \in [1,\infty)$, and $\nu \in [0,\infty)$, there holds

$$\sup_{P\in\mathcal{H}^{d,\beta}_7} \boldsymbol{E}_{P^{\otimes n}}\left[\left\|\hat{f}^{\mathbf{FNN}}_n - f^*_{\phi,P}\right\|^2_{\mathcal{L}^2_{P_X}}\right] \lesssim \left(\frac{(\log n)^3}{n}\right)^{\frac{2\beta}{2\beta+d}}, \tag{3.7}$$

where the estimator $\hat{f}^{\mathbf{FNN}}_n \in \mathcal{F}^{\mathbf{FNN}}_d(G,N,S,B,F)$ is defined by (2.14) with

$$\begin{aligned}&\mathrm{c}\log n \le G \asymp \log n,\ N \asymp \left(\frac{n}{\log^3 n}\right)^{\frac{d}{d+2\beta}},\ S \asymp \left(\frac{n}{\log^3 n}\right)^{\frac{d}{d+2\beta}} \cdot \log n,\\ &1 \le B \lesssim n^\nu,\ \text{and } 1 \le F \le \mu.\end{aligned} \tag{3.8}$$

Though we restrict the weights and biases to be bounded by B, both the convergence rate and the network complexities in the result above refine the previous estimates established in (3.4) and (3.5). In particular, since $\frac{6\beta}{2\beta+d} < 3 < 4$, the convergence rate in (3.7) is indeed faster than that in (3.4) due to a smaller power exponent of the term $\log n$. The proof of this claim is in Appendix C.7. We also remark that the convergence rate in (3.7) achieves the minimax optimal rate established in Stone (1982) up to log factors (so does the rate in (3.4)), which confirms that generalization analysis developed in this paper is also rate-optimal for bounded $f^*_{\phi,P}$.

In our work, we have established generalization bounds for ERMs over hypothesis spaces consisting of neural networks. However, such ERMs cannot be obtained in practice because the correspoding optimization problems (e.g., (2.2)) cannot be solved explicitly. Instead, practical neural network estimators are obtained from algorithms which numerically solve the empirical risk minimization problem. Therefore, it is better to conduct generalization analysis for estimators obtained from such algorithms. One typical work in this direction is Ji et al. (2021).

In Ji et al. (2021), for classification tasks, the authors establish excess ϕ-risk bounds to show that classifiers obtained from solving empirical risk minimization with respect to the logistic loss over shallow neural networks using gradient descent with (or without) early stopping are consistent. Note that the setting of Ji et al. (2021) is quite different from ours: We consider deep neural network models in our work, while Ji et al. (2021) considers shallow ones. Besides, we use the smoothness of the conditional probability function $\eta(\cdot) = P(\{1\}\,|\cdot)$ to characterize the regularity (or complexity) of the data distribution P. Instead, in Ji et al. (2021), for each $\overline{U}_\infty : \mathbb{R}^d \to \mathbb{R}^d$, the authors construct a function

$$f(\ \cdot\ ; \overline{U}_\infty) : \mathbb{R}^d \to \mathbb{R}, x \mapsto \int_{\mathbb{R}^d} x^\top \overline{U}_\infty(v) \cdot \mathbb{1}_{[0,\infty)}(v^\top x) \cdot \frac{1}{(2\pi)^{n/2}} \cdot \exp(-\frac{\|v\|_2^2}{2})\mathrm{d}v$$

called infinite-width random feature model. Then they use the norm of $\overline{U}_\infty$ which makes $\mathcal{E}_P^\phi(f(\ \cdot\ ; \overline{U}_\infty))$ small to characterize the regularity of data: the data distribution is regarded as simple if there is a $\overline{U}_\infty$ with $\mathcal{E}_P^\phi(f(\ \cdot\ ; \overline{U}_\infty)) \approx 0$ and moreover has a low norm. More rigorously, the slower the quantity

$$\inf\left\{ \left\|\overline{U}_\infty\right\|_{\mathbb{R}^d} \middle| \mathcal{E}_P^\phi(f(\ \cdot\ ; \overline{U}_\infty)) \leq \varepsilon \right\} \tag{3.9}$$

grows as $\varepsilon \to 0$, the more regular (simpler) the data distribution P is. In Ji et al. (2021), the established excess ϕ-risk bounds depend on the quantity $\mathcal{E}_P^\phi(f(\ \cdot\ ; \overline{U}_\infty))$ and the norm $\left\|\overline{U}_\infty\right\|_{\mathbb{R}^d}$. Hence by assuming certain growth rates of the quantity in (3.9) as $\varepsilon \to 0$, we can obtain specific rates of convergence from the excess ϕ-risk bounds in Ji et al. (2021). It is natural to ask is there any relation between these two characterizations of data regularity, that is, the smoothness of conditional probability function, and the rate of growth of the quantity in (3.9) as $\varepsilon \to 0$. For example, will Hölder smoothness of the conditional probability function imply certain growth rates of the quantity in (3.9) as $\varepsilon \to 0$? This question is worth considering because once we prove the equivalence of these two characterizations, then the generalization analysis in Ji et al. (2021) will be able to be used in other settings requiring smoothness of the conditional probability function and vice versa. In addition, it is also interesting to study how can we use our new techniques developed in this paper to establish generalization bounds for deep neural network estimators obtained from learning algorithms (e.g., gradient descent) within the settings in this paper.

4. Conclusion

In this paper, we develop a novel generalization analysis for binary classification with DNNs and logistic loss. The unboundedness of the target function in logistic classification poses challenges for the estimates of sample error and approximation error when deriving generalization bounds. To overcome these difficulties, we introduce a bivariate function $\psi : [0,1]^d \times \{-1,1\} \to \mathbb{R}$ to establish an elegant oracle-type inequality, aiming to bound the excess risk with respect to the logistic loss. This inequality incorporates the estimation of sample error and enables us to propose a framework for generalization analysis, which avoids using the explicit form of the target function. By properly choosing ψ under this framework, we can eliminate the boundedness restriction of the target function and establish sharp rates of convergence. In particular, for fully connected DNN classifiers trained by minimizing the empirical logistic risk, we obtain an optimal (up to some logarithmic factor) rate of convergence of the excess logistic risk (which further yields a rate of convergence of the excess misclassification error via the calibration inequality) merely under the Hölder smoothness assumption on the conditional probability function. If we instead assume that the conditional probability function is the composition of several vector-valued multivariate functions of which each component function is either a maximum value function of some of its input variables or a Hölder smooth function only depending on a small number of its input variables, we can even establish dimension-free optimal (up to some logarithmic factor) convergence rates for the excess logistic risk of fully connected DNN classifiers, further leading to dimension-free rates of convergence of their excess misclassification error through the calibration inequality. This result serves to elucidate the remarkable achievements of DNNs in high-dimensional real-world classification tasks. In other circumstances such as when the data distribution has a piecewise smooth decision boundary and the input data are bounded away from it (i.e., $s_2 = \infty$ in (2.51)), dimension-free rates of convergence can also be derived. Besides the novel oracle-type inequality, the sharp estimates presented in our paper also owe to a tight error bound for approximating the natural logarithm function (which is unbounded near zero) by fully connected DNNs. All the claims for the optimality of rates in our paper are justified by corresponding minimax lower bounds. As far as we know, all these results are new to the literature, which further enrich the theoretical understanding of classification using deep neural networks. At last, we would like to emphasize that our framework of generalization analysis is very general and can be extended to many other settings (e.g., when the loss function, the hypothesis space, or the assumption on the data distribution is different from that in this current paper). In particular, in our forthcoming research Zhang et al. (2024), we have investigated generalization analysis for CNN classifiers trained with the logistic loss, exponential loss, or LUM loss on spheres under the Sobolev smooth conditional probability assumption. Motivated by recent work Guo et al. (2020, 2017); Lin and Zhou (2018); Zhou (2018), we will also study more efficient implementations of deep logistic classification for dealing with big data.

Acknowledgments

The work described in this paper is supported partially by InnoHK initiative, the Government of the HKSAR, Laboratory for AI-Powered Financial Technologies, the Research Grants Council of Hong Kong (Projects No. CityU 11308121, 11306220, 11308020), the Germany/Hong Kong Joint Research Scheme (Project No. G-CityU101/20), the NSFC/RGC Joint Research Scheme (Project No. 12061160462 and N_CityU102/20). Lei Shi is also supported by Shanghai Science and Technology Program (Project No. 21JC1400600). The first version of the paper was written when Ding-Xuan Zhou worked at City University of Hong Kong.

Appendix A. Covering Numbers of Spaces of Fully Connected DNNs

In this appendix, we provide upper bounds for the covering numbers of spaces of fully connected DNNs. Recall that if $\mathcal{F}$ consists of bounded real-valued functions defined on a domain containing $[0,1]^d$, the covering number of $\mathcal{F}$ with respect to the radius γ and the metric $\mathcal{F}\times\mathcal{F}\ni(f,g)\mapsto\sup_{x\in[0,1]^d}|f(x)-g(x)|\in[0,\infty)$ is denoted by $\mathcal{N}(\mathcal{F},\gamma)$. For the space $\mathcal{F}_d^{\mathbf{FNN}}(G,N,S,B,F)$ defined by (1.15), the covering number $\mathcal{N}\left(\mathcal{F}_d^{\mathbf{FNN}}(G,N,S,B,F),\gamma\right)$ can be bounded from above in terms of G,N,S,B, and the radius of covering γ. The related results are stated below.

Theorem A.1 *For $G \in [1, \infty)$, $(N, S, B) \in [0, \infty)^3$, and $\gamma \in (0, 1)$, there holds*

$$\begin{aligned}&\log\Big(\mathcal{N}\left(\mathcal{F}_d^{\mathbf{FNN}}(G, N, S, B, \infty), \gamma\right)\Big)\\&\leq (S + Gd + 1)(2G + 5) \cdot \log \frac{(\max\{N, d\} + 1)(B \vee 1)(G + 1)}{\gamma}.\end{aligned}$$

Theorem A.1 can be proved in the same manner as in the proof of Lemma 5 in Schmidt-Hieber (2020). Therefore, we omit the proof here. Similar results are also presented in Proposition A.1 of Kim et al. (2021) and Lemma 3 of Suzuki (2019). Corollary A.1 follows immediately from Theorem A.1 and Lemma 10.6 of Anthony and Bartlett (2009).

Corollary A.1 *For $G \in [1, \infty)$, $(N, S, B) \in [0, \infty)^3$, $F \in [0, \infty]$ and $\gamma \in (0, 1)$, there holds*

$$\begin{aligned}&\log\Big(\mathcal{N}\left(\mathcal{F}_d^{\mathbf{FNN}}(G, N, S, B, F), \gamma\right)\Big)\\&\leq (S + Gd + 1)(2G + 5) \cdot \log \frac{(\max\{N, d\} + 1)(B \vee 1)(2G + 2)}{\gamma}.\end{aligned}$$

Appendix B. Approximation Theory of Fully Connected DNNs

Theorem B.1 below gives error bounds for approximating Hölder continuous functions by fully connected DNNs. Since it can be derived straightforwardly from Theorem 5 of Schmidt-Hieber (2020), we omit its proof.

Theorem B.1 *Suppose that $f \in \mathcal{B}_r^\beta\left([0,1]^d\right)$ with some $(\beta, r) \in (0, \infty)^2$. Then for any positive integers m and M' with $M' \geq \max\left\{(\beta+1)^d, \left(r\sqrt{d}\lceil\beta\rceil^d + 1\right)\mathrm{e}^d\right\}$, there exists*

$$\tilde{f} \in \mathcal{F}_d^{\mathbf{FNN}}\Big(14m(2 + \log_2(d \vee \beta)), 6\left(d + \lceil\beta\rceil\right)M', 987(2d+\beta)^{4d}M'm, 1, \infty\Big)$$

such that

$$\begin{aligned}&\sup_{x\in[0,1]^d}\left|f(x) - \tilde{f}(x)\right|\\&\leq r\sqrt{d}\lceil\beta\rceil^d \cdot 3^\beta M'^{-\beta/d} + \left(1 + 2r\sqrt{d}\lceil\beta\rceil^d\right) \cdot 6^d \cdot (1 + d^2 + \beta^2) \cdot M' \cdot 2^{-m}.\end{aligned}$$

Corollary B.1 follows directly from Theorem B.1.

Corollary B.1 *Suppose that $f \in \mathcal{B}_r^\beta\left([0,1]^d\right)$ with some $(\beta, r) \in (0, \infty)^2$. Then for any $\varepsilon \in (0, 1/2]$, there exists*

$$\tilde{f} \in \mathcal{F}_d^{\mathbf{FNN}}\left(D_1 \log\frac{1}{\varepsilon}, D_2\varepsilon^{-\frac{d}{\beta}}, D_3\varepsilon^{-\frac{d}{\beta}}\log\frac{1}{\varepsilon}, 1, \infty\right)$$

such that

$$\sup_{x\in[0,1]^d}\left|f(x) - \tilde{f}(x)\right| \leq \varepsilon,$$

where $(D_1, D_2, D_3) \in (0, \infty)^3$ are constants depending only on d, β and r.

Proof Let

$$E_1 = \max\left\{(\beta+1)^d, \left(r\sqrt{d}\lceil\beta\rceil^d+1\right)\mathrm{e}^d, \left(\frac{1}{2r}\cdot 3^{-\beta}\cdot\frac{1}{\sqrt{d}\lceil\beta\rceil^d}\right)^{-d/\beta}\right\},$$
$$E_2 = 3\max\left\{1+\frac{d}{\beta}, \frac{\log\left(4E_1\cdot\left(1+2r\sqrt{d}\lceil\beta\rceil^d\right)(1+d^2+\beta^2)\cdot 6^d\right)}{\log 2}\right\},$$

and

$$D_1 = 14\cdot(2+\log_2(d\vee\beta))\cdot(E_2+2),$$
$$D_2 = 6\cdot(d+\lceil\beta\rceil)\cdot(E_1+1),$$
$$D_3 = 987\cdot(2d+\beta)^{4d}\cdot(E_1+1)\cdot(E_2+2).$$

Then D_1, D_2, D_3 are constants only depending on d, β, r.

For $f\in\mathcal{B}_r^\beta\left([0,1]^d\right)$ and $\varepsilon\in(0,1/2]$, choose $M' = \left\lceil E_1\cdot\varepsilon^{-d/\beta}\right\rceil$ and $m = \lceil E_2\log(1/\varepsilon)\rceil$. Then m and M' are positive integers satisfying that

$$\begin{aligned}&1\le\max\left\{(\beta+1)^d, \left(r\sqrt{d}\lceil\beta\rceil^d+1\right)\mathrm{e}^d\right\}\le E_1\le E_1\cdot\varepsilon^{-d/\beta}\\&\le M'\le 1+E_1\cdot\varepsilon^{-d/\beta}\le(E_1+1)\cdot\varepsilon^{-d/\beta},\end{aligned}\tag{B.1}$$

$$M'^{-\beta/d}\le\left(E_1\cdot\varepsilon^{-d/\beta}\right)^{-\beta/d}\le\varepsilon\cdot\frac{1}{2r}\cdot 3^{-\beta}\cdot\frac{1}{\sqrt{d}\lceil\beta\rceil^d},\tag{B.2}$$

and

$$m\le E_2\log(1/\varepsilon)+2\log 2\le E_2\log(1/\varepsilon)+2\log(1/\varepsilon) = (2+E_2)\cdot\log(1/\varepsilon).\tag{B.3}$$

Moreover, we have that

$$\begin{aligned}&2\cdot\left(1+2r\sqrt{d}\lceil\beta\rceil^d\right)\cdot 6^d\cdot(1+d^2+\beta^2)\cdot M'\cdot\frac{1}{\varepsilon}\\&\le 2\cdot\left(1+2r\sqrt{d}\lceil\beta\rceil^d\right)\cdot 6^d\cdot(1+d^2+\beta^2)\cdot(E_1+1)\cdot\varepsilon^{-1-d/\beta}\\&\le 2\cdot\left(1+2r\sqrt{d}\lceil\beta\rceil^d\right)\cdot 6^d\cdot(1+d^2+\beta^2)\cdot 2E_1\cdot\varepsilon^{-1-d/\beta}\\&\le 2^{\frac{1}{3}E_2}\cdot\varepsilon^{-1-d/\beta}\le 2^{\frac{1}{3}E_2}\cdot\varepsilon^{-\frac{1}{3}E_2}\le\varepsilon^{-\frac{1}{3}E_2}\cdot\varepsilon^{-\frac{1}{3}E_2}\\&\le\varepsilon^{-E_2\cdot\log 2} = 2^{E_2\log(1/\varepsilon)}\le 2^m.\end{aligned}\tag{B.4}$$

Therefore, from (B.1), (B.2), (B.3), (B.4), and Theorem B.1, we conclude that there exists

$$\begin{aligned}&\tilde{f}\in\mathcal{F}_d^{\mathbf{FNN}}(14m(2+\log_2(d\vee\beta)), 6(d+\lceil\beta\rceil)M', 987(2d+\beta)^{4d}M'm, 1, \infty)\\&=\mathcal{F}_d^{\mathbf{FNN}}\left(\frac{D_1}{E_2+2}\cdot m, \frac{D_2}{E_1+1}\cdot M', \frac{D_3}{(E_1+1)\cdot(E_2+2)}\cdot M'm, 1, \infty\right)\\&\subset\mathcal{F}_d^{\mathbf{FNN}}\left(D_1\log\frac{1}{\varepsilon}, D_2\varepsilon^{-\frac{d}{\beta}}, D_3\varepsilon^{-\frac{d}{\beta}}\log\frac{1}{\varepsilon}, 1, \infty\right)\end{aligned}$$

such that

$$\sup_{x\in[0,1]^d}\left|f(x)-\tilde{f}(x)\right|$$
$$\leq r\sqrt{d}\,\lceil\beta\rceil^d\cdot 3^\beta M'^{-\beta/d}+\left(1+2r\sqrt{d}\,\lceil\beta\rceil^d\right)\cdot 6^d\cdot(1+d^2+\beta^2)\cdot M'\cdot 2^{-m}\leq\frac{\varepsilon}{2}+\frac{\varepsilon}{2}=\varepsilon.$$

Thus we complete the proof. ■

Appendix C. Proofs of Results in the Main Body

The proofs in this appendix will be organized in logical order in the sense that each result in this appendix is proved without relying on results that are presented after it.

Throughout this appendix, we use

$$C_{\mathtt{Parameter}_1,\mathtt{Parameter}_2,\cdots,\mathtt{Parameter}_m}$$

to denote a positive constant only depending on $\mathtt{Parameter}_1$, $\mathtt{Parameter}_2$, $\cdots$, $\mathtt{Parameter}_m$. For example, we may use $C_{d,\beta}$ to denote a positive constant only depending on (d,β). The values of such constants appearing in the proofs may be different from line to line or even in the same line. Besides, we may use the same symbol with different meanings in different proofs. For example, the symbol I may denote a number in one proof, and denote a set in another proof. To avoid confusion, we will explicitly redefine these symbols in each proof.

C.1 Proofs of Some Properties of the Target Function

The following lemma justifies our claim in (1.7).

Lemma C.1 *Let $d\in\mathbb{N}$, P be a probability measure on $[0,1]^d\times\{-1,1\}$, and $\phi:\mathbb{R}\to[0,\infty)$ be a measurable function. Define*

$$\overline{\phi}:[-\infty,\infty]\to[0,\infty],\ z\mapsto\begin{cases}\overline{\lim\limits_{t\to+\infty}}\,\phi(t), & \text{if } z=\infty,\\ \phi(z), & \text{if } z\in\mathbb{R},\\ \overline{\lim\limits_{t\to-\infty}}\,\phi(t), & \text{if } z=-\infty,\end{cases}$$

which is an extension of ϕ to $[-\infty,\infty]$. Suppose $f^*:[0,1]^d\to[-\infty,\infty]$ is a measurable function satisfying that

$$f^*(x)\in\mathop{\arg\min}_{z\in[-\infty,\infty]}\int_{\{-1,1\}}\overline{\phi}(yz)\mathrm{d}P(y|x)\text{ for } P_X\text{-almost all } x\in[0,1]^d. \tag{C.1}$$

Then there holds

$$\int_{[0,1]^d\times\{-1,1\}}\overline{\phi}(yf^*(x))\mathrm{d}P(x,y)=\inf\left\{\mathcal{R}_P^\phi(g)\,\middle|\,g:[0,1]^d\to\mathbb{R}\text{ is measurable}\right\}.$$

Proof Let $\Omega_0 := \left\{ x \in [0,1]^d \middle| f^*(x) \in \mathbb{R} \right\} \times \{-1,1\}$. Then for any $m \in \mathbb{N}$ and any $(i,j) \in \{-1,1\}^2$, define

$$f_m : [0,1]^d \to \mathbb{R},\ x \mapsto \begin{cases} m, & \text{if } f^*(x) = \infty, \\ f^*(x), & \text{if } f^*(x) \in \mathbb{R}, \\ -m, & \text{if } f^*(x) = -\infty, \end{cases}$$

and $\Omega_{i,j} = \left\{ x \in [0,1]^d \middle| f^*(x) = i \cdot \infty \right\} \times \{j\}$. Obviously, $yf^*(x) = ij \cdot \infty$ and $yf_m(x) = ijm$ for any $(i,j) \in \{-1,1\}^2$, any $m \in \mathbb{N}$, and any $(x,y) \in \Omega_{i,j}$. Therefore,

$$\begin{aligned} &\varlimsup_{m \to +\infty} \int_{\Omega_{i,j}} \phi(yf_m(x)) \mathrm{d}P(x,y) \\ &= \varlimsup_{m \to +\infty} \int_{\Omega_{i,j}} \phi(ijm) \mathrm{d}P(x,y) = P(\Omega_{i,j}) \cdot \varlimsup_{m \to +\infty} \phi(ijm) \\ &\leq P(\Omega_{i,j}) \cdot \varlimsup_{t \to ij \cdot \infty} \phi(t) = P(\Omega_{i,j}) \cdot \overline{\phi}(ij \cdot \infty) = \int_{\Omega_{i,j}} \overline{\phi}(ij \cdot \infty) \mathrm{d}P(x,y) \\ &= \int_{\Omega_{i,j}} \overline{\phi}(yf^*(x)) \mathrm{d}P(x,y),\ \forall\, (i,j) \in \{-1,1\}^2 . \end{aligned} \tag{C.2}$$

Besides, it is easy to verify that $yf_m(x) = yf^*(x) \in \mathbb{R}$ for any $(x,y) \in \Omega_0$ and any $m \in \mathbb{N}$, which means that

$$\int_{\Omega_0} \phi(yf_m(x)) \mathrm{d}P(x,y) = \int_{\Omega_0} \overline{\phi}(yf^*(x)) \mathrm{d}P(x,y),\ \forall\, m \in \mathbb{N}. \tag{C.3}$$

Combining (C.2) and (C.3), we obtain

$$\begin{aligned} &\inf \left\{ \mathcal{R}_P^\phi(g) \middle| g : [0,1]^d \to \mathbb{R} \text{ is measurable} \right\} \\ &\leq \varlimsup_{m \to +\infty} \mathcal{R}_P^\phi(f_m) = \varlimsup_{m \to +\infty} \int_{[0,1]^d \times \{-1,1\}} \phi(yf_m(x)) \mathrm{d}P(x,y) \\ &= \varlimsup_{m \to +\infty} \left(\int_{\Omega_0} \phi(yf_m(x)) \mathrm{d}P(x,y) + \sum_{i \in \{-1,1\}} \sum_{j \in \{-1,1\}} \int_{\Omega_{i,j}} \phi(yf_m(x)) \mathrm{d}P(x,y) \right) \\ &\leq \varlimsup_{m \to +\infty} \int_{\Omega_0} \phi(yf_m(x)) \mathrm{d}P(x,y) + \sum_{i \in \{-1,1\}} \sum_{j \in \{-1,1\}} \varlimsup_{m \to +\infty} \int_{\Omega_{i,j}} \phi(yf_m(x)) \mathrm{d}P(x,y) \\ &\leq \int_{\Omega_0} \overline{\phi}(yf^*(x)) \mathrm{d}P(x,y) + \sum_{i \in \{-1,1\}} \sum_{j \in \{-1,1\}} \int_{\Omega_{i,j}} \overline{\phi}(yf^*(x)) \mathrm{d}P(x,y) \\ &= \int_{[0,1]^d \times \{-1,1\}} \overline{\phi}(yf^*(x)) \mathrm{d}P(x,y). \end{aligned} \tag{C.4}$$

On the other hand, for any measurable $g : [0,1]^d \to \mathbb{R}$, it follows from (C.1) that

$$\int_{\{-1,1\}} \overline{\phi}(yf^*(x)) \mathrm{d}P(y|x) = \inf_{z \in [-\infty,\infty]} \int_{\{-1,1\}} \overline{\phi}(yz) \mathrm{d}P(y|x) \leq \int_{\{-1,1\}} \overline{\phi}(yg(x)) \mathrm{d}P(y|x)$$

$$= \int_{\{-1,1\}} \phi(yg(x))\mathrm{d}P(y|x) \text{ for } P_X\text{-almost all } x \in [0,1]^d.$$

Integrating both sides, we obtain

$$\begin{aligned}
&\int_{[0,1]^d\times\{-1,1\}} \overline{\phi}(yf^*(x))\mathrm{d}P(x,y) = \int_{[0,1]^d}\int_{\{-1,1\}} \overline{\phi}(yf^*(x))\mathrm{d}P(y|x)\mathrm{d}P_X(x)\\
&\leq \int_{[0,1]^d}\int_{\{-1,1\}} \phi(yg(x))\mathrm{d}P(y|x)P_X(x) = \int_{[0,1]^d\times\{-1,1\}} \phi(yg(x))\mathrm{d}P(x,y) = \mathcal{R}_P^\phi(g).
\end{aligned}$$

Since g is arbitrary, we deduce that

$$\int_{[0,1]^d\times\{-1,1\}} \overline{\phi}(yf^*(x))\mathrm{d}P(x,y) \leq \inf\left\{ \mathcal{R}_P^\phi(g) \middle| g : [0,1]^d \to \mathbb{R} \text{ is measurable}\right\},$$

which, together with (C.4), proves the desired result. ■

The next lemma gives the explicit form of the target function of the logistic risk.

Lemma C.2 *Let $\phi(t) = \log(1+\mathrm{e}^{-t})$ be the logistic loss, $d \in \mathbb{N}$, P be a probability measure on $[0,1]^d \times \{-1,1\}$, and η be the conditional probability function $P(\{1\}\,|\cdot)$ of P. Define*

$$f^* : [0,1]^d \to [-\infty,\infty],\ x \mapsto \begin{cases} \infty, & \textit{if } \eta(x) = 1,\\ \log\frac{\eta(x)}{1-\eta(x)}, & \textit{if } \eta(x) \in (0,1),\\ -\infty, & \textit{if } \eta(x) = 0, \end{cases} \tag{C.5}$$

which is a natural extension of the map

$$\left\{ z \in [0,1]^d \middle| \eta(z) \in (0,1)\right\} \ni x \mapsto \log\frac{\eta(x)}{1-\eta(x)} \in \mathbb{R}$$

to all of $[0,1]^d$. Then f^* is a target function of the ϕ-risk under P, i.e., (1.6) holds. In addition, the target function of the ϕ-risk under P is unique up to a P_X-null set. In other words, for any target function $f^\star$ of the ϕ-risk under P, we must have

$$P_X\left(\left\{ x \in [0,1]^d \middle| f^*(x) \neq f^\star(x)\right\}\right) = 0.$$

Proof Define

$$\overline{\phi} : [-\infty,\infty] \to [0,\infty],\ z \mapsto \begin{cases} 0, & \text{if } z = \infty,\\ \phi(z), & \text{if } z \in \mathbb{R},\\ \infty, & \text{if } z = -\infty, \end{cases} \tag{C.6}$$

which is a natural extension of the logistic loss ϕ to $[-\infty,\infty]$, and define

$$V_a : [-\infty,\infty] \to [0,\infty], z \mapsto a\overline{\phi}(z) + (1-a)\overline{\phi}(-z)$$

for any $a \in [0,1]$. Then we have that

$$\begin{aligned}&\int_{\{-1,1\}} \overline{\phi}(yz)\mathrm{d}P(y|x) = \eta(x)\overline{\phi}(z) + (1-\eta(x))\overline{\phi}(-z) \\ &= V_{\eta(x)}(z),\ \forall\, x \in [0,1]^d,\ z \in [-\infty,\infty].\end{aligned} \tag{C.7}$$

For any $a \in [0,1]$, we have that V_a is smooth on $\mathbb{R}$, and an elementary calculation gives

$$V_a''(t) = \frac{1}{2+\mathrm{e}^t+\mathrm{e}^{-t}} > 0,\ \forall\, t \in \mathbb{R}.$$

Therefore, V_a is strictly convex on $\mathbb{R}$ and

$$\begin{aligned}&\operatorname*{arg\,min}_{z\in\mathbb{R}} V_a(z) = \left\{z \in \mathbb{R} \middle| V_a'(z) = 0\right\} = \left\{z \in \mathbb{R} \,\middle|\, a\phi'(z) - (1-a)\phi'(-z) = 0\right\} \\ &= \left\{ z \in \mathbb{R} \middle| -a + \frac{\mathrm{e}^z}{1+\mathrm{e}^z} = 0 \right\} = \begin{cases} \left\{\log \frac{a}{1-a}\right\}, & \text{if } a \in (0,1), \\ \varnothing, & \text{if } a \in \{0,1\}. \end{cases}\end{aligned} \tag{C.8}$$

Besides, it is easy to verify that

$$V_a(z) = \infty,\ \forall\, a \in (0,1),\ \forall\, z \in \{\infty, -\infty\},$$

which, together with (C.8), yields

$$\operatorname*{arg\,min}_{z\in[-\infty,\infty]} V_a(z) = \operatorname*{arg\,min}_{z\in\mathbb{R}} V_a(z) = \left\{\log \frac{a}{1-a}\right\},\ \forall\, a \in (0,1). \tag{C.9}$$

In addition, it follows from

$$\overline{\phi}(z) > 0 = \overline{\phi}(\infty),\ \forall\, z \in [-\infty, \infty)$$

that

$$\operatorname*{arg\,min}_{z\in[-\infty,\infty]} V_1(z) = \operatorname*{arg\,min}_{z\in[-\infty,\infty]} \overline{\phi}(z) = \{\infty\} \tag{C.10}$$

and

$$\operatorname*{arg\,min}_{z\in[-\infty,\infty]} V_0(z) = \operatorname*{arg\,min}_{z\in[-\infty,\infty]} \overline{\phi}(-z) = \{-\infty\}. \tag{C.11}$$

Combining (C.7), (C.10) and (C.11), we obtain

$$\begin{aligned}&\operatorname*{arg\,min}_{z\in[-\infty,\infty]} \int_{\{-1,1\}} \overline{\phi}(yz)\mathrm{d}P(y|x) = \operatorname*{arg\,min}_{z\in[-\infty,\infty]} V_{\eta(x)}(z) = \begin{cases} \{+\infty\}, & \text{if } \eta(x) = 1, \\ \left\{\log \frac{\eta(x)}{1-\eta(x)}\right\}, & \text{if } \eta(x) \in (0,1), \\ \{-\infty\}, & \text{if } \eta(x) = 0 \end{cases} \\ &= \{f^*(x)\},\ \forall\, x \in [0,1]^d,\end{aligned}$$

which implies (1.6). Therefore, f^* is a target function of the ϕ-risk under the distribution P. Moreover, the uniqueness of the target function of the ϕ-risk under P follows immediately from the fact that for all $x \in [0,1]^d$ the set

$$\operatorname*{arg\,min}_{z\in[-\infty,\infty]} \int_{\{-1,1\}} \overline{\phi}(yz)\mathrm{d}P(y|x) = \{f^*(x)\}$$

contains exactly one point and the uniqueness (up to some P_X-null set) of the conditional distribution $P(\cdot|\cdot)$ of P. This completes the proof. ■

The Lemma C.3 below provides a formula for computing the infimum of the logistic risk over all real-valued measurable functions.

Lemma C.3 *Let $\phi(t) = \log(1 + \mathrm{e}^{-t})$ be the logistic loss, $\delta \in (0, 1/2]$, $d \in \mathbb{N}$, P be a probability measure on $[0,1]^d \times \{-1,1\}$, η be the conditional probability function $P(\{1\}\,|\cdot)$ of P, f^* be defined by* (C.5)*, $\overline{\phi}$ be defined by* (C.6)*, H be defined by*

$$H : [0,1] \to [0,\infty), t \mapsto \begin{cases} t\log\left(\frac{1}{t}\right) + (1-t)\log\left(\frac{1}{1-t}\right), & \text{if } t \in (0,1), \\ 0, & \text{if } t \in \{0,1\}, \end{cases}$$

and ψ be defined by

$$\begin{aligned} &\psi : [0,1]^d \times \{-1,1\} \to [0,\infty), \\ &(x,y) \mapsto \begin{cases} \phi\left(y\log\frac{\eta(x)}{1-\eta(x)}\right), & \text{if } \eta(x) \in [\delta, 1-\delta], \\ 0, & \text{if } \eta(x) \in \{0,1\}, \\ \eta(x)\log\frac{1}{\eta(x)} + (1-\eta(x))\log\frac{1}{1-\eta(x)}, & \text{if } \eta(x) \in (0,\delta)\cup(1-\delta,1). \end{cases} \end{aligned}$$

Then there holds

$$\begin{aligned} &\inf\left\{\mathcal{R}_P^\phi(g)\middle| g : [0,1]^d \to \mathbb{R} \text{ is measurable}\right\} = \int_{[0,1]^d\times\{-1,1\}} \overline{\phi}(yf^*(x))\mathrm{d}P(x,y) \\ &= \int_{[0,1]^d} H(\eta(x))\mathrm{d}P_X(x) = \int_{[0,1]^d\times\{-1,1\}} \psi(x,y)\mathrm{d}P(x,y). \end{aligned}$$

Proof According to Lemma C.2, f^* is a target function of the ϕ-risk under the distribution P, meaning that

$$f^*(x) \in \operatorname*{arg\,min}_{z\in[-\infty,\infty]} \int_{\{-1,1\}} \overline{\phi}(yz)\mathrm{d}P(y|x) \text{ for } P_X\text{-almost all } x \in [0,1]^d.$$

Then it follows from Lemma C.1 that

$$\begin{aligned} &\inf\left\{\mathcal{R}_P^\phi(g)\middle| g : [0,1]^d \to \mathbb{R} \text{ is measurable}\right\} = \int_{[0,1]^d\times\{-1,1\}} \overline{\phi}(yf^*(x))\mathrm{d}P(x,y) \\ &= \int_{[0,1]^d}\int_{\{-1,1\}} \overline{\phi}(yf^*(x))\mathrm{d}P(y|x)\mathrm{d}P_X(x) \\ &= \int_{[0,1]^d} \Big(\eta(x)\overline{\phi}(f^*(x)) + (1-\eta(x))\overline{\phi}(-f^*(x))\Big)\mathrm{d}P_X(x). \end{aligned} \tag{C.12}$$

For any $x \in [0,1]^d$, if $\eta(x) = 1$, then we have

$$\eta(x)\overline{\phi}(f^*(x)) + (1-\eta(x))\overline{\phi}(-f^*(x)) = \overline{\phi}(f^*(x)) = \overline{\phi}(+\infty) = 0 = H(\eta(x)) = 0$$

$$= 1 \cdot 0 + (1-1) \cdot 0 = \eta(x)\psi(x,1) + (1-\eta(x))\psi(x,-1) = \int_{\{-1,1\}} \psi(x,y)\mathrm{d}P(y|x);$$

If $\eta(x) = 0$, then we have

$$\begin{aligned}
&\eta(x)\overline{\phi}(f^*(x)) + (1-\eta(x))\overline{\phi}(-f^*(x)) = \overline{\phi}(-f^*(x)) = \overline{\phi}(+\infty) = 0 = H(\eta(x)) = 0 \\
&= 0 \cdot 0 + (1-0) \cdot 0 = \eta(x)\psi(x,1) + (1-\eta(x))\psi(x,-1) = \int_{\{-1,1\}} \psi(x,y)\mathrm{d}P(y|x);
\end{aligned}$$

If $\eta(x) \in (0,\delta) \cup (1-\delta,1)$, then we have

$$\begin{aligned}
&\eta(x)\overline{\phi}\left(f^*(x)\right) + (1-\eta(x))\overline{\phi}(-f^*(x)) \\
&= \eta(x)\phi\left(\log \frac{\eta(x)}{1-\eta(x)}\right) + (1-\eta(x))\phi\left(-\log \frac{\eta(x)}{1-\eta(x)}\right) \\
&= \eta(x)\log\left(1 + \frac{1-\eta(x)}{\eta(x)}\right) + (1-\eta(x))\log\left(1 + \frac{\eta(x)}{1-\eta(x)}\right) \\
&= \eta(x)\log\frac{1}{\eta(x)} + (1-\eta(x))\log\frac{1}{1-\eta(x)} \\
&= H(\eta(x)) = \int_{\{-1,1\}} \left(\eta(x)\log\frac{1}{\eta(x)} + (1-\eta(x))\log\frac{1}{1-\eta(x)}\right)\mathrm{d}P(y|x) \\
&= \int_{\{-1,1\}} \psi(x,y)\mathrm{d}P(y|x);
\end{aligned}$$

If $\eta(x) \in [\delta, 1-\delta]$, then we have that

$$\begin{aligned}
&\eta(x)\overline{\phi}\left(f^*(x)\right) + (1-\eta(x))\overline{\phi}(-f^*(x)) \\
&= \eta(x)\phi\left(\log \frac{\eta(x)}{1-\eta(x)}\right) + (1-\eta(x))\phi\left(-\log \frac{\eta(x)}{1-\eta(x)}\right) \\
&= \eta(x)\log\left(1 + \frac{1-\eta(x)}{\eta(x)}\right) + (1-\eta(x))\log\left(1 + \frac{\eta(x)}{1-\eta(x)}\right) \\
&= \eta(x)\log\frac{1}{\eta(x)} + (1-\eta(x))\log\frac{1}{1-\eta(x)} \\
&= H(\eta(x)) = \eta(x)\phi\left(\log \frac{\eta(x)}{1-\eta(x)}\right) + (1-\eta(x))\phi\left(-\log \frac{\eta(x)}{1-\eta(x)}\right) \\
&= \eta(x)\psi(x,1) + (1-\eta(x))\psi(x,-1) = \int_{\{-1,1\}} \psi(x,y)\mathrm{d}P(y|x).
\end{aligned}$$

In conclusion, we always have that

$$\eta(x)\overline{\phi}(f^*(x)) + (1-\eta(x))\overline{\phi}(-f^*(x)) = H(\eta(x)) = \int_{\{-1,1\}} \psi(x,y)\mathrm{d}P(y|x).$$

Since x is arbitrary, we deduce that

$$\int_{[0,1]^d} \Big(\eta(x)\overline{\phi}(f^*(x)) + (1-\eta(x))\overline{\phi}(-f^*(x))\Big)\mathrm{d}P_X(x) = \int_{[0,1]^d} H(\eta(x))\mathrm{d}P_X(x)$$

$$= \int_{[0,1]^d} \int_{\{-1,1\}} \psi(x,y)\mathrm{d}P(y|x)\mathrm{d}P_X(x) = \int_{[0,1]^d\times\{-1,1\}} \psi(x,y)\mathrm{d}P(x,y),$$

which, together with (C.12), proves the desired result. ■

C.2 Proof of Theorem 2.1

Appendix C.2 is devoted to the proof of Theorem 2.1.
Proof [Proof of Theorem 2.1] Throughout this proof, we denote

$$\Psi := \int_{[0,1]^d\times\{-1,1\}} \psi(x,y)\mathrm{d}P(x,y).$$

Then it follows from (2.3) and (2.4) that $0 \le \mathcal{R}_P^\phi\left(\hat{f}_n\right) - \Psi \le 2M < \infty$. Let $\{(X_k', Y_k')\}_{k=1}^n$ be an i.i.d. sample from distribution P which is independent of $\{(X_k, Y_k)\}_{k=1}^n$. By independence, we have

$$\mathbb{E}\left[\mathcal{R}_P^\phi\left(\hat{f}_n\right) - \Psi\right] = \frac{1}{n}\sum_{i=1}^n \mathbb{E}\left[\phi\left(Y_i'\hat{f}_n(X_i')\right) - \psi\left(X_i', Y_i'\right)\right]$$

with its empirical counterpart given by

$$\hat{R} := \frac{1}{n}\sum_{i=1}^n \mathbb{E}\left[\phi\left(Y_i\hat{f}_n(X_i)\right) - \psi(X_i, Y_i)\right].$$

Then we have

$$\begin{aligned}\hat{R} - \left(\mathcal{R}_P^\phi(g) - \Psi\right) &= \frac{1}{n}\sum_{i=1}^n \mathbb{E}\left[\phi\left(Y_i\hat{f}_n(X_i)\right) - \phi(Y_i g(X_i))\right] \\ &= \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^n \phi\left(Y_i\hat{f}_n(X_i)\right) - \frac{1}{n}\sum_{i=1}^n \phi\left(Y_i g(X_i)\right)\right] \le 0,\ \forall\, g \in \mathcal{F},\end{aligned}$$

where the last inequality follows from the fact that $\hat{f}_n$ is an empirical ϕ-risk minimizer which minimizes $\frac{1}{n}\sum_{i=1}^n \phi\left(Y_i g(X_i)\right)$ over all $g \in \mathcal{F}$. Hence $\hat{R} \le \inf_{g\in\mathcal{F}}\left(\mathcal{R}_P^\phi(g) - \Psi\right)$, which means that

$$\begin{aligned}\mathbb{E}\left[\mathcal{R}_P^\phi\left(\hat{f}_n\right) - \Psi\right] &= \left(\mathbb{E}\left[\mathcal{R}_P^\phi\left(\hat{f}_n\right) - \Psi\right] - (1+\varepsilon)\cdot\hat{R}\right) + (1+\varepsilon)\cdot\hat{R} \\ &\le \left(\mathbb{E}\left[\mathcal{R}_P^\phi\left(\hat{f}_n\right) - \Psi\right] - (1+\varepsilon)\cdot\hat{R}\right) + (1+\varepsilon)\cdot\inf_{g\in\mathcal{F}}\left(\mathcal{R}_P^\phi(g) - \Psi\right),\ \forall\,\varepsilon \in [0,1).\end{aligned} \tag{C.13}$$

We then establish an upper bound for $\mathbb{E}\left[\mathcal{R}_P^\phi\left(\hat{f}_n\right) - \Psi\right] - (1+\varepsilon)\cdot\hat{R}$ by using a similar argument to that in the proof of Lemma 4 of Schmidt-Hieber (2020). The desired inequality (2.6) will follow from this bound and (C.13). Recall that $W = \max\{3,\ \mathcal{N}(\mathcal{F},\gamma)\}$. From the definition of W, there exist $f_1, \cdots, f_W \in \mathcal{F}$ such that for any $f \in \mathcal{F}$, there exists some

$j \in \{1, \cdots, W\}$, such that $\|f - f_j\|_\infty \leq \gamma$. Therefore, there holds $\left\|\hat{f}_n - f_{j^*}\right\|_{[0,1]^d} \leq \gamma$ where j^* is a $\{1, \cdots, W\}$-valued statistic from the sample $\{(X_i, Y_i)\}_{i=1}^n$. Denote

$$A := M \cdot \sqrt{\frac{\log W}{\Gamma n}}. \tag{C.14}$$

And for $j = 1, 2, \cdots, W$, let

$$\begin{aligned}
&h_{j,1} := \mathcal{R}_P^\phi(f_j) - \Psi, \\
&h_{j,2} := \int_{[0,1]^d \times \{-1,1\}} (\phi(y f_j(x)) - \psi(x,y))^2 \,\mathrm{d}P(x,y), \\
&V_j := \left|\sum_{i=1}^n \left(\phi\left(Y_i f_j(X_i)\right) - \psi\left(X_i, Y_i\right) - \phi\left(Y_i' f_j(X_i')\right) + \psi\left(X_i', Y_i'\right)\right)\right|, \\
&r_j := A \vee \sqrt{h_{j,1}}.
\end{aligned} \tag{C.15}$$

Then define

$$T := \max_{j=1,\cdots,W} \frac{V_j}{r_j}.$$

Denote by $\mathbb{E}\left[\cdot \middle| (X_i, Y_i)_{i=1}^n\right]$ the conditional expectation with respect to $\{(X_i, Y_i)\}_{i=1}^n$. Then we have that

$$\begin{aligned}
r_{j^*} &= A \vee \sqrt{h_{j^*,1}} \\
&\leq A + \sqrt{h_{j^*,1}} \\
&= A + \sqrt{\mathbb{E}\left[\phi\left(Y' f_{j^*}(X')\right) - \psi(X', Y') \middle| (X_i, Y_i)_{i=1}^n\right]} \\
&\leq A + \sqrt{\gamma + \mathbb{E}\left[\phi\left(Y' \hat{f}_n(X')\right) - \psi(X', Y') \middle| (X_i, Y_i)_{i=1}^n\right]} \\
&= A + \sqrt{\gamma + \mathcal{R}_P^\phi\left(\hat{f}_n\right) - \Psi} \\
&\leq A + \sqrt{\gamma} + \sqrt{\mathcal{R}_P^\phi\left(\hat{f}_n\right) - \Psi},
\end{aligned}$$

where (X', Y') is an i.i.d. copy of $(X_i, Y_i)\,(1 \leq i \leq n)$ and the second inequality follows from

$$|\phi(t_1) - \phi(t_2)| \leq |t_1 - t_2|\,, \ \forall\, t_1, t_2 \in \mathbb{R} \tag{C.16}$$

and $\left\|f_{j^*}-\hat{f}\right\|_{[0,1]^d}\leq\gamma$. Consequently,

$$\begin{aligned}
&\mathbb{E}\left[\mathcal{R}_P^\phi\left(\hat{f}_n\right)-\Psi\right]-\hat{R}\leq\left|\hat{R}-\mathbb{E}\left[\mathcal{R}_P^\phi\left(\hat{f}_n\right)-\Psi\right]\right|\\
&=\frac{1}{n}\left|\mathbb{E}\left[\sum_{i=1}^n\left(\phi\left(Y_i\hat{f}_n(X_i)\right)-\psi(X_i,Y_i)-\phi\left(Y_i'\hat{f}_n(X_i')\right)+\psi(X_i',Y_i')\right)\right]\right|\\
&\leq\frac{1}{n}\mathbb{E}\left[\left|\sum_{i=1}^n\left(\phi\left(Y_if_{j^*}(X_i)\right)-\psi(X_i,Y_i)-\phi\left(Y_i'f_{j^*}(X_i')\right)+\psi(X_i',Y_i')\right)\right|\right]+2\gamma\\
&=\frac{1}{n}\mathbb{E}\left[V_{j^*}\right]+2\gamma\leq\frac{1}{n}\mathbb{E}\left[T\cdot r_{j^*}\right]+2\gamma\\
&\leq\frac{1}{n}\mathbb{E}\left[T\cdot\sqrt{\mathcal{R}_P^\phi\left(\hat{f}_n\right)-\Psi}\right]+\frac{A+\sqrt{\gamma}}{n}\cdot\mathbb{E}\left[T\right]+2\gamma\\
&\leq\frac{1}{n}\sqrt{\mathbb{E}\left[T^2\right]}\cdot\sqrt{\mathbb{E}\left[\mathcal{R}_P^\phi\left(\hat{f}_n\right)-\Psi\right]}+\frac{A+\sqrt{\gamma}}{n}\cdot\mathbb{E}\left[T\right]+2\gamma\\
&\leq\frac{\varepsilon\mathbb{E}\left[\mathcal{R}_P^\phi\left(\hat{f}_n\right)-\Psi\right]}{2+2\varepsilon}+\frac{(1+\varepsilon)\mathbb{E}\left[T^2\right]}{2\varepsilon\cdot n^2}+\frac{A+\sqrt{\gamma}}{n}\mathbb{E}\left[T\right]+2\gamma,\ \forall\,\varepsilon\in(0,1),
\end{aligned}\tag{C.17}$$

where the last inequality follows from $2\sqrt{ab}\leq\frac{\epsilon}{1+\epsilon}a+\frac{1+\epsilon}{\epsilon}b$, $\forall a>0,b>0$. We then bound $\mathbb{E}\left[T\right]$ and $\mathbb{E}\left[T^2\right]$ by Bernstein's inequality (see e.g., Chapter 3.1 of Cucker and Zhou (2007) and Chapter 6.2 of Steinwart and Christmann (2008)). Indeed, it follows from (2.5) and (C.15) that

$$h_{j,2}\leq\Gamma\cdot h_{j,1}\leq\Gamma\cdot(r_j)^2\,,\ \forall\,j\in\{1,\cdots,W\}.$$

For any $j\in\{1,\cdots,W\}$ and $t\geq 0$, we apply Bernstein's inequality to the zero mean i.i.d. random variables

$$\left\{\phi\left(Y_if_j(X_i)\right)-\psi(X_i,Y_i)-\phi\left(Y_i'f_j(X_i')\right)+\psi(X_i',Y_i')\right\}_{i=1}^n$$

and obtain

$$\begin{aligned}
&\mathbb{P}(V_j\geq t)\\
&=\mathbb{P}\left(\left|\sum_{i=1}^n\left(\phi\left(Y_if_j(X_i)\right)-\psi(X_i,Y_i)-\phi\left(Y_i'f_j(X_i')\right)+\psi(X_i',Y_i')\right)\right|\geq t\right)\\
&\leq 2\exp\left(\frac{-t^2/2}{Mt+\sum_{i=1}^n\mathbb{E}\left[\left(\phi\left(Y_if_j(X_i)\right)-\psi(X_i,Y_i)-\phi\left(Y_i'f_j(X_i')\right)+\psi(X_i',Y_i')\right)^2\right]}\right)\\
&\leq 2\exp\left(\frac{-t^2/2}{Mt+2\sum_{i=1}^n\mathbb{E}\left[\left(\phi\left(Y_if_j(X_i)\right)-\psi(X_i,Y_i)\right)^2+\left(\phi\left(Y_i'f_j(X_i')\right)-\psi(X_i',Y_i')\right)^2\right]}\right)\\
&=2\exp\left(\frac{-t^2/2}{Mt+4\sum_{i=1}^n h_{j,2}}\right)=2\exp\left(\frac{-t^2}{2Mt+8nh_{j,2}}\right)\leq 2\exp\left(-\frac{t^2}{2Mt+8n\Gamma\cdot(r_j)^2}\right).
\end{aligned}$$

Hence

$$\begin{aligned}
\mathbb{P}(T \geq t) &\leq \sum_{j=1}^{W} \mathbb{P}(V_j/r_j \geq t) = \sum_{j=1}^{W} \mathbb{P}(V_j \geq tr_j) \\
&\leq 2\sum_{j=1}^{W} \exp\left(-\frac{(tr_j)^2}{2Mtr_j + 8n\Gamma \cdot r_j^2}\right) = 2\sum_{j=1}^{W} \exp\left(-\frac{t^2}{2Mt/r_j + 8n\Gamma}\right) \\
&\leq 2\sum_{j=1}^{W} \exp\left(-\frac{t^2}{2Mt/A + 8n\Gamma}\right) = 2W\exp\left(-\frac{t^2}{2Mt/A + 8n\Gamma}\right), \ \forall\, t \in [0,\infty).
\end{aligned}$$

Therefore, for any $\theta \in \{1,2\}$, by taking

$$B := \left(\frac{M}{A} \cdot \log W + \sqrt{\left(\frac{M}{A} \cdot \log W\right)^2 + 8n\Gamma \log W}\right)^{\theta} = 4^{\theta} \cdot (n\Gamma \log W)^{\theta/2},$$

we derive

$$\begin{aligned}
\mathbb{E}\left[T^{\theta}\right] &= \int_0^{\infty} \mathbb{P}\left(T \geq t^{1/\theta}\right) \mathrm{d}t \leq B + \int_B^{\infty} \mathbb{P}\left(T \geq t^{1/\theta}\right) \mathrm{d}t \\
&\leq B + \int_B^{\infty} \left(2W \exp\left(-\frac{t^{2/\theta}}{2Mt^{1/\theta}/A + 8n\Gamma}\right)\right) \mathrm{d}t \\
&\leq B + \int_B^{\infty} \left(2W \exp\left(-\frac{B^{1/\theta} \cdot t^{1/\theta}}{2MB^{1/\theta}/A + 8n\Gamma}\right)\right) \mathrm{d}t \\
&= B + 2WB\theta \cdot (\log W)^{-\theta} \int_{\log W}^{\infty} \mathrm{e}^{-u} u^{\theta-1} \mathrm{d}u \\
&\leq B + 2WB\theta \cdot (\log W)^{-\theta} \cdot \theta \cdot \mathrm{e}^{-\log W} (\log W)^{\theta-1} \\
&\leq 5\theta B \leq 5\theta \cdot 4^{\theta} \cdot (n\Gamma \log W)^{\theta/2}.
\end{aligned}$$

Plugging the inequality above and (C.14) into (C.17), we obtain

$$\begin{aligned}
&\mathbb{E}\left[\mathcal{R}_P^{\phi}\left(\hat{f}_n\right) - \Psi\right] - \hat{R} \leq \left|\hat{R} - \mathbb{E}\left[\mathcal{R}_P^{\phi}\left(\hat{f}_n\right) - \Psi\right]\right| \\
&\leq \frac{\varepsilon \mathbb{E}\left[\mathcal{R}_P^{\phi}\left(\hat{f}_n\right) - \Psi\right]}{2 + 2\varepsilon} + \frac{(1+\varepsilon)\mathbb{E}\left[T^2\right]}{2\varepsilon \cdot n^2} + \frac{A + \sqrt{\gamma}}{n}\mathbb{E}\left[T\right] + 2\gamma \\
&\leq \frac{\varepsilon}{1+\varepsilon}\mathbb{E}\left[\mathcal{R}_P^{\phi}\left(\hat{f}_n\right) - \Psi\right] + 20 \cdot \sqrt{\gamma} \cdot \sqrt{\frac{\Gamma \log W}{n}} \\
&\quad + 20M \cdot \frac{\log W}{n} + 80 \cdot \frac{\Gamma \log W}{n} \cdot \frac{1+\varepsilon}{\varepsilon} + 2\gamma, \ \forall\, \varepsilon \in (0,1).
\end{aligned}$$

Multiplying the above inequality by $(1+\varepsilon)$ and then rearranging, we obtain that

$$\begin{aligned}
&\mathbb{E}\left[\mathcal{R}_P^{\phi}\left(\hat{f}_n\right) - \Psi\right] - (1+\varepsilon) \cdot \hat{R} \leq 20 \cdot (1+\varepsilon) \cdot \sqrt{\gamma} \cdot \sqrt{\frac{\Gamma \log W}{n}} \\
&\quad + 20 \cdot (1+\varepsilon) \cdot M \cdot \frac{\log W}{n} + 80 \cdot \frac{\Gamma \log W}{n} \cdot \frac{(1+\varepsilon)^2}{\varepsilon} + (2+2\varepsilon) \cdot \gamma, \ \forall\, \varepsilon \in (0,1).
\end{aligned} \tag{C.18}$$

Combining (C.18) and (C.13), we deduce that

$$\mathbb{E}\left[\mathcal{R}_P^\phi\left(\hat{f}_n\right)-\Psi\right] \leq (1+\varepsilon)\cdot\inf_{g\in\mathcal{F}}\left(\mathcal{R}_P^\phi(g)-\Psi\right)+20\cdot(1+\varepsilon)\cdot\sqrt{\gamma}\cdot\sqrt{\frac{\Gamma\log W}{n}}$$
$$+20\cdot(1+\varepsilon)\cdot M\cdot\frac{\log W}{n}+80\cdot\frac{\Gamma\log W}{n}\cdot\frac{(1+\varepsilon)^2}{\varepsilon}+(2+2\varepsilon)\cdot\gamma,\ \forall\,\varepsilon\in(0,1).$$

This proves the desired inequality (2.6) and completes the proof of Theorem 2.1. ■

C.3 Proof of Theorem 2.4

To prove Theorem 2.4, we need the following Lemma C.4 and Lemma C.5.

Lemma C.4, which describes neural networks that approximate the multiplication operator, can be derived directly from Lemma A.2 of Schmidt-Hieber (2020). Thus we omit its proof. One can also find a similar result to Lemma C.4 in the earlier paper Yarotsky (2017) (cf. Proposition 3 therein).

Lemma C.4 *For any $\varepsilon\in(0,1/2]$, there exists a neural network*

$$\mathrm{M}\in\mathcal{F}_2^{\mathbf{FNN}}\left(15\log\frac{1}{\varepsilon},6,900\log\frac{1}{\varepsilon},1,1\right)$$

such that for any $t,t'\in[0,1]$, there hold $\mathrm{M}(t,t')\in[0,1]$, $\mathrm{M}(t,0)=\mathrm{M}(0,t')=0$ and

$$\left|\mathrm{M}(t,t')-t\cdot t'\right|\leq\varepsilon.$$

In Lemma C.5, we construct a neural network which performs the operation of multiplying the inputs by 2^k.

Lemma C.5 *Let k be a positive integer and f be a univariate function given by $f(x)=2^k\cdot\max\{x,0\}$. Then*

$$f\in\mathcal{F}_1^{\mathbf{FNN}}(k,2,4k,1,\infty).$$

Proof For any $1\leq i\leq k-1$, let $v_i=(0,0)^\top$ and

$$\boldsymbol{W}_i=\begin{pmatrix}1&1\\1&1\end{pmatrix}.$$

In addition, take

$$\boldsymbol{W}_0=(1,1)^\top,\boldsymbol{W}_k=(1,1),\text{ and }v_k=(0,0)^\top.$$

Then we have

$$f=x\mapsto\boldsymbol{W}_k\sigma_{v_k}\boldsymbol{W}_{k-1}\sigma_{v_{k-1}}\cdots\boldsymbol{W}_1\sigma_{v_1}\boldsymbol{W}_0x\in\mathcal{F}_1^{\mathbf{FNN}}(k,2,4k,1,\infty),$$

which proves this lemma. ■

Now we are in the position to prove Theorem 2.4.

Proof [Proof of Theorem 2.4] Given $a \in (0, 1/2]$, let $I := \lceil -\log_2 a \rceil$ and $J_k := \left[\frac{1}{3\cdot 2^k}, \frac{1}{2^k}\right]$ for $k = 0, 1, 2, \cdots$. Then $1 \le I \le 1 - \log_2 a \le 4 \log \frac{1}{a}$. The idea of proof is to construct neural networks $\left\{\tilde{h}_k\right\}_k$ which satisfy that $0 \le \tilde{h}_k(t) \le 1$ and $(8 \log a) \cdot \tilde{h}_k$ approximates the natural logarithm function on J_k. Then the function

$$x \mapsto (8 \log a) \cdot \sum_k \mathrm{M}\left(\tilde{h}_k(x), \tilde{f}_k(x)\right)$$

is the desired neural network in Theorem 2.4, where M is the neural network that approximates multiplication operators given in Lemma C.4 and $\{\tilde{f}_k\}_k$ are neural networks representing piecewise linear function supported on J_k which constitutes a partition of unity.

Specifically, given $\alpha \in (0, \infty)$, there exists some $r_\alpha > 0$ only depending on α such that

$$x \mapsto \log\left(\frac{2x}{3} + \frac{1}{3}\right) \in \mathcal{B}^\alpha_{r_\alpha}([0, 1]).$$

Hence it follows from Corollary B.1 that there exists

$$\begin{aligned}\tilde{g}_1 &\in \mathcal{F}^{\mathbf{FNN}}_1\left(C_\alpha \log \frac{2}{\varepsilon}, C_\alpha \left(\frac{2}{\varepsilon}\right)^{1/\alpha}, C_\alpha \left(\frac{2}{\varepsilon}\right)^{1/\alpha} \log \frac{2}{\varepsilon}, 1, \infty\right) \\ &\subset \mathcal{F}^{\mathbf{FNN}}_1\left(C_\alpha \log \frac{1}{\varepsilon}, C_\alpha \left(\frac{1}{\varepsilon}\right)^{1/\alpha}, C_\alpha \left(\frac{1}{\varepsilon}\right)^{1/\alpha} \log \frac{1}{\varepsilon}, 1, \infty\right)\end{aligned}$$

such that

$$\sup_{x \in [0,1]} \left|\tilde{g}_1(x) - \log\left(\frac{2x}{3} + \frac{1}{3}\right)\right| \le \varepsilon/2.$$

Recall that the ReLU function is given by $\sigma(t) = \max\{t, 0\}$. Let

$$\tilde{g}_2 : \mathbb{R} \to \mathbb{R}, \quad x \mapsto -\sigma\left(-\sigma\left(\tilde{g}_1(x) + \log 3\right) + \log 3\right).$$

Then

$$\tilde{g}_2 \in \mathcal{F}^{\mathbf{FNN}}_1\left(C_\alpha \log \frac{1}{\varepsilon}, C_\alpha \left(\frac{1}{\varepsilon}\right)^{1/\alpha}, C_\alpha \left(\frac{1}{\varepsilon}\right)^{1/\alpha} \log \frac{1}{\varepsilon}, 1, \infty\right), \tag{C.19}$$

and for $x \in \mathbb{R}$, there holds

$$-\log 3 \le \tilde{g}_2(x) = \begin{cases} -\log 3, & \text{if } \tilde{g}_1(x) < -\log 3, \\ \tilde{g}_1(x), & \text{if } -\log 3 \le \tilde{g}_1(x) \le 0, \\ 0, & \text{if } \tilde{g}_1(x) > 0. \end{cases}$$

Moreover, since $-\log 3 \le \log\left(\frac{2x}{3} + \frac{1}{3}\right) \le 0$ whenever $x \in [0, 1]$, we have

$$\sup_{x \in [0,1]} \left|\tilde{g}_2(x) - \log\left(\frac{2x}{3} + \frac{1}{3}\right)\right| \le \sup_{x \in [0,1]} \left|\tilde{g}_1(x) - \log\left(\frac{2x}{3} + \frac{1}{3}\right)\right| \le \varepsilon/2.$$

Let $x = \frac{3\cdot 2^k\cdot t-1}{2}$ in the above inequality, we obtain

$$\sup_{t\in J_k}\left|\tilde{g}_2\left(\frac{3\cdot 2^k\cdot t-1}{2}\right)-k\log 2-\log t\right|\leq \varepsilon/2,\quad \forall\, k=0,1,2,\cdots. \tag{C.20}$$

For any $0\leq k\leq I$, define

$$\tilde{h}_k:\mathbb{R}\to\mathbb{R},\quad t\mapsto \sigma\left(\frac{\sigma\left(-\tilde{g}_2\left(\sigma\left(\frac{3}{4\cdot 2^{I-k}}\cdot 2^{I+1}\cdot\sigma(t)-\frac{1}{2}\right)\right)\right)}{8\log\frac{1}{a}}+\frac{k\log 2}{8\log\frac{1}{a}}\right).$$

Then we have

$$\begin{aligned}0\leq \tilde{h}_k(t)&\leq\left|\frac{\sigma\left(-\tilde{g}_2\left(\sigma\left(\frac{3}{4\cdot 2^{I-k}}\cdot 2^{I+1}\cdot\sigma(t)-\frac{1}{2}\right)\right)\right)}{8\log\frac{1}{a}}+\frac{k\log 2}{8\log\frac{1}{a}}\right|\\&\leq\left|\frac{-\tilde{g}_2\left(\sigma\left(\frac{3}{4\cdot 2^{I-k}}\cdot 2^{I+1}\cdot\sigma(t)-\frac{1}{2}\right)\right)}{8\log\frac{1}{a}}\right|+\frac{k\log 2}{8\log\frac{1}{a}}\\&\leq\frac{\sup_{x\in\mathbb{R}}|\tilde{g}_2(x)|}{8\log\frac{1}{a}}+\frac{I}{8\log\frac{1}{a}}\leq\frac{\log 3+4\log\frac{1}{a}}{8\log\frac{1}{a}}\leq 1,\ \forall\, t\in\mathbb{R}.\end{aligned} \tag{C.21}$$

Therefore, it follows from (C.19), the definition of $\tilde{h}_k$, and Lemma C.5 that (cf. Figure C.1)

$$\begin{aligned}\tilde{h}_k&\in\mathcal{F}_1^{\mathbf{FNN}}\left(C_\alpha\log\frac{1}{\varepsilon}+I, C_\alpha\left(\frac{1}{\varepsilon}\right)^{\frac{1}{\alpha}}, C_\alpha\left(\frac{1}{\varepsilon}\right)^{\frac{1}{\alpha}}\log\frac{1}{\varepsilon}+4I,1,1\right)\\&\subset\mathcal{F}_1^{\mathbf{FNN}}\left(C_\alpha\log\frac{1}{\varepsilon}+4\log\frac{1}{a}, C_\alpha\left(\frac{1}{\varepsilon}\right)^{\frac{1}{\alpha}}, C_\alpha\left(\frac{1}{\varepsilon}\right)^{\frac{1}{\alpha}}\log\frac{1}{\varepsilon}+16\log\frac{1}{a},1,1\right)\end{aligned} \tag{C.22}$$

for all $0\leq k\leq I$. Besides, according to (C.20), it is easy to verify that for $0\leq k\leq I$, there holds

$$\left|(8\log a)\cdot\tilde{h}_k(t)-\log t\right|=\left|\tilde{g}_2\left(\frac{3}{2}\cdot 2^k\cdot t-1/2\right)-k\log 2-\log t\right|\leq\varepsilon/2,\quad \forall\, t\in J_k.$$

Define

$$\tilde{f}_0:\mathbb{R}\to[0,1],\quad x\mapsto\begin{cases}0, & \text{if } x\in(-\infty,1/3),\\ 6\cdot\left(x-\frac{1}{3}\right), & \text{if } x\in[1/3,1/2],\\ 1, & \text{if } x\in(1/2,\infty),\end{cases}$$

and for $k\in\mathbb{N}$,

$$\tilde{f}_k:\mathbb{R}\to[0,1],\quad x\mapsto\begin{cases}0, & \text{if } x\in\mathbb{R}\setminus J_k,\\ 6\cdot 2^k\cdot\left(x-\frac{1}{3\cdot 2^k}\right), & \text{if } x\in\left[\frac{1}{3\cdot 2^k},\frac{1}{2^{k+1}}\right),\\ 1, & \text{if } x\in\left[\frac{1}{2^{k+1}},\frac{1}{3\cdot 2^{k-1}}\right],\\ -3\cdot 2^k\cdot\left(x-\frac{1}{2^k}\right), & \text{if } x\in\left(\frac{1}{3\cdot 2^{k-1}},\frac{1}{2^k}\right].\end{cases}$$

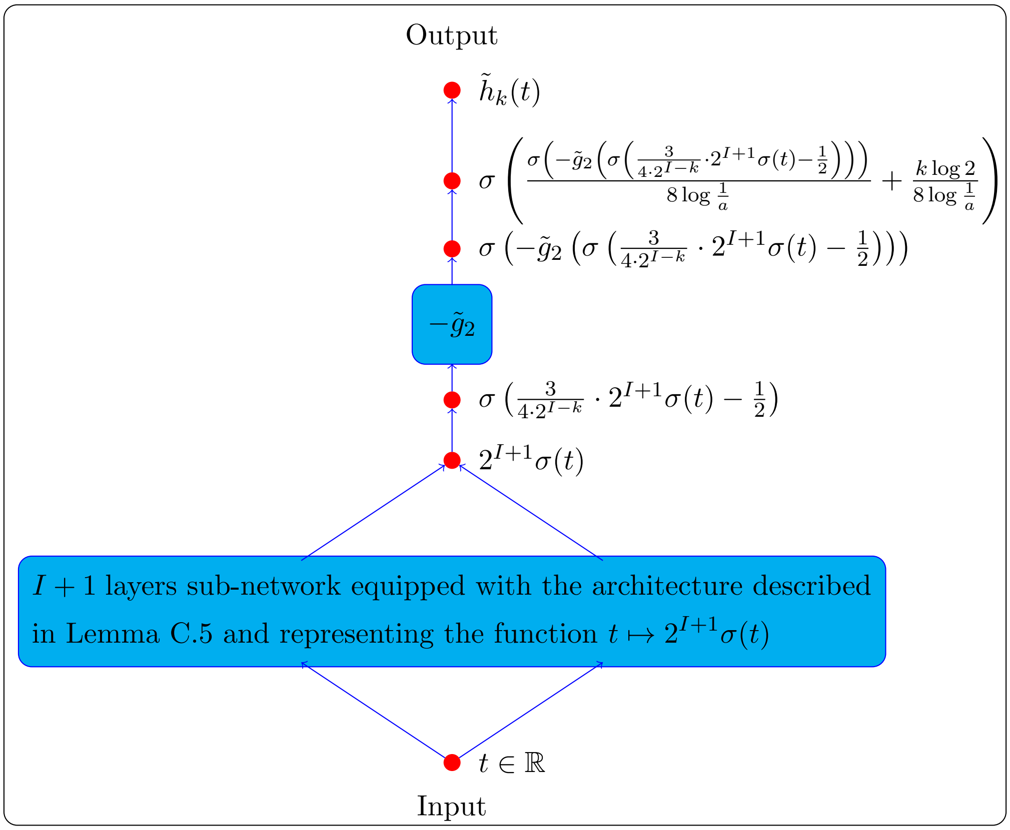

Figure C.1: Networks representing functions $\tilde{h}_k$.

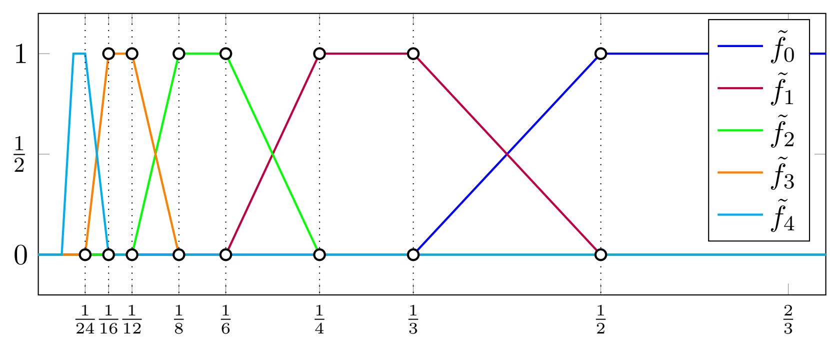

Figure C.2: Graphs of functions $\tilde{f}_k$.

Then it is easy to show that for any $x \in \mathbb{R}$ and $k \in \mathbb{N}$, there hold

$$\begin{aligned}\tilde{f}_k(x) = &\frac{6}{2^{I-k+3}} \cdot 2^{I+3} \cdot \sigma\left(x - \frac{1}{3 \cdot 2^k}\right) - \frac{6}{2^{I-k+3}} \cdot 2^{I+3} \cdot \sigma\left(x - \frac{1}{2^{k+1}}\right)\\ &+ \frac{6}{2^{I-k+4}} \cdot 2^{I+3} \cdot \sigma\left(x - \frac{1}{2^k}\right) - \frac{6}{2^{I-k+3}} \cdot 2^{I+3} \cdot \sigma\left(x - \frac{1}{3 \cdot 2^{k-1}}\right),\end{aligned}$$

and

$$\tilde{f}_0(x) = \frac{6}{2^{I+3}} \cdot 2^{I+3} \cdot \sigma(x - 1/3) - \frac{6}{2^{I+3}} \cdot 2^{I+3} \cdot \sigma(x - 1/2).$$

Hence it follows from Lemma C.5 that (cf. Figure C.3)

$$\begin{aligned}\tilde{f}_k &\in \mathcal{F}_1^{\mathbf{FNN}}(I+5, 8, 16I+60, 1, \infty) \\ &\subset \mathcal{F}_1^{\mathbf{FNN}}\left(12\log\frac{1}{a}, 8, 152\log\frac{1}{a}, 1, \infty\right), \quad \forall\, 0 \le k \le I.\end{aligned} \tag{C.23}$$

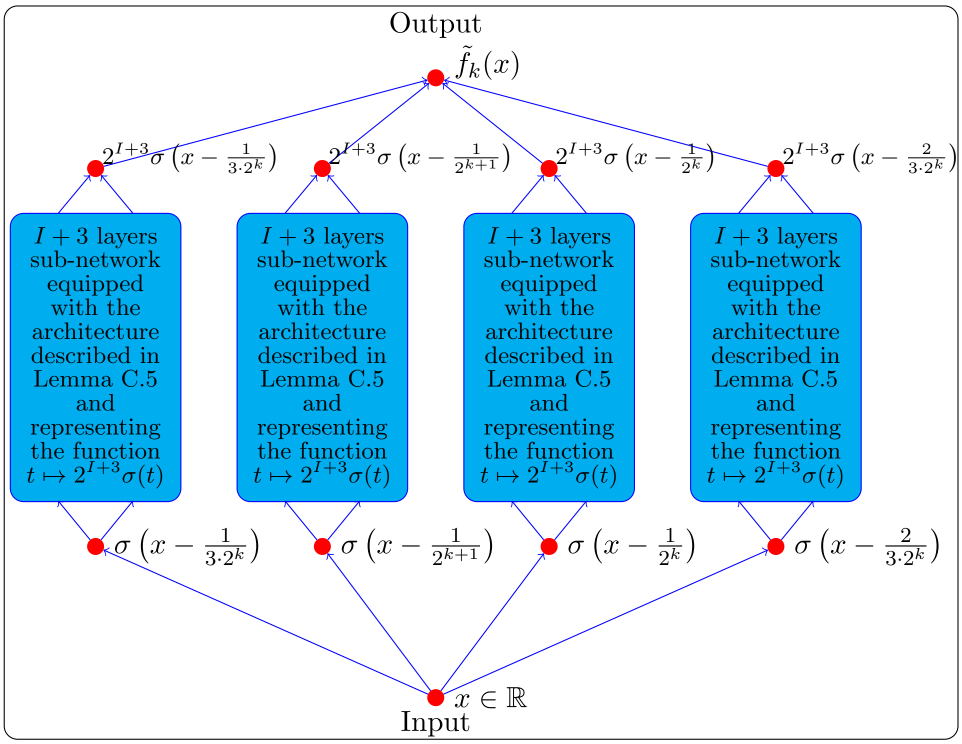

Figure C.3: Networks representing functions $\tilde{f}_k$.

Next, we show that

$$\sup_{t\in[a,1]}\left|\log(t) + 8\log\left(\frac{1}{a}\right)\sum_{k=0}^{I}\tilde{h}_k(t)\tilde{f}_k(t)\right| \le \varepsilon/2. \tag{C.24}$$

Indeed, we have the following inequalities:

$$\begin{aligned}\left|\log(t) + 8\log\left(\frac{1}{a}\right)\sum_{k=0}^{I}\tilde{h}_k(t)\tilde{f}_k(t)\right| &= \left|\log t + 8\log\left(\frac{1}{a}\right)\tilde{h}_0(t)\tilde{f}_0(t)\right| \\ &= \left|\log t + 8\log\left(\frac{1}{a}\right)\tilde{h}_0(t)\right| \le \varepsilon/2,\ \forall\, t \in [1/2, 1];\end{aligned} \tag{C.25}$$

$$\left|\log(t) + 8\log\left(\frac{1}{a}\right)\sum_{k=0}^{I}\tilde{h}_k(t)\tilde{f}_k(t)\right| = \left|\log(t) + 8\log\left(\frac{1}{a}\right)\tilde{h}_{m-1}(t)\right| \le \varepsilon/2, \\ \forall t \in \left[\frac{1}{2^m}, \frac{1}{3\cdot 2^{m-2}}\right] \cap [a,1] \text{ with } 2 \le m \le I; \tag{C.26}$$

and

$$\begin{aligned} &\left|\log(t) + 8\log\left(\frac{1}{a}\right)\sum_{k=0}^{I}\tilde{h}_k(t)\tilde{f}_k(t)\right| \\ &= \left|\log(t)\left(\tilde{f}_m(t) + \tilde{f}_{m-1}(t)\right) - 8\log(a)\left(\tilde{h}_m(t)\tilde{f}_m(t) + \tilde{h}_{m-1}(t)\tilde{f}_{m-1}(t)\right)\right| \\ &\le \tilde{f}_m(t)\left|\log(t) - 8\log(a)\,\tilde{h}_m(t)\right| + \tilde{f}_{m-1}(t)\left|\log(t) - 8\log(a)\,\tilde{h}_{m-1}(t)\right| \\ &\le \tilde{f}_m(t)\cdot\frac{\varepsilon}{2} + \tilde{f}_{m-1}(t)\cdot\frac{\varepsilon}{2} = \frac{\varepsilon}{2}, \quad \forall\, t \in \left[\frac{1}{3\cdot 2^{m-1}}, \frac{1}{2^m}\right] \cap [a,1] \text{ with } 1 \le m \le I. \end{aligned} \tag{C.27}$$

Note that

$$[a,1] \subset [1/2,1] \cup \left(\bigcup_{m=1}^{I}\left[\frac{1}{3\cdot 2^{m-1}}, \frac{1}{2^m}\right]\right) \cup \left(\bigcup_{m=2}^{I}\left[\frac{1}{2^m}, \frac{1}{3\cdot 2^{m-2}}\right]\right).$$

Consequently, (C.24) follows immediately from (C.25), (C.26) and (C.27).

From Lemma C.4 we know that there exists

$$\mathrm{M} \in \mathcal{F}_2^{\mathbf{FNN}}\left(15\log\frac{96(\log a)^2}{\varepsilon}, 6, 900\log\frac{96(\log a)^2}{\varepsilon}, 1, 1\right) \tag{C.28}$$

such that for any $t, t' \in [0,1]$, there hold $\mathrm{M}(t,t') \in [0,1]$, $\mathrm{M}(t,0) = \mathrm{M}(0,t') = 0$ and

$$\left|\mathrm{M}(t,t') - t\cdot t'\right| \le \frac{\varepsilon}{96(\log a)^2}. \tag{C.29}$$

Define

$$\tilde{g}_3 : \mathbb{R} \to \mathbb{R}, \quad x \mapsto \sum_{k=0}^{I}\mathrm{M}\left(\tilde{h}_k(x), \tilde{f}_k(x)\right),$$

and

$$\begin{aligned} &\tilde{f} : \mathbb{R} \to \mathbb{R}, \\ &\quad x \mapsto \sum_{k=1}^{8I}\left[\frac{\log(a)}{I}\cdot\sigma\left(\frac{\log b}{8\log a} + \sigma\left(\sigma(\tilde{g}_3(x)) - \frac{\log b}{8\log a}\right) - \sigma\left(\sigma(\tilde{g}_3(x)) - \frac{1}{8}\right)\right)\right]. \end{aligned}$$

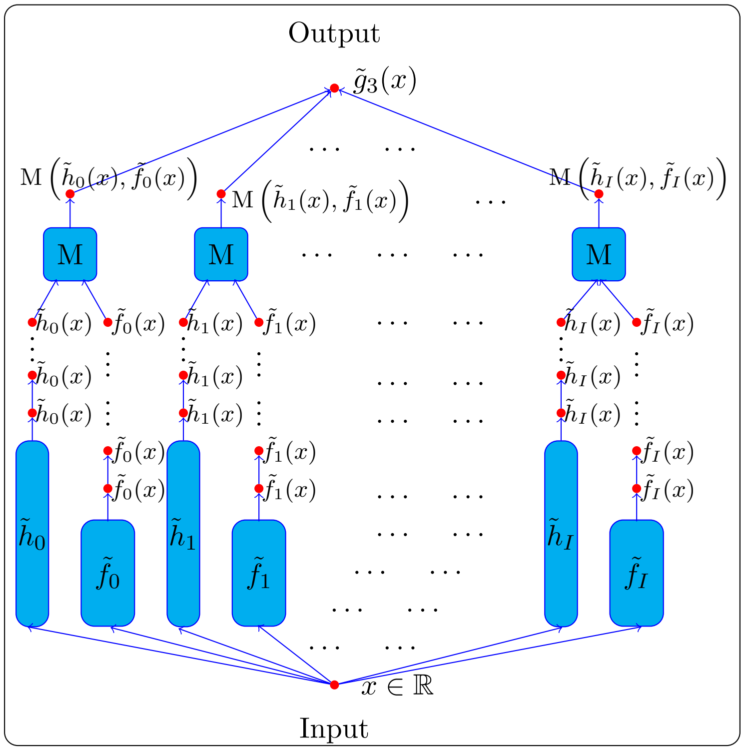

Figure C.4: The network representing the function $\tilde{g}_3$.

Then it follows from (C.21),(C.29), (C.24), the definitions of $\tilde{f}_k$ and $\tilde{g}_3$ that

$$
\begin{aligned}
&|\log t - 8\log(a)\cdot\tilde{g}_3(t)| \\
&\leq 8\log\left(\frac{1}{a}\right)\cdot\left|\tilde{g}_3(t)-\sum_{k=0}^{I}\tilde{h}_k(t)\tilde{f}_k(t)\right| + \left|\log t + 8\log\left(\frac{1}{a}\right)\sum_{k=0}^{I}\tilde{h}_k(t)\tilde{f}_k(t)\right| \\
&\leq 8\log\left(\frac{1}{a}\right)\cdot\left|\tilde{g}_3(t)-\sum_{k=0}^{I}\tilde{h}_k(t)\tilde{f}_k(t)\right| + \varepsilon/2 \\
&\leq \varepsilon/2 + |8\log a|\cdot\sum_{k=0}^{I}\left|\mathrm{M}\left(\tilde{h}_k(t),\tilde{f}_k(t)\right)-\tilde{h}_k(t)\tilde{f}_k(t)\right| \\
&\leq \varepsilon/2 + |8\log a|\cdot(I+1)\cdot\frac{\varepsilon}{96\,(\log a)^2} \leq \varepsilon,\ \forall\, t\in[a,1].
\end{aligned}
\tag{C.30}
$$

However, for any $t \in \mathbb{R}$, by the definition of $\tilde{f}$, we have

$$
\tilde{f}(t) = \begin{cases} 8\log(a)\cdot\tilde{g}_3(t), & \text{if } 8\log(a)\cdot\tilde{g}_3(t)\in[\log a,\log b], \\ \log a, & \text{if } 8\log(a)\cdot\tilde{g}_3(t) < \log a, \\ \log b, & \text{if } 8\log(a)\cdot\tilde{g}_3(t) > \log b, \end{cases}
\tag{C.31}
$$

satisfying $\log a \leq \tilde{f}(t) \leq \log b \leq 0$.

Then by (C.30), (C.31) and the fact that $\log t \in [\log a, \log b]$, $\forall\, t \in [a,b]$, we obtain

$$\left|\log t - \tilde{f}(t)\right| \le |\log t - 8\log(a)\cdot \tilde{g}_3(t)| \le \varepsilon,\ \forall\, t \in [a,b].$$

That is,

$$\sup_{t\in[a,b]} \left|\log t - \tilde{f}(t)\right| \le \varepsilon. \tag{C.32}$$

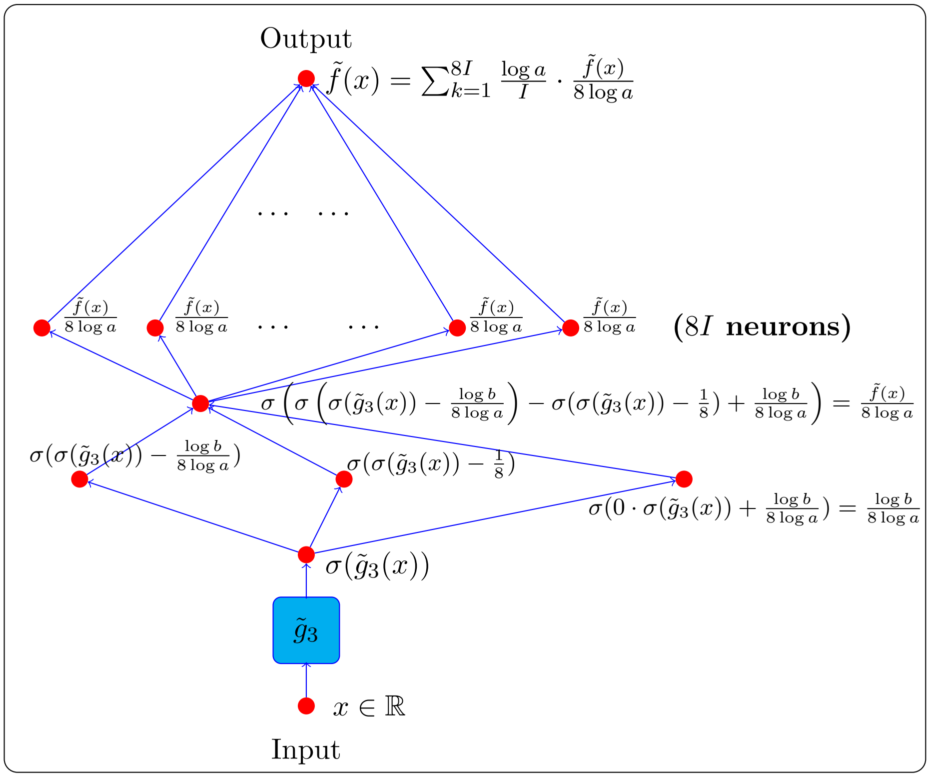

Figure C.5: The network representing the function $\tilde{f}$.

On the other hand, it follows from (C.22), (C.23), (C.28), the definition of $\tilde{g}_3$, and $1 \le I \le 4\log\frac{1}{a}$ that

$$\begin{aligned}
\tilde{g}_3 \in \mathcal{F}_1^{\mathbf{FNN}}\Bigg(& C_\alpha \log\frac{1}{\varepsilon} + I + 15\log\left(96\,(\log a)^2\right), C_\alpha\left(\frac{1}{\varepsilon}\right)^{\frac{1}{\alpha}} I,\\
&(I+1)\cdot\left(20I + C_\alpha\left(\frac{1}{\varepsilon}\right)^{\frac{1}{\alpha}}\cdot\log\frac{1}{\varepsilon} + 900\log\left(96\,(\log a)^2\right)\right), 1, \infty\Bigg)\\
\subset \mathcal{F}_1^{\mathbf{FNN}}\Bigg(& C_\alpha \log\frac{1}{\varepsilon} + 139\log\frac{1}{a}, C_\alpha\left(\frac{1}{\varepsilon}\right)^{\frac{1}{\alpha}}\log\frac{1}{a},\\
&C_\alpha\left(\frac{1}{\varepsilon}\right)^{\frac{1}{\alpha}}\cdot\left(\log\frac{1}{\varepsilon}\right)\cdot\left(\log\frac{1}{a}\right) + 65440\,(\log a)^2, 1, \infty\Bigg).
\end{aligned}$$

Then by the definition of $\tilde{f}$ we obtain (cf. Figure C.5)

$$\tilde{f} \in \mathcal{F}_1^{\mathbf{FNN}}\Bigg(C_\alpha \log\frac{1}{\varepsilon} + 139\log\frac{1}{a}, C_\alpha\left(\frac{1}{\varepsilon}\right)^{\frac{1}{\alpha}}\log\frac{1}{a},$$
$$C_\alpha\left(\frac{1}{\varepsilon}\right)^{\frac{1}{\alpha}}\cdot\left(\log\frac{1}{\varepsilon}\right)\cdot\left(\log\frac{1}{a}\right) + 65440\left(\log a\right)^2, 1, \infty\Bigg).$$

This, together with (C.31) and (C.32), completes the proof of Theorem 2.4. ■

C.4 Proof of Theorem 2.2 and Theorem 2.3

Appendix C.4 is devoted to the proof of Theorem 2.2 and Theorem 2.3. We will first establish several lemmas. We then use these lemmas to prove Theorem 2.3. Finally, we derive Theorem 2.2 by applying Theorem 2.3 with $q = 0$, $d_* = d$ and $d_\star = K = 1$.

Lemma C.6 *Let $\phi(t) = \log(1 + \mathrm{e}^{-t})$ be the logistic loss. Suppose real numbers a, f, A, B satisfy that $0 < a < 1$ and $A \le \min\left\{f, \log\frac{a}{1-a}\right\} \le \max\left\{f, \log\frac{a}{1-a}\right\} \le B$. Then there holds*

$$\begin{aligned}
&\min\left\{\frac{1}{4 + 2\mathrm{e}^A + 2\mathrm{e}^{-A}}, \frac{1}{4 + 2\mathrm{e}^B + 2\mathrm{e}^{-B}}\right\}\cdot\left|f - \log\frac{a}{1-a}\right|^2 \\
&\le a\phi(f) + (1-a)\phi(-f) - a\log\frac{1}{a} - (1-a)\log\frac{1}{1-a} \\
&\le \sup\left\{\frac{1}{4 + 2\mathrm{e}^z + 2\mathrm{e}^{-z}}\middle| z \in [A, B]\right\}\cdot\left|f - \log\frac{a}{1-a}\right|^2 \le \frac{1}{8}\cdot\left|f - \log\frac{a}{1-a}\right|^2.
\end{aligned}$$

Proof Consider the map $G : \mathbb{R} \to [0, \infty), z \mapsto a\phi(z) + (1-a)\phi(-z)$. Obviously G is twice continuously differentiable on $\mathbb{R}$ with $G'\left(\log\frac{a}{1-a}\right) = 0$ and $G''(z) = \frac{1}{2+\mathrm{e}^z+\mathrm{e}^{-z}}$ for any real number z. Then it follows from Taylor's theorem that there exists a real number ξ between $\log\frac{a}{1-a}$ and f, such that

$$\begin{aligned}
&a\phi(f) + (1-a)\phi(-f) - a\log\frac{1}{a} - (1-a)\log\frac{1}{1-a} = G(f) - G\left(\log\frac{a}{1-a}\right) \\
&= \left(f - \log\frac{a}{1-a}\right)\cdot G'\left(\log\frac{a}{1-a}\right) + \frac{G''(\xi)}{2}\cdot\left|f - \log\frac{a}{1-a}\right|^2 \\
&= \frac{G''(\xi)}{2}\cdot\left|f - \log\frac{a}{1-a}\right|^2 = \frac{\left|f - \log\frac{a}{1-a}\right|^2}{4 + 2\mathrm{e}^\xi + 2\mathrm{e}^{-\xi}}.
\end{aligned} \tag{C.33}$$

Since $A \le \min\left\{f, \log\frac{a}{1-a}\right\} \le \max\left\{f, \log\frac{a}{1-a}\right\} \le B$, we must have $\xi \in [A,B]$, which, together with (C.33), yields

$$\begin{aligned}
&\min\left\{\frac{1}{4+2\mathrm{e}^{A}+2\mathrm{e}^{-A}}, \frac{1}{4+2\mathrm{e}^{B}+2\mathrm{e}^{-B}}\right\}\cdot\left|f-\log\frac{a}{1-a}\right|^2 \\
&=\left(\inf_{t\in[A,B]}\frac{1}{4+2\mathrm{e}^{t}+\mathrm{e}^{-t}}\right)\cdot\left|f-\log\frac{a}{1-a}\right|^2 \le \frac{\left|f-\log\frac{a}{1-a}\right|^2}{4+2\mathrm{e}^{\xi}+2\mathrm{e}^{-\xi}} \\
&= a\phi(f)+(1-a)\phi(-f)-a\log\frac{1}{a}-(1-a)\log\frac{1}{1-a}=\frac{\left|f-\log\frac{a}{1-a}\right|^2}{4+2\mathrm{e}^{\xi}+2\mathrm{e}^{-\xi}} \\
&\le \sup\left\{\frac{1}{4+2\mathrm{e}^{z}+2\mathrm{e}^{-z}}\middle|z\in[A,B]\right\}\cdot\left|f-\log\frac{a}{1-a}\right|^2 \le \frac{1}{8}\cdot\left|f-\log\frac{a}{1-a}\right|^2.
\end{aligned} \tag{C.34}$$

This completes the proof. ■

Lemma C.7 *Let $\phi(t)=\log\left(1+\mathrm{e}^{-t}\right)$ be the logistic loss, f be a real number, $d\in\mathbb{N}$, and P be a Borel probability measure on $[0,1]^d\times\{-1,1\}$ of which the conditional probability function $[0,1]^d \ni z\mapsto P(\{1\}|z)\in[0,1]$ is denoted by η. Then for $x\in[0,1]^d$ such that $\eta(x)\notin\{0,1\}$, there holds*

$$\begin{aligned}
&\left|\inf_{t\in\left[f\wedge\log\frac{\eta(x)}{1-\eta(x)},f\vee\log\frac{\eta(x)}{1-\eta(x)}\right]}\frac{1}{2(2+\mathrm{e}^{t}+\mathrm{e}^{-t})}\right|\cdot\left|f-\log\frac{\eta(x)}{1-\eta(x)}\right|^2 \\
&\le \int_{\{-1,1\}}\left(\phi\left(yf\right)-\phi\left(y\log\frac{\eta(x)}{1-\eta(x)}\right)\right)\mathrm{d}P(y|x) \\
&\le \left|\sup_{t\in\left[f\wedge\log\frac{\eta(x)}{1-\eta(x)},f\vee\log\frac{\eta(x)}{1-\eta(x)}\right]}\frac{1}{2(2+\mathrm{e}^{t}+\mathrm{e}^{-t})}\right|\cdot\left|f-\log\frac{\eta(x)}{1-\eta(x)}\right|^2 \le \frac{1}{4}\left|f-\log\frac{\eta(x)}{1-\eta(x)}\right|^2.
\end{aligned}$$

Proof Given $x\in[0,1]^d$ such that $\eta(x)\notin\{0,1\}$, define

$$V_x:\mathbb{R}\to(0,\infty),\quad t\mapsto\eta(x)\phi(t)+(1-\eta(x))\phi(-t).$$

Then it is easy to verify that

$$\int_{\{-1,1\}}\phi\left(yt\right)\mathrm{d}P(y|x)=\phi(t)P(Y=1|X=x)+\phi(-t)P(Y=-1|X=x)=V_x(t)$$

for all $t\in\mathbb{R}$. Consequently,

$$\begin{aligned}
&\int_{\{-1,1\}}\left(\phi\left(yf\right)-\phi\left(y\log\frac{\eta(x)}{1-\eta(x)}\right)\right)\mathrm{d}P(y|x)=V_x(f)-V_x\left(\log\frac{\eta(x)}{1-\eta(x)}\right) \\
&=\eta(x)\phi(f)+(1-\eta(x))\phi(-f)-\eta(x)\log\frac{1}{\eta(x)}-(1-\eta(x))\log\frac{1}{1-\eta(x)}.
\end{aligned}$$

The desired inequalities then follow immediately by applying Lemma C.6. ■

Lemma C.8 *Let $\phi(t) = \log\left(1+\mathrm{e}^{-t}\right)$ be the logistic loss, $d \in \mathbb{N}$, $f : [0,1]^d \to \mathbb{R}$ be a measurable function, and P be a Borel probability measure on $[0,1]^d \times \{-1,1\}$ of which the conditional probability function $[0,1]^d \ni z \mapsto P(\{1\}\,|z) \in [0,1]$ is denoted by η. Assume that there exist constants $(a,b) \in \mathbb{R}^2$, $\delta \in (0,1/2)$, and a measurable function $\hat{\eta} : [0,1]^d \to \mathbb{R}$, such that $\hat{\eta} = \eta$, P_X-a.s.,*

$$\log\frac{\delta}{1-\delta} \leq f(x) \leq -a,\ \forall\, x \in [0,1]^d \text{ satisfying } 0 \leq \hat{\eta}(x) = \eta(x) < \delta,$$

and

$$b \leq f(x) \leq \log\frac{1-\delta}{\delta},\ \forall\, x \in [0,1]^d \text{ satisfying } 1-\delta < \hat{\eta}(x) = \eta(x) \leq 1.$$

Then

$$\begin{aligned}
&\mathcal{E}_P^{\phi}(f) - \phi(a)P_X(\Omega_2) - \phi(b)P_X(\Omega_3)\\
&\leq \int_{\Omega_1} \sup\left\{\left.\frac{\left|f(x)-\log\frac{\eta(x)}{1-\eta(x)}\right|^2}{2(2+\mathrm{e}^t+\mathrm{e}^{-t})}\right| t \in \left[f(x)\wedge\log\frac{\eta(x)}{1-\eta(x)}, f(x)\vee\log\frac{\eta(x)}{1-\eta(x)}\right]\right\}\mathrm{d}P_X(x)\\
&\leq \int_{\Omega_1}\left|f(x)-\log\frac{\eta(x)}{1-\eta(x)}\right|^2\mathrm{d}P_X(x),
\end{aligned}$$

where

$$\begin{aligned}
\Omega_1 &:= \left\{x \in [0,1]^d \,\middle|\, \delta \leq \hat{\eta}(x) = \eta(x) \leq 1-\delta\right\},\\
\Omega_2 &:= \left\{x \in [0,1]^d \,\middle|\, 0 \leq \hat{\eta}(x) = \eta(x) < \delta\right\},\\
\Omega_3 &:= \left\{x \in [0,1]^d \,\middle|\, 1-\delta < \hat{\eta}(x) = \eta(x) \leq 1\right\}.
\end{aligned} \tag{C.35}$$

Proof Define

$$\psi : [0,1]^d \times \{-1,1\} \to [0,\infty),$$
$$(x,y) \mapsto \begin{cases} \phi\left(y\log\frac{\eta(x)}{1-\eta(x)}\right), & \text{if } \eta(x) \in [\delta, 1-\delta],\\ 0, & \text{if } \eta(x) \in \{0,1\},\\ \eta(x)\log\frac{1}{\eta(x)} + (1-\eta(x))\log\frac{1}{1-\eta(x)}, & \text{if } \eta(x) \in (0,\delta)\cup(1-\delta,1).\end{cases}$$

Since $\hat{\eta} = \eta \in [0,1]$, P_X-a.s., we have that $P_X([0,1]^d \setminus (\Omega_1\cup\Omega_2\cup\Omega_3)) = 0$. Then it follows from lemma C.3 that

$$\begin{aligned}
&\mathcal{E}_P^{\phi}(f) = \mathcal{R}_P^{\phi}(f) - \inf\left\{\mathcal{R}_P^{\phi}(g)\,\middle|\, g : [0,1]^d \to \mathbb{R} \text{ is measurable}\right\}\\
&= \int_{[0,1]^d\times\{-1,1\}} \phi(yf(x))\mathrm{d}P(x,y) - \int_{[0,1]^d\times\{-1,1\}} \psi(x,y)\mathrm{d}P(x,y) = I_1 + I_2 + I_3,
\end{aligned} \tag{C.36}$$

where

$$I_i := \int_{\Omega_i\times\{-1,1\}} \left(\phi\left(yf(x)\right) - \psi(x,y)\right)\mathrm{d}P(x,y),\ i = 1,2,3.$$

According to Lemma C.7, we have

$$\begin{aligned} I_1 &= \int_{\Omega_1}\int_{\{-1,1\}}\left(\phi\left(yf(x)\right)-\phi\left(y\log\frac{\eta(x)}{1-\eta(x)}\right)\right)\mathrm{d}P(y|x)\mathrm{d}P_X(x) \\ &\leq \int_{\Omega_1}\sup\left\{\frac{\left|f(x)-\log\frac{\eta(x)}{1-\eta(x)}\right|^2}{2(2+\mathrm{e}^t+\mathrm{e}^{-t})}\,\middle|\, \begin{array}{c} t\in\left[f(x)\wedge\log\dfrac{\eta(x)}{1-\eta(x)},\infty\right) \text{ and} \\ t\in\left(-\infty, f(x)\vee\log\dfrac{\eta(x)}{1-\eta(x)}\right] \end{array}\right\}\mathrm{d}P_X(x). \end{aligned} \tag{C.37}$$

Then it remains to bound I_2 and I_3.

Indeed, for any $x\in\Omega_2$, if $\eta(x)=0$, then

$$\int_{\{-1,1\}}\left(\phi(yf(x))-\psi(x,y)\right)\mathrm{d}P(y|x)=\phi(-f(x))\leq\phi(a).$$

Otherwise, we have

$$\begin{aligned} &\int_{\{-1,1\}}\left(\phi(yf(x))-\psi(x,y)\right)\mathrm{d}P(y|x) \\ &=\left(\phi(f(x))-\log\frac{1}{\eta(x)}\right)\eta(x)+\left(\phi(-f(x))-\log\frac{1}{1-\eta(x)}\right)(1-\eta(x)) \\ &=\left(\phi\left(f(x)\right)-\phi\left(\log\frac{\eta(x)}{1-\eta(x)}\right)\right)\eta(x)+\left(\phi\left(-f(x)\right)-\phi\left(-\log\frac{\eta(x)}{1-\eta(x)}\right)\right)(1-\eta(x)) \\ &\leq\left(\phi\left(\log\frac{\delta}{1-\delta}\right)-\phi\left(\log\frac{\eta(x)}{1-\eta(x)}\right)\right)\eta(x)+\phi(-f(x))(1-\eta(x)) \\ &\leq\phi(-f(x))(1-\eta(x))\leq\phi(-f(x))\leq\phi(a). \end{aligned}$$

Therefore, no matter whether $\eta(x)=0$ or $\eta(x)\neq 0$, there always holds

$$\int_{\{-1,1\}}\left(\phi(yf(x))-\psi(x,y)\right)\mathrm{d}P(y|x)\leq\phi(a),$$

which means that

$$\begin{aligned} I_2 &= \int_{\Omega_2}\int_{\{-1,1\}}\left(\phi(yf(x))-\psi(x,y)\right)\mathrm{d}P(y|x)\mathrm{d}P_X(x) \\ &\leq\int_{\Omega_2}\phi(a)\mathrm{d}P_X(x)=\phi(a)P_X(\Omega_2). \end{aligned} \tag{C.38}$$

Similarly, for any $x\in\Omega_3$, if $\eta(x)=1$, then

$$\int_{\{-1,1\}}\left(\phi(yf(x))-\psi(x,y)\right)\mathrm{d}P(y|x)=\phi(f(x))\leq\phi(b).$$

Otherwise, we have

$$\int_{\{-1,1\}}\left(\phi(yf(x))-\psi(x,y)\right)\mathrm{d}P(y|x)$$

$$
\begin{aligned}
&= \left(\phi(f(x)) - \log\frac{1}{\eta(x)}\right)\eta(x) + \left(\phi(-f(x)) - \log\frac{1}{1-\eta(x)}\right)(1-\eta(x)) \\
&= \left(\phi\left(f(x)\right) - \phi\left(\log\frac{\eta(x)}{1-\eta(x)}\right)\right)\eta(x) + \left(\phi\left(-f(x)\right) - \phi\left(-\log\frac{\eta(x)}{1-\eta(x)}\right)\right)(1-\eta(x)) \\
&\leq \phi(f(x))\eta(x) + \left(\phi\left(\log\frac{\delta}{1-\delta}\right) - \phi\left(\log\frac{1-\eta(x)}{\eta(x)}\right)\right)(1-\eta(x)) \\
&\leq \phi(f(x))\eta(x) \leq \phi(f(x)) \leq \phi(b).
\end{aligned}
$$

Therefore, no matter whether $\eta(x) = 1$ or $\eta(x) \neq 1$, we have

$$\int_{\{-1,1\}} (\phi(yf(x)) - \psi(x,y))\,\mathrm{d}P(y|x) \leq \phi(b),$$

which means that

$$
\begin{aligned}
I_3 &= \int_{\Omega_3}\int_{\{-1,1\}} (\phi(yf(x)) - \psi(x,y))\,\mathrm{d}P(y|x)\mathrm{d}P_X(x) \\
&\leq \int_{\Omega_3} \phi(b)\mathrm{d}P_X(x) = \phi(b)P_X(\Omega_3).
\end{aligned}
\tag{C.39}
$$

The desired inequality then follows immediately from (C.37), (C.38), (C.39) and (C.36). Thus we complete the proof. ■

Lemma C.9 *Let* $\delta \in (0, 1/2)$, $a \in [\delta, 1-\delta]$, $f \in \left[-\log\frac{1-\delta}{\delta}, \log\frac{1-\delta}{\delta}\right]$, *and* $\phi(t) = \log(1+\mathrm{e}^{-t})$ *be the logistic loss. Then there hold*

$$H(a,f) \leq \Gamma \cdot G(a,f)$$

with $\Gamma = 5000\left|\log\delta\right|^2$,

$$H(a,f) := a\cdot\left|\phi(f) - \phi\left(\log\frac{a}{1-a}\right)\right|^2 + (1-a)\cdot\left|\phi(-f) - \phi\left(-\log\frac{a}{1-a}\right)\right|^2,$$

and

$$
\begin{aligned}
G(a,f) &:= a\phi(f) + (1-a)\phi(-f) - a\phi\left(\log\frac{a}{1-a}\right) - (1-a)\phi\left(-\log\frac{a}{1-a}\right) \\
&= a\phi(f) + (1-a)\phi(-f) - a\log\frac{1}{a} - (1-a)\log\frac{1}{1-a}.
\end{aligned}
$$

Proof In this proof, we will frequently use elementary inequalities

$$x\log\frac{1}{x} \leq \min\left\{1-x, (1-x)\cdot\log\frac{1}{1-x}\right\},\ \forall\, x \in [1/2, 1), \tag{C.40}$$

and

$$
\begin{aligned}
&-\log\frac{1}{1-x} - 2 < -\log 7 \leq -\log\left(\exp\left(\frac{3-3x}{x}\log\frac{1}{1-x}\right) - 1\right) \\
&< \log\frac{x}{1-x} < 2 + \log\frac{1}{1-x},\ \forall\, x \in [1/2, 1).
\end{aligned}
\tag{C.41}
$$

We first show that

$$G(a,f)\geq\frac{a\phi(f)}{3}\quad\text{provided } \frac{1}{2}\leq a\leq 1-\delta \text{ and } f\leq -\log\left(\exp\left(\frac{3-3a}{a}\log\frac{1}{1-a}\right)-1\right). \tag{C.42}$$

Indeed, if $1/2\leq a\leq 1-\delta$ and $f\leq -\log\left(\exp\left(\frac{3-3a}{a}\log\frac{1}{1-a}\right)-1\right)$, then

$$\begin{aligned}&\frac{2}{3}\cdot a\phi(f)\geq\frac{2}{3}\cdot a\phi\left(-\log\left(\exp\left(\frac{3-3a}{a}\log\frac{1}{1-a}\right)-1\right)\right)=(2-2a)\cdot\log\frac{1}{1-a}\\&\geq a\log\frac{1}{a}+(1-a)\log\frac{1}{1-a},\end{aligned}$$

which means that

$$G(a,f)\geq a\phi(f)-a\log\frac{1}{a}-(1-a)\log\frac{1}{1-a}\geq\frac{a\phi(f)}{3}.$$

This proves (C.42).

We next show that

$$G(a,f)\geq\frac{1-a}{18}\left|f-\log\frac{a}{1-a}\right|^2\quad\text{provided } \frac{1}{2}\leq a\leq 1-\delta \text{ and } -2-\log\frac{1}{1-a}\leq f\leq 2+\log\frac{1}{1-a}. \tag{C.43}$$

Indeed, if $1/2\leq a\leq 1-\delta$ and $-2-\log\frac{1}{1-a}\leq f\leq 2+\log\frac{1}{1-a}$, then it follows from Lemma C.6 that

$$\begin{aligned}G(a,f)&\geq\frac{\left|f-\log\frac{a}{1-a}\right|^2}{4+2\exp\left(2+\log\frac{1}{1-a}\right)+2\exp\left(-2-\log\frac{1}{1-a}\right)}\\&\geq\frac{\left|f-\log\frac{a}{1-a}\right|^2}{5+15\cdot\frac{1}{1-a}}\geq\frac{(1-a)\cdot\left|f-\log\frac{a}{1-a}\right|^2}{5-5a+15}\geq\frac{(1-a)\cdot\left|f-\log\frac{a}{1-a}\right|^2}{18},\end{aligned}$$

which proves (C.43).

We then show

$$H(a,f)\leq\Gamma\cdot G(a,f)\text{ provided } 1/2\leq a\leq 1-\delta\text{ and } -\log\frac{1-\delta}{\delta}\leq f\leq\log\frac{1-\delta}{\delta} \tag{C.44}$$

by considering the following four cases.

Case I. $1/2\leq a\leq 1-\delta$ and $2+\log\frac{1}{1-a}\leq f\leq\log\frac{1-\delta}{\delta}$. In this case we have

$$\begin{aligned}&\log\frac{1}{\delta}=\phi\left(\log\frac{\delta}{1-\delta}\right)\geq\phi(-f)=\log(1+\mathrm{e}^f)\geq f\geq 2+\log\frac{1}{1-a}\\&>\phi\left(-\log\frac{a}{1-a}\right)=\log\frac{1}{1-a}\geq\log\frac{1}{a}>0,\end{aligned} \tag{C.45}$$

which, together with (C.40), yields

$$a\log\frac{1}{a}+(1-a)\log\frac{1}{1-a}\le(1-a)\cdot\left(1+\log\frac{1}{1-a}\right)\le(1-a)\cdot\frac{1+\log\frac{1}{1-a}}{2+\log\frac{1}{1-a}}\cdot\phi(-f).$$

Consequently,

$$\begin{aligned}G(a,f)&\ge(1-a)\cdot\phi(-f)-a\log\frac{1}{a}-(1-a)\log\frac{1}{1-a}\\&\ge(1-a)\cdot\phi(-f)-(1-a)\cdot\frac{1+\log\frac{1}{1-a}}{2+\log\frac{1}{1-a}}\cdot\phi(-f)\\&=\frac{(1-a)\cdot\phi(-f)}{2+\log\frac{1}{1-a}}\ge\frac{(1-a)\cdot\phi(-f)}{4\log\frac{1}{\delta}}.\end{aligned}\tag{C.46}$$

On the other hand, it follows from $f\ge 2+\log\frac{1}{1-a}>\log\frac{a}{1-a}$ that

$$0\le\phi\left(\log\frac{a}{1-a}\right)-\phi(f)<\phi\left(\log\frac{a}{1-a}\right),$$

which, together with (C.40) and (C.45), yields

$$\begin{aligned}a\cdot\left|\phi(f)-\phi\left(\log\frac{a}{1-a}\right)\right|^2&\le a\cdot\left|\phi\left(\log\frac{a}{1-a}\right)\right|^2\\&=a\cdot\left|\log\frac{1}{a}\right|^2\le(1-a)\cdot\log\frac{1}{a}\le(1-a)\cdot\phi(-f).\end{aligned}\tag{C.47}$$

Besides, it follows from (C.46) that $0\le\phi(-f)-\phi\left(-\log\frac{a}{1-a}\right)\le\phi(-f)$. Consequently,

$$(1-a)\cdot\left|\phi(-f)-\phi\left(-\log\frac{a}{1-a}\right)\right|^2\le(1-a)\cdot\phi(-f)^2\le(1-a)\cdot\phi(-f)\cdot\log\frac{1}{\delta}.\tag{C.48}$$

Combining (C.46), (C.47) and (C.48), we deduce that

$$H(a,f)\le(1-a)\cdot\phi(-f)\cdot\left|1+\log\frac{1}{\delta}\right|\le(1-a)\cdot\phi(-f)\cdot\frac{\Gamma}{4\log\frac{1}{\delta}}\le\Gamma\cdot G(a,f),$$

which proves the desired inequality.

Case II. $1/2\le a\le 1-\delta$ and $-\log\left(\exp\left(\frac{3-3a}{a}\log\frac{1}{1-a}\right)-1\right)\le f<2+\log\frac{1}{1-a}$. In this case, we have $-2-\log\frac{1}{1-a}\le f\le 2+\log\frac{1}{1-a}$, where we have used (C.41). Therefore, it follows from (C.43) that $G(a,f)\ge\frac{1-a}{18}\left|f-\log\frac{a}{1-a}\right|^2$. On the other hand, it follow from (C.41) and Taylor's Theorem that there exists

$$\begin{aligned}&-\log 7\le-\log\left(\exp\left(\frac{3-3a}{a}\log\frac{1}{1-a}\right)-1\right)\\&\le f\wedge\log\frac{a}{1-a}\le\xi\le f\vee\log\frac{a}{1-a}\le 2+\log\frac{1}{1-a},\end{aligned}$$

such that

$$
\begin{aligned}
& a\cdot\left|\phi(f)-\phi\left(\log\frac{a}{1-a}\right)\right|^2\\
& = a\cdot\left|\phi'(\xi)\right|^2\cdot\left|f-\log\frac{a}{1-a}\right|^2\leq a\cdot\mathrm{e}^{-2\xi}\cdot\left|f-\log\frac{a}{1-a}\right|^2\\
& \leq a\cdot\exp(\log 7)\cdot\exp\left(\log\left(\exp\left(\frac{3-3a}{a}\log\frac{1}{1-a}\right)-1\right)\right)\cdot\left|f-\log\frac{a}{1-a}\right|^2\\
& = 7a\cdot\int_0^{\frac{3-3a}{a}\log\frac{1}{1-a}}\mathrm{e}^t\mathrm{d}t\cdot\left|f-\log\frac{a}{1-a}\right|^2\\
& \leq 7a\cdot\left|\frac{3-3a}{a}\log\frac{1}{1-a}\right|\cdot\exp\left(\frac{3-3a}{a}\log\frac{1}{1-a}\right)\cdot\left|f-\log\frac{a}{1-a}\right|^2\\
& \leq 7a\cdot\left|\frac{3-3a}{a}\log\frac{1}{1-a}\right|\cdot(1+\exp(\log 7))\cdot\left|f-\log\frac{a}{1-a}\right|^2\\
& \leq 168\cdot\left|(1-a)\cdot\log\frac{1}{1-a}\right|\cdot\left|f-\log\frac{a}{1-a}\right|^2\\
& \leq 168\cdot\left|(1-a)\cdot\log\frac{1}{\delta}\right|\cdot\left|f-\log\frac{a}{1-a}\right|^2.
\end{aligned}
\tag{C.49}
$$

Besides, we have

$$
\begin{aligned}
& (1-a)\cdot\left|\phi(-f)-\phi\left(-\log\frac{a}{1-a}\right)\right|^2\\
& \leq|1-a|\cdot\left\|\phi'\right\|_{\mathbb{R}}\cdot\left|f-\log\frac{a}{1-a}\right|^2\leq|1-a|\cdot\left|f-\log\frac{a}{1-a}\right|^2.
\end{aligned}
\tag{C.50}
$$

Combining (C.49), (C.50) and the fact that $G(a,f)\geq\frac{1-a}{18}\left|f-\log\frac{a}{1-a}\right|^2$, we deduce that

$$
\begin{aligned}
& H(a,f)\leq 168\cdot\left|(1-a)\cdot\log\frac{1}{\delta}\right|\cdot\left|f-\log\frac{a}{1-a}\right|^2+|1-a|\cdot\left|f-\log\frac{a}{1-a}\right|^2\\
& \leq 170\cdot\left|(1-a)\cdot\log\frac{1}{\delta}\right|\cdot\left|f-\log\frac{a}{1-a}\right|^2\leq\Gamma\cdot\frac{1-a}{18}\cdot\left|f-\log\frac{a}{1-a}\right|^2\leq\Gamma\cdot G(a,f),
\end{aligned}
$$

which proves the desired inequality.

Case III. $1/2\leq a\leq 1-\delta$ and $-\log\frac{a}{1-a}\leq f<-\log\left(\exp\left(\frac{3-3a}{a}\log\frac{1}{1-a}\right)-1\right)$. In this case, we still have (C.50). Besides, it follows from (C.42) that $G(a,f)\geq\frac{a\phi(f)}{3}$. Moreover, by (C.41) we obtain $-2-\log\frac{1}{1-a}<f<2+\log\frac{1}{1-a}$, which, together with (C.43), yields $G(a,f)\geq\frac{1-a}{18}\left|f-\log\frac{a}{1-a}\right|^2$. In addition, since $f<-\log\left(\exp\left(\frac{3-3a}{a}\log\frac{1}{1-a}\right)-1\right)\leq$

$\log \frac{a}{1-a}$, we have that $0 < \phi(f) - \phi\left(\log \frac{a}{1-a}\right) < \phi(f)$, which means that

$$\begin{aligned} & a \cdot \left|\phi(f) - \phi\left(\log \frac{a}{1-a}\right)\right|^2 \leq a \cdot |\phi(f)|^2 \\ & \leq a\phi(f)\phi\left(-\log \frac{a}{1-a}\right) = a\phi(f)\log\frac{1}{1-a} \leq a\phi(f)\log\frac{1}{\delta}. \end{aligned} \tag{C.51}$$

Combining all these inequalities, we obtain

$$\begin{aligned} H(a,f) &\leq a\phi(f)\cdot\left|\log\frac{1}{\delta}\right| + |1-a|\cdot\left|f - \log\frac{a}{1-a}\right|^2 \\ &\leq \frac{\Gamma a\phi(f)}{6} + \Gamma\cdot\frac{1-a}{36}\cdot\left|f - \log\frac{a}{1-a}\right|^2 \\ &\leq \frac{\Gamma\cdot G(a,f)}{2} + \frac{\Gamma\cdot G(a,f)}{2} = \Gamma\cdot G(a,f), \end{aligned}$$

which proves the desired inequality.

Case IV. $-\log\frac{1-\delta}{\delta} \leq f < \min\left\{-\log\frac{a}{1-a}, -\log\left(\exp\left(\frac{3-3a}{a}\log\frac{1}{1-a}\right)-1\right)\right\}$ and $1/2 \leq a \leq 1-\delta$. In this case, we still have $G(a,f) \geq \frac{a\phi(f)}{3}$ according to (C.42). Besides, it follows from

$$f < \min\left\{-\log\frac{a}{1-a}, -\log\left(\exp\left(\frac{3-3a}{a}\log\frac{1}{1-a}\right)-1\right)\right\} \leq -\log\frac{a}{1-a} \leq \log\frac{a}{1-a}$$

that

$$\begin{aligned} 0 &\leq \min\left\{\phi\left(-\log\frac{a}{1-a}\right) - \phi(-f), \phi(f) - \phi\left(\log\frac{a}{1-a}\right)\right\} \\ &\leq \max\left\{\phi\left(-\log\frac{a}{1-a}\right) - \phi(-f), \phi(f) - \phi\left(\log\frac{a}{1-a}\right)\right\} \\ &\leq \max\left\{\phi\left(-\log\frac{a}{1-a}\right), \phi(f)\right\} = \phi(f). \end{aligned} \tag{C.52}$$

Combining (C.52) and the fact that $G(a,f) \geq \frac{a\phi(f)}{3}$, we deduce that

$$\begin{aligned} H(a,f) &\leq a\cdot|\phi(f)|^2 + (1-a)\cdot|\phi(f)|^2 \leq \phi(f)\phi\left(-\log\frac{1-\delta}{\delta}\right) \\ &= \phi(f)\log\frac{1}{\delta} \leq \frac{\Gamma a\phi(f)}{3} \leq \Gamma\cdot G(a,f), \end{aligned}$$

which proves the desired inequality.

Combining all these four cases, we conclude that (C.44) has been proved. Furthermore, (C.44) yields that

$$H(a,f) = H(1-a,-f) \leq \Gamma\cdot G(1-a,-f) = \Gamma\cdot G(a,f)$$

provided $\delta \leq a \leq 1/2$ and $-\log\frac{1-\delta}{\delta} \leq f \leq \log\frac{1-\delta}{\delta}$, which, together with (C.44), proves this lemma. ∎

Lemma C.10 *Let $\phi(t) = \log\left(1 + \mathrm{e}^{-t}\right)$ be the logistic loss, $\delta_0 \in (0, 1/3)$, $d \in \mathbb{N}$ and P be a Borel probability measure on $[0,1]^d \times \{-1,1\}$ of which the conditional probability function $[0,1]^d \ni z \mapsto P(\{1\}\,|z) \in [0,1]$ is denoted by η. Then there exists a measurable function*

$$\psi : [0,1]^d \times \{-1,1\} \to \left[0, \log \frac{10\log(1/\delta_0)}{\delta_0}\right]$$

such that

$$\int_{[0,1]^d\times\{-1,1\}} \psi\,(x,y)\mathrm{d}P(x,y) = \inf\left\{\mathcal{R}_P^{\phi}(g)\middle|\, g : [0,1]^d \to \mathbb{R} \text{ is measurable}\right\} \tag{C.53}$$

and

$$\begin{aligned}&\int_{[0,1]^d\times\{-1,1\}} (\phi\,(yf(x)) - \psi(x,y))^2 \mathrm{d}P(x,y)\\ &\leq 125000\,|\log\delta_0|^2 \cdot \int_{[0,1]^d\times\{-1,1\}} (\phi\,(yf(x)) - \psi(x,y))\,\mathrm{d}P(x,y)\end{aligned} \tag{C.54}$$

for any measurable $f : [0,1]^d \to \left[\log\frac{\delta_0}{1-\delta_0}, \log\frac{1-\delta_0}{\delta_0}\right]$.

Proof Let

$$H : [0,1] \to [0,\infty), \quad t \mapsto \begin{cases} t\log\left(\dfrac{1}{t}\right) + (1-t)\log\left(\dfrac{1}{1-t}\right), & \text{if } \in (0,1),\\ 0, & \text{if } t \in \{0,1\}.\end{cases}$$

Then it is easy to show that $H\left(\frac{\delta_0}{10\log(1/\delta_0)}\right) \leq \frac{4}{5}\log\left(\frac{1}{1-\delta_0}\right) \leq H\left(\frac{\delta_0}{\log(1/\delta_0)}\right)$. Thus there exists $\delta_1 \in \left(0, \frac{1}{3}\right)$ such that

$$H(\delta_1) \leq \frac{4}{5}\log\left(\frac{1}{1-\delta_0}\right)$$

and

$$0 < \frac{\delta_0}{10\log\left(1/\delta_0\right)} \leq \delta_1 \leq \frac{\delta_0}{\log(1/\delta_0)} \leq \delta_0 < 1/3.$$

Take

$$\psi : [0,1]^d \times \{-1,1\} \to \mathbb{R}, \quad (x,y) \mapsto \begin{cases} \phi\left(y\log\dfrac{\eta(x)}{1-\eta(x)}\right), & \text{if } \eta(x) \in [\delta_1, 1-\delta_1],\\ H(\eta(x)), & \text{if } \eta(x) \notin [\delta_1, 1-\delta_1],\end{cases}$$

which can be further expressed as

$$\begin{aligned}&\psi : [0,1]^d \times \{-1,1\} \to \mathbb{R},\\ &(x,y) \mapsto \begin{cases} \phi\left(y\log\dfrac{\eta(x)}{1-\eta(x)}\right), & \text{if } \eta(x) \in [\delta_1, 1-\delta_1],\\ 0, & \text{if } \eta(x) \in \{0,1\},\\ \eta(x)\log\dfrac{1}{\eta(x)} + (1-\eta(x))\log\dfrac{1}{1-\eta(x)}, & \text{if } \eta(x) \in (0,\delta_1) \cup (1-\delta_1, 1).\end{cases}\end{aligned}$$

Obviously, ψ is a measurable function such that

$$0 \le \psi(x,y) \le \log\frac{1}{\delta_1} \le \log\frac{10\log(1/\delta_0)}{\delta_0}, \quad \forall\,(x,y)\in[0,1]^d\times\{-1,1\},$$

and it follows immediately from Lemma C.3 that (C.53) holds. We next show (C.54).

For any measurable function $f:[0,1]^d\to\left[\log\frac{\delta_0}{1-\delta_0},\log\frac{1-\delta_0}{\delta_0}\right]$ and any $x\in[0,1]^d$, if $\eta(x)\notin[\delta_1,1-\delta_1]$, then we have

$$\begin{aligned}0 &\le \psi(x,y) = H(\eta(x)) \le H(\delta_1) \le \frac{4}{5}\log\frac{1}{1-\delta_0}\\ &= \frac{4}{5}\phi\left(\log\frac{1-\delta_0}{\delta_0}\right) \le \frac{4}{5}\phi(yf(x)) \le \phi(yf(x)), \quad \forall\, y\in\{-1,1\}.\end{aligned}$$

Hence $0\le\frac{1}{5}\phi(yf(x))\le\phi(yf(x))-\psi(x,y)\le\phi(yf(x))$, $\forall\, y\in\{-1,1\}$, which means that

$$\begin{aligned}&(\phi(yf(x))-\psi(x,y))^2 \le \phi(yf(x))^2 \le \phi(yf(x))\phi\left(-\log\frac{1-\delta_0}{\delta_0}\right)\\ &= \frac{1}{5}\phi(yf(x))\cdot 5\log\frac{1}{\delta_0} \le (\phi(yf(x))-\psi(x,y))\cdot 5000\left|\log\delta_1\right|^2, \quad \forall\, y\in\{-1,1\}.\end{aligned}$$

Integrating both sides with respect to y, we obtain

$$\begin{aligned}&\int_{\{-1,1\}}(\phi(yf(x))-\psi(x,y))^2\,\mathrm{d}P(y|x)\\ &\le 5000\left|\log\delta_1\right|^2\cdot\int_{\{-1,1\}}(\phi(yf(x))-\psi(x,y))\,\mathrm{d}P(y|x).\end{aligned} \tag{C.55}$$

If $\eta(x)\in[\delta_1,1-\delta_1]$, then it follows from Lemma C.9 that

$$\begin{aligned}&\int_{\{-1,1\}}(\phi(yf(x))-\psi(x,y))^2\,\mathrm{d}P(y|x)\\ &= \eta(x)\left|\phi(f(x))-\phi\left(\log\frac{\eta(x)}{1-\eta(x)}\right)\right|^2 + (1-\eta(x))\left|\phi(-f(x))-\phi\left(-\log\frac{\eta(x)}{1-\eta(x)}\right)\right|^2\\ &\le 5000\left|\log\delta_1\right|^2\cdot\Bigg(\eta(x)\phi(f(x))+(1-\eta(x))\phi(-f(x))\\ &\qquad\qquad -\eta(x)\phi\Big(\log\frac{\eta(x)}{1-\eta(x)}\Big)-(1-\eta(x))\phi\Big(-\log\frac{\eta(x)}{1-\eta(x)}\Big)\Bigg)\\ &= 5000\left|\log\delta_1\right|^2\int_{\{-1,1\}}(\phi(yf(x))-\psi(x,y))\,\mathrm{d}P(y|x),\end{aligned}$$

which means that (C.55) still holds. Therefore, (C.55) holds for all $x\in[0,1]^d$. We then integrate both sides of (C.55) with respect to x and obtain

$$\int_{[0,1]^d\times\{-1,1\}}(\phi(yf(x))-\psi(x,y))^2\mathrm{d}P(x,y)$$

$$\le 5000\,|\log\delta_1|^2\int_{[0,1]^d\times\{-1,1\}}(\phi(yf(x))-\psi(x,y))\,\mathrm{d}P(x,y)$$
$$\le 125000\,|\log\delta_0|^2\int_{[0,1]^d\times\{-1,1\}}(\phi(yf(x))-\psi(x,y))\,\mathrm{d}P(x,y),$$

which yields (C.54). In conclusion, the function ψ defined above has all the desired properties. Thus we complete the proof. ■

The following Lemma C.11 is similar to Lemma 3 of Schmidt-Hieber (2020).

Lemma C.11 *Let $(d,d_\star,d_*,K)\in\mathbb{N}^4$, $\beta\in(0,\infty)$, $r\in[1,\infty)$, and $q\in\mathbb{N}\cup\{0\}$. Suppose $h_0,h_1,\ldots,h_q,\tilde h_0,\tilde h_1,\ldots,\tilde h_q$ are functions satisfying that*

(i) $\mathbf{dom}(h_i)=\mathbf{dom}(\tilde h_i)=[0,1]^K$ *for* $0<i\le q$ *and* $\mathbf{dom}(h_0)=\mathbf{dom}(\tilde h_0)=[0,1]^d$*;*

(ii) $\mathbf{ran}(h_i)\cup\mathbf{ran}(\tilde h_i)\subset[0,1]^K$ *for* $0\le i<q$ *and* $\mathbf{ran}(h_q)\cup\mathbf{ran}(\tilde h_q)\subset\mathbb{R}$*;*

(iii) $h_q\in\mathcal{G}_\infty^{\mathbf{H}}(d_*,\beta,r)\cup\mathcal{G}_\infty^{\mathbf{M}}(d_\star)$*;*

(iv) For $0\le i<q$ and $1\le j\le K$, the j-th coordinate function of h_i given by $\mathbf{dom}(h_i)\ni x\mapsto(h_i(x))_j\in\mathbb{R}$ belongs to $\mathcal{G}_\infty^{\mathbf{H}}(d_*,\beta,r)\cup\mathcal{G}_\infty^{\mathbf{M}}(d_\star)$.

Then there holds

$$\begin{aligned}&\left\|h_q\circ h_{q-1}\circ\cdots\circ h_1\circ h_0-\tilde h_q\circ\tilde h_{q-1}\circ\cdots\circ\tilde h_1\circ\tilde h_0\right\|_{[0,1]^d}\\&\le\left|r\cdot d_*^{1\wedge\beta}\right|^{\sum_{k=0}^{q-1}(1\wedge\beta)^k}\cdot\sum_{k=0}^{q}\left\|\tilde h_k-h_k\right\|_{\mathbf{dom}(h_k)}^{(1\wedge\beta)^{q-k}}.\end{aligned}\tag{C.56}$$

Proof We will prove this lemma by induction on q. The case $q=0$ is trivial. Now assume that $q>0$ and that the desired result holds for $q-1$. Consider the case q. For each $0\le i<q$ and $1\le j\le K$, denote

$$\tilde h_{i,j}:\mathbf{dom}(\tilde h_i)\to\mathbb{R},\ \ x\mapsto\big(\tilde h_i(x)\big)_j,$$

and

$$h_{i,j}:\mathbf{dom}(h_i)\to\mathbb{R},\ \ x\mapsto\big(h_i(x)\big)_j.$$

Obviously, $\mathbf{ran}(\tilde h_{i,j})\cup\mathbf{ran}(h_{i,j})\subset[0,1]$. By induction hypothesis (that is, the case $q-1$ of this lemma), we have that

$$\begin{aligned}&\left\|h_{q-1,j}\circ h_{q-2}\circ h_{q-3}\circ\cdots\circ h_0-\tilde h_{q-1,j}\circ\tilde h_{q-2}\circ\tilde h_{q-3}\circ\cdots\circ\tilde h_0\right\|_{[0,1]^d}\\&\le\left|r\cdot d_*^{1\wedge\beta}\right|^{\sum_{k=0}^{q-2}(1\wedge\beta)^k}\cdot\left(\left\|\tilde h_{q-1,j}-h_{q-1,j}\right\|_{\mathbf{dom}(h_{q-1,j})}+\sum_{k=0}^{q-2}\left\|\tilde h_k-h_k\right\|_{\mathbf{dom}(h_k)}^{(1\wedge\beta)^{q-1-k}}\right)\end{aligned}$$

$$\leq \left| r \cdot d_*^{1\wedge\beta} \right|^{\sum_{k=0}^{q-2}(1\wedge\beta)^k} \cdot \sum_{k=0}^{q-1} \left\| \tilde{h}_k - h_k \right\|_{\mathbf{dom}(h_k)}^{(1\wedge\beta)^{q-1-k}}, \; \forall\, j \in \mathbb{Z} \cap (0, K].$$

Therefore,

$$\begin{aligned}
&\left\| h_{q-1} \circ h_{q-2} \circ h_{q-3} \circ \cdots \circ h_0 - \tilde{h}_{q-1} \circ \tilde{h}_{q-2} \circ \tilde{h}_{q-3} \circ \cdots \circ \tilde{h}_0 \right\|_{[0,1]^d} \\
&= \sup_{j \in \mathbb{Z}\cap(0,K]} \left\| h_{q-1,j} \circ h_{q-2} \circ h_{q-3} \circ \cdots \circ h_0 - \tilde{h}_{q-1,j} \circ \tilde{h}_{q-2} \circ \tilde{h}_{q-3} \circ \cdots \circ \tilde{h}_0 \right\|_{[0,1]^d} \\
&\leq \left| r \cdot d_*^{1\wedge\beta} \right|^{\sum_{k=0}^{q-2}(1\wedge\beta)^k} \cdot \sum_{k=0}^{q-1} \left\| \tilde{h}_k - h_k \right\|_{\mathbf{dom}(h_k)}^{(1\wedge\beta)^{q-1-k}}.
\end{aligned} \tag{C.57}$$

We next show that

$$\left| h_q(x) - h_q(x') \right| \leq r \cdot d_*^{1\wedge\beta} \cdot \left\| x - x' \right\|_\infty^{1\wedge\beta}, \; \forall\, x, x' \in [0,1]^K \tag{C.58}$$

by considering three cases.

Case I: $h_q \in \mathcal{G}_\infty^{\mathbf{H}}(d_*, \beta, r)$ and $\beta > 1$. In this case, we must have that $h_q \in \mathcal{G}_K^{\mathbf{H}}(d_*, \beta, r)$ since $\mathbf{dom}(h_q) = [0,1]^K$. Therefore, there exist $I \subset \{1, 2, \ldots, K\}$ and $g \in \mathcal{B}_r^\beta\left([0,1]^{d_*}\right)$ such that $\#(I) = d_*$ and $h_q(x) = g((x)_I)$ for all $x \in [0,1]^K$. Denote $\lambda := \beta + 1 - \lceil \beta \rceil$. We then use Taylor's formula to deduce that

$$\begin{aligned}
&\left| h_q(x) - h_q(x') \right| = \left| g((x)_I) - g((x')_I) \right| \overset{\exists\, \xi \in [0,1]^{d_*}}{=\!=\!=\!=} \left| \nabla g(\xi) \cdot \left((x)_I - (x')_I \right) \right| \\
&\leq \left\| \nabla g(\xi) \right\|_\infty \cdot \left\| (x)_I - (x')_I \right\|_1 \leq \left\| \nabla g \right\|_{[0,1]^d} \cdot d_* \cdot \left\| (x)_I - (x')_I \right\|_\infty \\
&\leq \left\| g \right\|_{\mathcal{C}^{\beta-\lambda,\lambda}([0,1]^d)} \cdot d_* \cdot \left\| (x)_I - (x')_I \right\|_\infty \leq r \cdot d_* \cdot \left\| (x)_I - (x')_I \right\|_\infty \\
&\leq r \cdot d_*^{1\wedge\beta} \cdot \left\| x - x' \right\|_\infty^{1\wedge\beta}, \; \forall\, x, x' \in [0,1]^K,
\end{aligned}$$

which yields (C.58).

Case II: $h_q \in \mathcal{G}_\infty^{\mathbf{H}}(d_*, \beta, r)$ and $\beta \leq 1$. In this case, we still have that $h_q \in \mathcal{G}_K^{\mathbf{H}}(d_*, \beta, r)$. Therefore, there exist $I \subset \{1, 2, \ldots, K\}$ and $g \in \mathcal{B}_r^\beta\left([0,1]^{d_*}\right)$ such that $\#(I) = d_*$ and $h_q(x) = g((x)_I)$ for all $x \in [0,1]^K$. Consequently,

$$\begin{aligned}
&\left| h_q(x) - h_q(x') \right| = \left| g((x)_I) - g((x')_I) \right| \leq \left\| (x)_I - (x')_I \right\|_2^\beta \cdot \sup_{[0,1]^{d_*} \ni z \neq z' \in [0,1]^{d_*}} \frac{\left| g(z) - g(z') \right|}{\left\| z - z' \right\|_2^\beta} \\
&\leq \left\| (x)_I - (x')_I \right\|_2^\beta \cdot \left\| g \right\|_{\mathcal{C}^{0,\beta}([0,1]^d)} \leq \left\| (x)_I - (x')_I \right\|_2^\beta \cdot r \leq r \cdot \left| \sqrt{d_*} \cdot \left\| x - x' \right\|_\infty \right|^\beta \\
&\leq r \cdot d_*^{1\wedge\beta} \cdot \left\| x - x' \right\|_\infty^{1\wedge\beta}, \; \forall\, x, x' \in [0,1]^K,
\end{aligned}$$

which yields (C.58).

Case III: $h_q \in \mathcal{G}_\infty^{\mathbf{M}}(d_\star)$. In this case, we have that there exists $I \subset \{1, 2, \ldots, K\}$ such that $1 \leq \#(I) \leq d_\star$ and $h_q(x) = \max\left\{ (x)_i \middle| i \in I \right\}$ for all $x \in [0,1]^K$. Consequently,

$$\begin{aligned}
&\left| h_q(x) - h_q(x') \right| = \left| \max\left\{ (x)_i \middle| i \in I \right\} - \max\left\{ (x')_i \middle| i \in I \right\} \right| \leq \left\| (x)_I - (x')_I \right\|_\infty \\
&\leq r \cdot d_*^{1\wedge\beta} \cdot \left\| x - x' \right\|_\infty \leq r \cdot d_*^{1\wedge\beta} \cdot \left\| x - x' \right\|_\infty^{1\wedge\beta}, \; \forall\, x, x' \in [0,1]^K,
\end{aligned}$$

which yields (C.58).

Combining the above three cases, we deduce that (C.58) always holds true. From (C.58) and (C.57) we obtain that

$$\begin{aligned}
&\left|h_q \circ h_{q-1} \circ \cdots \circ h_0(x) - \tilde{h}_q \circ \tilde{h}_{q-1} \circ \cdots \circ \tilde{h}_0(x)\right| \\
&\leq \left|h_q \circ h_{q-1} \circ \cdots \circ h_0(x) - h_q \circ \tilde{h}_{q-1} \circ \cdots \circ \tilde{h}_0(x)\right| \\
&\quad + \left|h_q \circ \tilde{h}_{q-1} \circ \cdots \circ \tilde{h}_0(x) - \tilde{h}_q \circ \tilde{h}_{q-1} \circ \cdots \circ \tilde{h}_0(x)\right| \\
&\leq r \cdot d_*^{1\wedge\beta} \cdot \left\|h_{q-1} \circ \cdots \circ h_0(x) - \tilde{h}_{q-1} \circ \cdots \circ \tilde{h}_0(x)\right\|_\infty^{1\wedge\beta} + \left\|h_q - \tilde{h}_q\right\|_{\mathbf{dom}(h_q)} \\
&\leq r \cdot d_*^{1\wedge\beta} \cdot \left\|h_{q-1} \circ \cdots \circ h_0 - \tilde{h}_{q-1} \circ \cdots \circ \tilde{h}_0\right\|_{[0,1]^d}^{1\wedge\beta} + \left\|h_q - \tilde{h}_q\right\|_{\mathbf{dom}(h_q)} \\
&\leq r \cdot d_*^{1\wedge\beta} \cdot \left|\left|r \cdot d_*^{1\wedge\beta}\right|^{\sum_{k=0}^{q-2}(1\wedge\beta)^k} \cdot \sum_{k=0}^{q-1} \left\|\tilde{h}_k - h_k\right\|_{\mathbf{dom}(h_k)}^{(1\wedge\beta)^{q-1-k}}\right|^{1\wedge\beta} + \left\|h_q - \tilde{h}_q\right\|_{\mathbf{dom}(h_q)} \\
&= r \cdot d_*^{1\wedge\beta} \cdot \left|r \cdot d_*^{1\wedge\beta}\right|^{\sum_{k=0}^{q-2}(1\wedge\beta)^{k+1}} \cdot \left|\sum_{k=0}^{q-1} \left\|\tilde{h}_k - h_k\right\|_{\mathbf{dom}(h_k)}^{(1\wedge\beta)^{q-1-k}}\right|^{1\wedge\beta} + \left\|h_q - \tilde{h}_q\right\|_{\mathbf{dom}(h_q)} \\
&\leq r \cdot d_*^{1\wedge\beta} \cdot \left|r \cdot d_*^{1\wedge\beta}\right|^{\sum_{k=0}^{q-2}(1\wedge\beta)^{k+1}} \cdot \sum_{k=0}^{q-1} \left|\left\|\tilde{h}_k - h_k\right\|_{\mathbf{dom}(h_k)}^{(1\wedge\beta)^{q-1-k}}\right|^{1\wedge\beta} + \left\|h_q - \tilde{h}_q\right\|_{\mathbf{dom}(h_q)} \\
&\leq r \cdot d_*^{1\wedge\beta} \cdot \left|r \cdot d_*^{1\wedge\beta}\right|^{\sum_{k=0}^{q-2}(1\wedge\beta)^{k+1}} \cdot \left|\sum_{k=0}^{q-1} \left|\left\|\tilde{h}_k - h_k\right\|_{\mathbf{dom}(h_k)}^{(1\wedge\beta)^{q-1-k}}\right|^{1\wedge\beta} + \left\|h_q - \tilde{h}_q\right\|_{\mathbf{dom}(h_q)}\right| \\
&= \left|r \cdot d_*^{1\wedge\beta}\right|^{\sum_{k=0}^{q-1}(1\wedge\beta)^k} \cdot \sum_{k=0}^{q} \left\|\tilde{h}_k - h_k\right\|_{\mathbf{dom}(h_k)}^{(1\wedge\beta)^{q-k}}, \ \forall\, x \in [0,1]^d.
\end{aligned}$$

Therefore,

$$\begin{aligned}
&\left\|h_q \circ h_{q-1} \circ \cdots \circ h_1 \circ h_0 - \tilde{h}_q \circ \tilde{h}_{q-1} \circ \cdots \circ \tilde{h}_1 \circ \tilde{h}_0\right\|_{[0,1]^d} \\
&= \sup_{x\in[0,1]^d} \left|h_q \circ h_{q-1} \circ \cdots \circ h_1 \circ h_0(x) - \tilde{h}_q \circ \tilde{h}_{q-1} \circ \cdots \circ \tilde{h}_1 \circ \tilde{h}_0(x)\right| \\
&\leq \left|r \cdot d_*^{1\wedge\beta}\right|^{\sum_{k=0}^{q-1}(1\wedge\beta)^k} \cdot \sum_{k=0}^{q} \left\|\tilde{h}_k - h_k\right\|_{\mathbf{dom}(h_k)}^{(1\wedge\beta)^{q-k}},
\end{aligned}$$

meaning that the desired result holds for q.

In conclusion, according to mathematical induction, we have that the desired result holds for all $q \in \mathbb{N} \cup \{0\}$. This completes the proof. ■

Lemma C.12 *Let k be an positive integer. Then there exists a neural network*

$$\tilde{f} \in \mathcal{F}_k^{\mathbf{FNN}}\left(1 + 2 \cdot \left\lceil \frac{\log k}{\log 2} \right\rceil, 2k, 26 \cdot 2^{\left\lceil \frac{\log k}{\log 2} \right\rceil} - 20 - 2 \cdot \left\lceil \frac{\log k}{\log 2} \right\rceil, 1, 1\right)$$

such that

$$\tilde{f}(x) = \|x\|_\infty\,,\ \forall\, x \in \mathbb{R}^k.$$

Proof We argue by induction.

Firstly, consider the case $k = 1$. Define

$$\tilde{f}_1 : \mathbb{R} \to \mathbb{R}, x \mapsto \sigma(x) + \sigma(-x).$$

Obviously,

$$\begin{aligned}\tilde{f}_1 &\in \mathcal{F}_1^{\mathbf{FNN}}(1,2,6,1,1)\\ &\subset \mathcal{F}_1^{\mathbf{FNN}}\left(1 + 2\cdot\left\lceil\frac{\log 1}{\log 2}\right\rceil, 2\cdot 1, 26\cdot 2^{\left\lceil\frac{\log 1}{\log 2}\right\rceil} - 20 - 2\cdot\left\lceil\frac{\log 1}{\log 2}\right\rceil, 1, 1\right)\end{aligned}$$

and $\tilde{f}(x) = \sigma(x) + \sigma(-x) = |x| = \|x\|_\infty$ for all $x \in \mathbb{R} = \mathbb{R}^1$. This proves the $k = 1$ case.

Now assume that the desired result holds for $k = 1, 2, 3, \ldots, m-1$ ($m \geq 2$), and consider the case $k = m$. Define

$$\begin{aligned}\tilde{g}_1 : \mathbb{R}^m &\to \mathbb{R}^{\lfloor \frac{m}{2} \rfloor},\\ x &\mapsto \left((x)_1, (x)_2, \cdots, (x)_{\lfloor \frac{m}{2}\rfloor - 1}, (x)_{\lfloor \frac{m}{2}\rfloor}\right),\end{aligned}$$

$$\begin{aligned}\tilde{g}_2 : \mathbb{R}^m &\to \mathbb{R}^{\lceil \frac{m}{2} \rceil},\\ x &\mapsto \left((x)_{\lfloor \frac{m}{2}\rfloor + 1}, (x)_{\lfloor \frac{m}{2}\rfloor + 2}, \cdots, (x)_{m-1}, (x)_m\right),\end{aligned}$$

and

$$\begin{aligned}\tilde{f}_m : \mathbb{R}^m &\to \mathbb{R},\\ x &\mapsto \sigma\left(\frac{1}{2}\cdot\sigma\Big(\tilde{f}_{\lfloor\frac{m}{2}\rfloor}(\tilde{g}_1(x))\Big) - \frac{1}{2}\cdot\sigma\Big(\tilde{f}_{\lceil\frac{m}{2}\rceil}(\tilde{g}_2(x))\Big)\right)\\ &\quad + \sigma\left(\frac{1}{2}\cdot\sigma\Big(\tilde{f}_{\lceil\frac{m}{2}\rceil}(\tilde{g}_2(x))\Big) - \frac{1}{2}\cdot\sigma\Big(\tilde{f}_{\lfloor\frac{m}{2}\rfloor}(\tilde{g}_1(x))\Big)\right)\\ &\quad + \frac{1}{2}\cdot\sigma\Big(\tilde{f}_{\lfloor\frac{m}{2}\rfloor}(\tilde{g}_1(x))\Big) + \frac{1}{2}\cdot\sigma\Big(\tilde{f}_{\lceil\frac{m}{2}\rceil}(\tilde{g}_2(x))\Big).\end{aligned} \tag{C.59}$$

It follows from the induction hypothesis that

$$\begin{aligned}\tilde{f}_{\lceil\frac{m}{2}\rceil}\circ\tilde{g}_2 &\in \mathcal{F}_m^{\mathbf{FNN}}\left(1 + 2\left\lceil\frac{\log\lceil\frac{m}{2}\rceil}{\log 2}\right\rceil, 2\left\lceil\frac{m}{2}\right\rceil, 26\cdot 2^{\left\lceil\frac{\log\lceil\frac{m}{2}\rceil}{\log 2}\right\rceil} - 20 - 2\left\lceil\frac{\log\lceil\frac{m}{2}\rceil}{\log 2}\right\rceil, 1, 1\right)\\ &= \mathcal{F}_m^{\mathbf{FNN}}\left(-1 + 2\left\lceil\frac{\log m}{\log 2}\right\rceil, 2\left\lceil\frac{m}{2}\right\rceil, 13\cdot 2^{\left\lceil\frac{\log m}{\log 2}\right\rceil} - 18 - 2\left\lceil\frac{\log m}{\log 2}\right\rceil, 1, 1\right)\end{aligned}$$

and

$$\tilde{f}_{\lfloor\frac{m}{2}\rfloor}\circ\tilde{g}_1 \in \mathcal{F}_m^{\mathbf{FNN}}\left(1 + 2\left\lceil\frac{\log\lfloor\frac{m}{2}\rfloor}{\log 2}\right\rceil, 2\left\lfloor\frac{m}{2}\right\rfloor, 26\cdot 2^{\left\lceil\frac{\log\lfloor\frac{m}{2}\rfloor}{\log 2}\right\rceil} - 20 - 2\left\lceil\frac{\log\lfloor\frac{m}{2}\rfloor}{\log 2}\right\rceil, 1, 1\right)$$

$$\subset \mathcal{F}_m^{\mathbf{FNN}}\left(-1+2\left\lceil\frac{\log m}{\log 2}\right\rceil, 2\left\lfloor\frac{m}{2}\right\rfloor, 13\cdot 2^{\left\lceil\frac{\log m}{\log 2}\right\rceil}-18-2\left\lceil\frac{\log m}{\log 2}\right\rceil, 1, 1\right),$$

which, together with (C.59), yield

$$\begin{aligned}\tilde{f}_m \in \mathcal{F}_m^{\mathbf{FNN}}\Bigg(&2-1+2\left\lceil\frac{\log m}{\log 2}\right\rceil, 2\left\lceil\frac{m}{2}\right\rceil+2\left\lfloor\frac{m}{2}\right\rfloor,\\ &2\cdot\left|13\cdot 2^{\left\lceil\frac{\log m}{\log 2}\right\rceil}-18-2\left\lceil\frac{\log m}{\log 2}\right\rceil\right|+2\left\lceil\frac{\log m}{\log 2}\right\rceil+16, 1, \infty\Bigg) \\ = \mathcal{F}_m^{\mathbf{FNN}}&\left(1+\left\lceil\frac{\log m}{\log 2}\right\rceil, 2m, 26\cdot 2^{\left\lceil\frac{\log m}{\log 2}\right\rceil}-20-2\left\lceil\frac{\log m}{\log 2}\right\rceil, 1, \infty\right)\end{aligned} \tag{C.60}$$

(cf. Figure C.6). Besides, it is easy to verify that

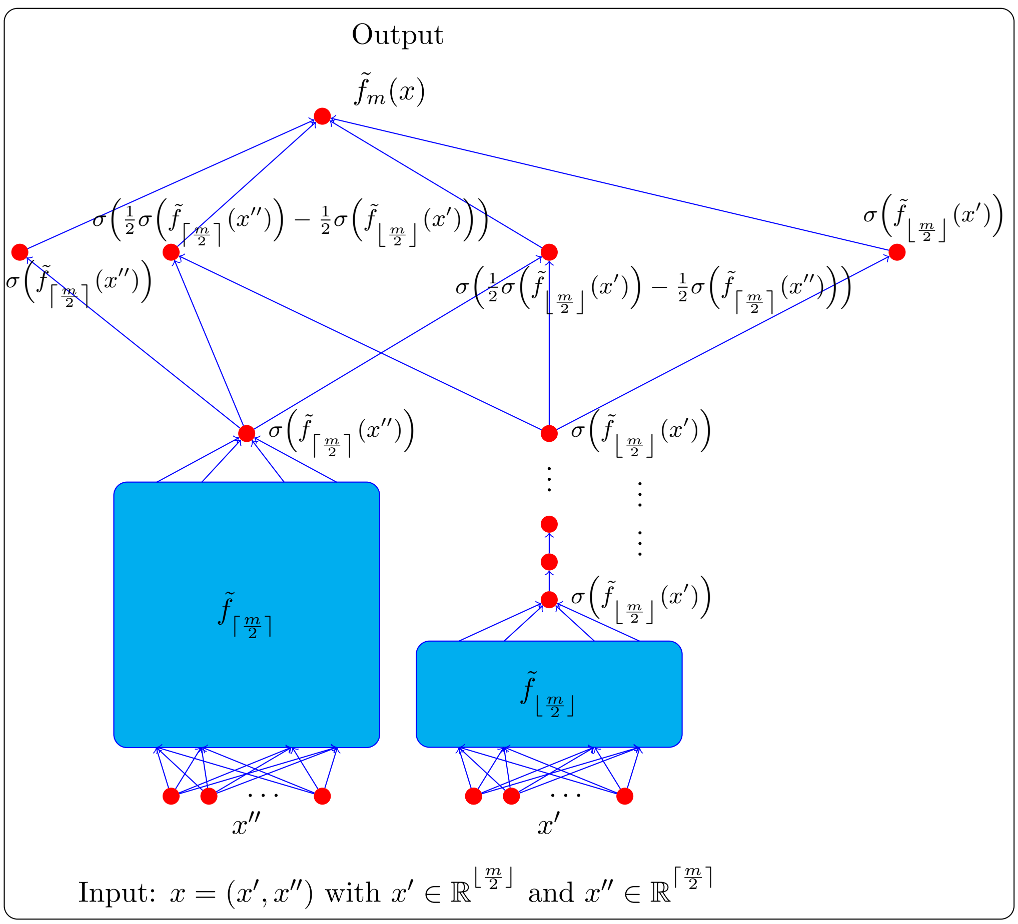

Figure C.6: The network $\tilde{f}_m$.

$$
\begin{aligned}
&\tilde{f}_m(x) = \max\left\{\sigma\Big(\tilde{f}_{\lfloor \frac{m}{2}\rfloor}(\tilde{g}_1(x))\Big), \sigma\Big(\tilde{f}_{\lceil \frac{m}{2}\rceil}(\tilde{g}_2(x))\Big)\right\}\\
&= \max\left\{\sigma\left(\left\|\Big((x)_1,\ldots,(x)_{\lfloor \frac{m}{2}\rfloor}\Big)\right\|_\infty\right), \sigma\left(\left\|\Big((x)_{\lfloor \frac{m}{2}\rfloor+1},\ldots,(x)_m\Big)\right\|_\infty\right)\right\}\\
&= \max\left\{\left\|\Big((x)_1,\ldots,(x)_{\lfloor \frac{m}{2}\rfloor}\Big)\right\|_\infty, \left\|\Big((x)_{\lfloor \frac{m}{2}\rfloor+1},\ldots,(x)_m\Big)\right\|_\infty\right\}\\
&= \max\left\{\max_{1\le i\le \lfloor \frac{m}{2}\rfloor}|(x)_i|, \max_{\lfloor \frac{m}{2}\rfloor+1\le i\le m}|(x)_i|\right\} = \max_{1\le i\le m}|(x)_i| = \|x\|_\infty, \ \forall\, x\in\mathbb{R}^m.
\end{aligned}
\tag{C.61}
$$

Combining (C.60) and (C.61), we deduce that the desired result holds for $k=m$. Therefore, according to mathematical induction, we have that the desired result hold for all positive integer k. This completes the proof. ■

Lemma C.13 *Let* $(\varepsilon, d, d_\star, d_*, \beta, r) \in (0,1/2]\times\mathbb{N}\times\mathbb{N}\times\mathbb{N}\times(0,\infty)\times(0,\infty)$ *and* f *be a function from* $[0,1]^d$ *to* $\mathbb{R}$. *Suppose* $f\in\mathcal{G}_\infty^{\mathbf{H}}(d_*,\beta,r\vee 1)\cup\mathcal{G}_\infty^{\mathbf{M}}(d_\star)$. *Then there exist constants* $E_1, E_2, E_3\in(0,\infty)$ *only depending on* (d_*,β,r) *and a neural network*

$$
\tilde{f}\in\mathcal{F}_d^{\mathbf{FNN}}\left(3\log d_\star + E_1\log\frac{1}{\varepsilon}, 2d_\star + E_2\varepsilon^{-\frac{d_*}{\beta}}, 52d_\star + E_3\varepsilon^{-\frac{d_*}{\beta}}\log\frac{1}{\varepsilon}, 1, \infty\right)
$$

such that

$$
\sup_{x\in[0,1]^d}\left|\tilde{f}(x)-f(x)\right| < 2\varepsilon.
$$

Proof According to Corollary B.1, there exist constants $E_1,E_2,E_3\in(6,\infty)$ only depending on (d_*,β,r), such that

$$
\begin{aligned}
&\inf\left\{\sup_{x\in[0,1]^{d_*}}|g(x)-\tilde{g}(x)|\,\middle|\,\tilde{g}\in\mathcal{F}_{d_*}^{\mathbf{FNN}}\left(E_1\log\frac{1}{t}, E_2 t^{-\frac{d_*}{\beta}}, E_3 t^{-\frac{d_*}{\beta}}\log\frac{1}{t}, 1, \infty\right)\right\}\\
&\le t, \ \forall\, g\in\mathcal{B}_{r\vee 1}^{\beta}\left([0,1]^{d_*}\right), \ \forall\, t\in(0,1/2].
\end{aligned}
\tag{C.62}
$$

We next consider two cases.

Case I: $f\in\mathcal{G}_\infty^{\mathbf{M}}(d_\star)$. In this case, we must have $f\in\mathcal{G}_d^{\mathbf{M}}(d_\star)$, since $\mathbf{dom}(f)=[0,1]^d$. Therefore, there exists $I\subset\{1,2,\ldots,d\}$, such that $1\le\#(I)\le d_\star$ and

$$
f(x)=\max\left\{(x)_i\middle|i\in I\right\}, \ \forall\, x\in[0,1]^d.
$$

According to Lemma C.12, there exists

$$
\begin{aligned}
&\tilde{g}\in\mathcal{F}_{\#(I)}^{\mathbf{FNN}}\left(1+2\cdot\left\lceil\frac{\log\#(I)}{\log 2}\right\rceil, 2\cdot\#(I), 26\cdot 2^{\left\lceil\frac{\log\#(I)}{\log 2}\right\rceil}-20-2\cdot\left\lceil\frac{\log\#(I)}{\log 2}\right\rceil, 1, 1\right)\\
&\subset\mathcal{F}_{\#(I)}^{\mathbf{FNN}}\left(1+2\cdot\left\lceil\frac{\log d_\star}{\log 2}\right\rceil, 2d_\star, 26\cdot 2^{\left\lceil\frac{\log d_\star}{\log 2}\right\rceil}, 1, 1\right)\\
&\subset\mathcal{F}_{\#(I)}^{\mathbf{FNN}}\left(3+3\log d_\star, 2d_\star, 52d_\star, 1, 1\right)
\end{aligned}
\tag{C.63}
$$

such that

$$\tilde{g}(x) = \|x\|_\infty, \ \forall\, x \in \mathbb{R}^{\#(I)}.$$

Define $\tilde{f} : \mathbb{R}^d \to \mathbb{R},\ x \mapsto \tilde{g}((x)_I)$. Then it follows from (C.63) that

$$\begin{aligned}\tilde{f} &\in \mathcal{F}_d^{\mathbf{FNN}}(3 + 3\log d_\star, 2d_\star, 52d_\star, 1, 1)\\ &\subset \mathcal{F}_d^{\mathbf{FNN}}\left(3\log d_\star + E_1 \log\frac{1}{\varepsilon}, 2d_\star + E_2\varepsilon^{-\frac{d_*}{\beta}}, 52d_\star + E_3\varepsilon^{-\frac{d_*}{\beta}}\log\frac{1}{\varepsilon}, 1, \infty\right)\end{aligned}$$

and

$$\begin{aligned}\sup_{x\in[0,1]^d}\left|f(x) - \tilde{f}(x)\right| &= \sup_{x\in[0,1]^d}\left|\max\left\{(x)_i \middle| i \in I\right\} - \tilde{g}((x)_I)\right|\\ &= \sup_{x\in[0,1]^d}\left|\max\left\{|(x)_i| \middle| i \in I\right\} - \|(x)_I\|_\infty\right| = 0 < 2\varepsilon,\end{aligned}$$

which yield the desired result.

Case II: $f \in \mathcal{G}_\infty^{\mathbf{H}}(d_*, \beta, r \vee 1)$. In this case, we must have $f \in \mathcal{G}_d^{\mathbf{H}}(d_*, \beta, r \vee 1)$, since $\mathbf{dom}(f) = [0,1]^d$. By definition, there exist $I \subset \{1, 2, \ldots, d\}$ and $g \in \mathcal{B}_{r\vee 1}^{\beta}\left([0,1]^{d_*}\right)$ such that $\#(I) = d_*$ and $f(x) = g\left((x)_I\right)$ for all $x \in [0,1]^d$. Then it follows from (C.62) that there exists $\tilde{g} \in \mathcal{F}_{d_*}^{\mathbf{FNN}}\left(E_1 \log\frac{1}{\varepsilon}, E_2\varepsilon^{-\frac{d_*}{\beta}}, E_3\varepsilon^{-\frac{d_*}{\beta}}\log\frac{1}{\varepsilon}, 1, \infty\right)$ such that

$$\sup_{x\in[0,1]^{d_*}} |g(x) - \tilde{g}(x)| < 2\varepsilon.$$

Define $\tilde{f} : \mathbb{R}^d \to \mathbb{R},\ x \mapsto \tilde{g}((x)_I)$. Then we have that

$$\begin{aligned}\tilde{f} &\in \mathcal{F}_d^{\mathbf{FNN}}\left(E_1 \log\frac{1}{\varepsilon}, E_2\varepsilon^{-\frac{d_*}{\beta}}, E_3\varepsilon^{-\frac{d_*}{\beta}}\log\frac{1}{\varepsilon}, 1, \infty\right)\\ &\subset \mathcal{F}_d^{\mathbf{FNN}}\left(3\log d_\star + E_1 \log\frac{1}{\varepsilon}, 2d_\star + E_2\varepsilon^{-\frac{d_*}{\beta}}, 52d_\star + E_3\varepsilon^{-\frac{d_*}{\beta}}\log\frac{1}{\varepsilon}, 1, \infty\right)\end{aligned}$$

and

$$\sup_{x\in[0,1]^d}\left|f(x) - \tilde{f}(x)\right| = \sup_{x\in[0,1]^d}|g((x)_I) - \tilde{g}((x)_I)| = \sup_{x\in[0,1]^{d_*}}|g(x) - \tilde{g}(x)| < 2\varepsilon.$$

These yield the desired result again.

In conclusion, the desired result always holds. Thus we completes the proof of this lemma. ■

Lemma C.14 *Let $\beta \in (0,\infty)$, $r \in (0,\infty)$, $q \in \mathbb{N} \cup \{0\}$, and $(d, d_\star, d_*, K) \in \mathbb{N}^4$ with $d_* \le \min\left\{d, K + \mathbb{1}_{\{0\}}(q)\cdot(d-K)\right\}$. Suppose $f \in \mathcal{G}_d^{\mathbf{CHOM}}(q, K, d_\star, d_*, \beta, r)$ and $\varepsilon \in (0, 1/2]$. Then there exist $E_7 \in (0,\infty)$ only depending on (d_*, β, r, q) and*

$$\begin{aligned}\tilde{f} \in \mathcal{F}_d^{\mathbf{FNN}}\Bigg((q+1)\cdot\left|3\log d_\star + E_7\log\frac{1}{\varepsilon}\right|, 2Kd_\star + KE_7\varepsilon^{-\frac{d_*}{\beta\cdot(1\wedge\beta)^q}},\\ (Kq+1)\cdot\left|63d_\star + E_7\varepsilon^{-\frac{d_*}{\beta\cdot(1\wedge\beta)^q}}\log\frac{1}{\varepsilon}\right|, 1, \infty\Bigg)\end{aligned} \tag{C.64}$$

such that

$$\sup_{x\in[0,1]^d}\left|f(x)-\tilde{f}(x)\right|\leq\frac{\varepsilon}{8}. \tag{C.65}$$

Proof By the definition of $\mathcal{G}_d^{\mathbf{CHOM}}(q,K,d_\star,d_*,\beta,r)$, there exist functions $h_0,h_1,\ldots,h_q$ such that

(i) $\mathbf{dom}(h_i)=[0,1]^K$ for $0<i\leq q$ and $\mathbf{dom}(h_0)=[0,1]^d$;

(ii) $\mathbf{ran}(h_i)\subset[0,1]^K$ for $0\leq i<q$ and $\mathbf{ran}(h_q)\subset\mathbb{R}$;

(iii) $h_q\in\mathcal{G}_\infty^{\mathbf{H}}(d_*,\beta,r\vee 1)\cup\mathcal{G}_\infty^{\mathbf{M}}(d_\star)$;

(iv) For $0\leq i<q$ and $1\leq j\leq K$, the j-th coordinate function of h_i given by $\mathbf{dom}(h_i)\ni x\mapsto(h_i(x))_j\in\mathbb{R}$ belongs to $\mathcal{G}_\infty^{\mathbf{H}}(d_*,\beta,r\vee 1)\cup\mathcal{G}_\infty^{\mathbf{M}}(d_\star)$;

(v) $f=h_q\circ h_{q-1}\circ\cdots\circ h_2\circ h_1\circ h_0$.

Define $\Omega:=\left\{(i,j)\in\mathbb{Z}^2\,\middle|\,0\leq i\leq q,1\leq j\leq K,\mathbb{1}_{\{q\}}(i)\leq\mathbb{1}_{\{1\}}(j)\right\}$. For each $(i,j)\in\Omega$, denote $d_{i,j}:=K+\mathbb{1}_{\{0\}}(i)\cdot(d-K)$ and

$$h_{i,j}:\mathbf{dom}(h_i)\to\mathbb{R},\ \ x\mapsto\big(h_i(x)\big)_j.$$

Then it is easy to verify that,

$$\mathbf{dom}(h_{i,j})=[0,1]^{d_{i,j}}\text{ and }h_{i,j}\in\mathcal{G}_\infty^{\mathbf{H}}(d_*,\beta,r\vee 1)\cup\mathcal{G}_\infty^{\mathbf{M}}(d_\star),\ \forall\,(i,j)\in\Omega, \tag{C.66}$$

and

$$\mathbf{ran}\left(h_{i,j}\right)\subset[0,1],\ \forall\,(i,j)\in\Omega\setminus\{(q,1)\}. \tag{C.67}$$

Fix $\varepsilon\in(0,1/2]$. Take

$$\delta:=\frac{1}{2}\cdot\left|\frac{\varepsilon}{8\cdot|(1\vee r)\cdot d_*|^q\cdot(q+1)}\right|^{\frac{1}{(1\wedge\beta)^q}}\leq\frac{\varepsilon/2}{8\cdot|(1\vee r)\cdot d_*|^q\cdot(q+1)}\leq\frac{\varepsilon}{8}\leq\frac{1}{16}.$$

According to (C.66) and Lemma C.13, there exists a constant $E_1\in(6,\infty)$ only depending on (d_*,β,r) and a set of functions $\left\{\tilde{g}_{i,j}:\mathbb{R}^{d_{i,j}}\to\mathbb{R}\right\}_{(i,j)\in\Omega}$, such that

$$\begin{aligned}\tilde{g}_{i,j}\in\mathcal{F}_{d_{i,j}}^{\mathbf{FNN}}\Bigg(&3\log d_\star+E_1\log\frac{1}{\delta},2d_\star+E_1\delta^{-\frac{d_*}{\beta}},\\&52d_\star+E_1\delta^{-\frac{d_*}{\beta}}\log\frac{1}{\delta},1,\infty\Bigg),\forall(i,j)\in\Omega\end{aligned} \tag{C.68}$$

and

$$\sup\left\{|\tilde{g}_{i,j}(x)-h_{i,j}(x)|\,\middle|\,x\in[0,1]^{d_{i,j}}\right\}\leq 2\delta,\ \forall\,(i,j)\in\Omega. \tag{C.69}$$

Define

$$E_4:=8\cdot|(1\vee r)\cdot d_*|^q\cdot(q+1),$$

$$
\begin{aligned}
E_5 &:= 2^{\frac{d_*}{\beta}} \cdot E_4^{\frac{d_*}{\beta \cdot (1 \wedge \beta)^q}}, \\
E_6 &:= \frac{1}{(1 \wedge \beta)^q} + \frac{2 \log E_4}{(1 \wedge \beta)^q} + 2 \log 2, \\
E_7 &:= E_1 E_6 + E_1 E_5 + 2 E_1 E_5 E_6 + 6,
\end{aligned}
$$

Obviously, E_4, E_5, E_6, E_7 are constants only depending on (d_*, β, r, q). Next, define

$$
\tilde{h}_{i,j} : \mathbb{R}^{d_{i,j}} \to \mathbb{R},\ x \mapsto \sigma\big(\sigma\left(\tilde{g}_{i,j}(x)\right)\big) - \sigma\big(\sigma\left(\tilde{g}_{i,j}(x)\right) - 1\big)
$$

for each $(i,j) \in \Omega \setminus \{(q,1)\}$, and define $\tilde{h}_{q,1} := \tilde{g}_{q,1}$. It follows from the fact

$$
\sigma\big(\sigma\left(z\right)\big) - \sigma\big(\sigma\left(z\right) - 1\big) \in [0,1],\ \forall\, z \in \mathbb{R}
$$

and (C.68) that

$$
\mathbf{ran}(\tilde{h}_{i,j}) \subset [0,1],\ \forall\, (i,j) \in \Omega \setminus (q,1) \tag{C.70}
$$

and

$$
\begin{aligned}
\tilde{h}_{i,j} \in \mathcal{F}^{\mathbf{FNN}}_{d_{i,j}} \Bigg(& 2 + 3 \log d_\star + E_1 \log \frac{1}{\delta}, 2 d_\star + E_1 \delta^{-\frac{d_*}{\beta}}, \\
& 58 d_\star + E_1 \delta^{-\frac{d_*}{\beta}} \log \frac{1}{\delta}, 1, \infty \Bigg), \forall (i,j) \in \Omega.
\end{aligned} \tag{C.71}
$$

Besides, it follows from the fact that

$$
\left|\sigma\big(\sigma\left(z\right)\big) - \sigma\big(\sigma\left(z\right) - 1\big) - w\right| \le \left|w - z\right|,\ \forall\, z \in \mathbb{R},\ \forall\, w \in [0,1]
$$

and (C.69) that

$$
\begin{aligned}
& \sup \left\{ \left|\tilde{h}_{i,j}(x) - h_{i,j}(x)\right| \middle| x \in [0,1]^{d_{i,j}} \right\} \\
& \le \sup \left\{ \left|\tilde{g}_{i,j}(x) - h_{i,j}(x)\right| \middle| x \in [0,1]^{d_{i,j}} \right\} \le 2\delta.
\end{aligned} \tag{C.72}
$$

We then define

$$
\tilde{h}_i : \mathbb{R}^{d_{i,1}} \to \mathbb{R}^K, x \mapsto \Big(\tilde{h}_{i,1}(x), \tilde{h}_{i,2}(x), \ldots, \tilde{h}_{i,K}(x)\Big)^\top
$$

for each $i \in \{0,1,\ldots,q-1\}$, and $\tilde{h}_q := \tilde{h}_{q,1}$. From (C.70) we obtain

$$
\mathbf{ran}(\tilde{h}_i) \subset [0,1]^K \subset \mathbf{dom}(\tilde{h}_{i+1}),\ \forall\, i \in \{0,1,\ldots,q-1\}. \tag{C.73}
$$

Thus we can well define the function $\tilde{f} := \tilde{h}_q \circ \tilde{h}_{q-1} \circ \cdots \circ \tilde{h}_1 \circ \tilde{h}_0$, which is from $\mathbb{R}^d$ to $\mathbb{R}$. Since all the functions $\tilde{h}_{i,j}$ ($(i,j) \in \Omega$) are neural networks satisfying (C.71), we deduce that $\tilde{f}$ is also a neural network, which is comprised of all those networks $\tilde{h}_{i,j}$ through series and parallel connection. Obviously, the depth of $\tilde{f}$ is less than or equal to

$$
\sum_i \left(1 + \max_j \Big(\text{the depth of } \tilde{h}_{i,j}\Big)\right),
$$

the width of $\tilde{f}$ is less than or equal to

$$\max_i \sum_j \Big(\text{the width of } \tilde{h}_{i,j}\Big),$$

the number of nonzero parameters of $\tilde{f}$ is less than or equal to

$$\sum_{i,j} \bigg(\Big(\text{the number of nonzero parameters } \tilde{h}_{i,j}\Big) + \max_k \Big(\text{the depth of } \tilde{h}_{i,k}\Big)\bigg),$$

and the parameters of $\tilde{f}$ is bounded by 1 in absolute value. Thus we have that

$$\begin{aligned}
\tilde{f} \in \mathcal{F}_d^{\mathbf{FNN}} \Bigg(&(q+1)\cdot\left|3 + 3\log d_\star + E_1 \log\frac{1}{\delta}\right|, 2Kd_\star + KE_1\delta^{-\frac{d_*}{\beta}},\\
&(Kq+1)\cdot\left|63d_\star + 2E_1\delta^{-\frac{d_*}{\beta}}\log\frac{1}{\delta}\right|, 1, \infty\Bigg)\\
= \mathcal{F}_d^{\mathbf{FNN}} \Bigg(&(q+1)\cdot\left|3 + 3\log d_\star + E_1\cdot\left(\log 2 + \frac{\log\frac{E_4}{\varepsilon}}{(1\wedge\beta)^q}\right)\right|, 2Kd_\star + KE_1E_5\varepsilon^{-\frac{d_*}{\beta\cdot(1\wedge\beta)^q}},\\
&(Kq+1)\cdot\left|63d_\star + 2E_1E_5\varepsilon^{-\frac{d_*}{\beta\cdot(1\wedge\beta)^q}}\cdot\left(\log 2 + \frac{\log\frac{E_4}{\varepsilon}}{(1\wedge\beta)^q}\right)\right|, 1, \infty\Bigg)\\
\subset \mathcal{F}_d^{\mathbf{FNN}} \Bigg(&(q+1)\cdot\left|3 + 3\log d_\star + E_1E_6 \log\frac{1}{\varepsilon}\right|, 2Kd_\star + KE_1E_5\varepsilon^{-\frac{d_*}{\beta\cdot(1\wedge\beta)^q}},\\
&(Kq+1)\cdot\left|63d_\star + 2E_1E_5\varepsilon^{-\frac{d_*}{\beta\cdot(1\wedge\beta)^q}}E_6\log\frac{1}{\varepsilon}\right|, 1, \infty\Bigg)\\
\subset \mathcal{F}_d^{\mathbf{FNN}} \Bigg(&(q+1)\cdot\left|3\log d_\star + E_7 \log\frac{1}{\varepsilon}\right|, 2Kd_\star + KE_7\varepsilon^{-\frac{d_*}{\beta\cdot(1\wedge\beta)^q}},\\
&(Kq+1)\cdot\left|63d_\star + E_7\varepsilon^{-\frac{d_*}{\beta\cdot(1\wedge\beta)^q}}\log\frac{1}{\varepsilon}\right|, 1, \infty\Bigg),
\end{aligned}$$

leading to (C.64). Moreover, it follows from (C.72) and Lemma C.11 that

$$\begin{aligned}
&\sup_{x\in[0,1]^d}\left|\tilde{f}(x) - f(x)\right| = \sup_{x\in[0,1]^d}\left|\tilde{h}_q\circ\cdots\circ\tilde{h}_0(x) - h_q\circ\cdots\circ h_0(x)\right|\\
&\leq \left|(1\vee r)\cdot d_*^{1\wedge\beta}\right|^{\sum_{i=0}^{q-1}(1\wedge\beta)^i}\cdot\sum_{i=0}^{q}\left|\sup_{x\in[0,1]^{d_{i,1}}}\left\|\tilde{h}_i(x) - h_i(x)\right\|_\infty\right|^{(1\wedge\beta)^{q-i}}\\
&\leq |(1\vee r)\cdot d_*|^q\cdot\sum_{i=0}^{q}|2\delta|^{(1\wedge\beta)^{q-i}} \leq |(1\vee r)\cdot d_*|^q\cdot\sum_{i=0}^{q}|2\delta|^{(1\wedge\beta)^q} = \frac{\varepsilon}{8},
\end{aligned}$$

which yields (C.65).

In conclusion, the constant E_7 and the neural network $\tilde{f}$ have all the desired properties. The proof of this lemma is then completed.

■

The next lemma aims to estimate the approximation error.

Lemma C.15 *Let $\phi(t)=\log\left(1+\mathrm{e}^{-t}\right)$ be the logistic loss, $q\in\mathbb{N}\cup\{0\}$, $(\beta,r)\in(0,\infty)^2$, $(d,d_\star,d_*,K)\in\mathbb{N}^4$ with $d_*\leq\min\left\{d,K+\mathbb{1}_{\{0\}}(q)\cdot(d-K)\right\}$, and P be a Borel probability measure on $[0,1]^d\times\{-1,1\}$. Suppose that there exists an $\hat{\eta}\in\mathcal{G}_d^{\mathbf{CHOM}}(q,K,d_\star,d_*,\beta,r)$ such that $P_X(\left\{x\in[0,1]^d\middle|\,\hat{\eta}(x)=P(\{1\}\,|x)\right\})=1$. Then there exist constants D_1,D_2,D_3 only depending on $(d_\star,d_*,\beta,r,q)$ such that for any $\delta\in(0,1/3)$,*

$$\begin{aligned}&\inf\left\{\mathcal{E}_P^{\phi}\left(f\right)\middle|f\in\mathcal{F}_d^{\mathbf{FNN}}\left(D_1\log\frac{1}{\delta},KD_2\delta^{\frac{-d_*/\beta}{(1\wedge\beta)^q}},KD_3\delta^{\frac{-d_*/\beta}{(1\wedge\beta)^q}}\cdot\log\frac{1}{\delta},1,\log\frac{1-\delta}{\delta}\right)\right\}\\&\leq 8\delta.\end{aligned}\tag{C.74}$$

Proof Denote by η the conditional probability function $[0,1]^d\ni x\mapsto P(\{1\}\,|x)\in[0,1]$. Fix $\delta\in(0,1/3)$. Then it follows from Lemma C.14 that there exists

$$\begin{aligned}\tilde{\eta}\in\mathcal{F}_d^{\mathbf{FNN}}\Bigg(&C_{d_\star,d_*,\beta,r,q}\log\frac{1}{\delta},KC_{d_\star,d_*,\beta,r,q}\delta^{-\frac{d_*}{\beta\cdot(1\wedge\beta)^q}},\\&KC_{d_\star,d_*,\beta,r,q}\delta^{-\frac{d_*}{\beta\cdot(1\wedge\beta)^q}}\log\frac{1}{\delta},1,\infty\Bigg)\end{aligned}\tag{C.75}$$

such that

$$\sup_{x\in[0,1]^d}|\tilde{\eta}(x)-\hat{\eta}(x)|\leq\delta/8.\tag{C.76}$$

Also, by Theorem 2.4 with $a=\varepsilon=\delta$, $b=1-\delta$ and $\alpha=\frac{2\beta}{d_*}$, there exists

$$\begin{aligned}\tilde{l}\in\mathcal{F}_1^{\mathbf{FNN}}\Bigg(&C_{d_*,\beta}\log\frac{1}{\delta}+139\log\frac{1}{\delta},C_{d_*,\beta}\cdot\left(\frac{1}{\delta}\right)^{\frac{1}{2\beta/d_*}}\log\frac{1}{\delta},\\&C_{d_*,\beta}\cdot\left(\frac{1}{\delta}\right)^{\frac{1}{2\beta/d_*}}\cdot\left(\log\frac{1}{\delta}\right)\cdot\left(\log\frac{1}{\delta}\right)+65440\left(\log\delta\right)^2,1,\infty\Bigg)\\\subset\mathcal{F}_1^{\mathbf{FNN}}&\left(C_{d_*,\beta}\log\frac{1}{\delta},C_{d_*,\beta}\delta^{-\frac{d_*}{\beta}},C_{d_*,\beta}\delta^{-\frac{d_*}{\beta}}\log\frac{1}{\delta},1,\infty\right)\\\subset\mathcal{F}_1^{\mathbf{FNN}}&\left(C_{d_*,\beta}\log\frac{1}{\delta},C_{d_*,\beta}\delta^{-\frac{d_*}{\beta\cdot(1\wedge\beta)^q}},C_{d_*,\beta}\delta^{-\frac{d_*}{\beta\cdot(1\wedge\beta)^q}}\log\frac{1}{\delta},1,\infty\right)\end{aligned}\tag{C.77}$$

such that

$$\sup_{t\in[\delta,1-\delta]}\left|\tilde{l}(t)-\log t\right|\leq\delta\tag{C.78}$$

and

$$\log\delta\leq\tilde{l}(t)\leq\log\left(1-\delta\right)<0,\ \forall\,t\in\mathbb{R}.\tag{C.79}$$

Recall that the clipping function Π_δ is given by

$$\Pi_\delta:\mathbb{R}\to[\delta,1-\delta],\quad t\mapsto\begin{cases}1-\delta, & \text{if } t>1-\delta,\\ \delta, & \text{if } t<\delta,\\ t, & \text{otherwise.}\end{cases}$$

Define $\tilde{f} : \mathbb{R} \to \mathbb{R}, x \mapsto \tilde{l}\left(\Pi_\delta\left(\tilde{\eta}(x)\right)\right) - \tilde{l}\left(1 - \Pi_\delta\left(\tilde{\eta}(x)\right)\right)$. Consequently, we know from (C.75), (C.77) and (C.79) that (cf. Figure C.7)

$$\tilde{f} \in \mathcal{F}_d^{\mathbf{FNN}}\Bigg(C_{d_\star,d_*,\beta,r,q}\log\frac{1}{\delta}, KC_{d_\star,d_*,\beta,r,q}\delta^{-\frac{d_*}{\beta\cdot(1\wedge\beta)^q}}, KC_{d_\star,d_*,\beta,r,q}\delta^{-\frac{d_*}{\beta\cdot(1\wedge\beta)^q}}\log\frac{1}{\delta}, 1, \log\frac{1-\delta}{\delta}\Bigg).$$

Let $\Omega_1, \Omega_2, \Omega_3$ be defined in (C.35). Then it follows from (C.76) that

$$\begin{aligned}&|\Pi_\delta(\tilde{\eta}(x)) - \eta(x)| = |\Pi_\delta(\tilde{\eta}(x)) - \Pi_\delta\left(\hat{\eta}(x)\right)| \le |\tilde{\eta}(x) - \hat{\eta}(x)| \\ &\le \frac{\delta}{8} \le \frac{\min\left\{\eta(x), 1-\eta(x)\right\}}{8}, \ \forall\, x \in \Omega_1,\end{aligned}$$

which means that

$$\min\left\{\frac{\Pi_\delta\left(\tilde{\eta}(x)\right)}{\eta(x)}, \frac{1-\Pi_\delta\left(\tilde{\eta}(x)\right)}{1-\eta(x)}\right\} \ge 7/8, \ \forall\, x \in \Omega_1. \tag{C.80}$$

Combining (C.78) and (C.80), we obtain that

$$\begin{aligned}&\left|\tilde{f}(x) - \log\frac{\eta(x)}{1-\eta(x)}\right| \\ &\le \left|\tilde{l}\left(\Pi_\delta\left(\tilde{\eta}(x)\right)\right) - \log\left(\eta(x)\right)\right| + \left|\tilde{l}\left(1-\Pi_\delta\left(\tilde{\eta}(x)\right)\right) - \log\left(1-\eta(x)\right)\right| \\ &\le \left|\tilde{l}\left(\Pi_\delta\left(\tilde{\eta}(x)\right)\right) - \log\left(\Pi_\delta\left(\tilde{\eta}(x)\right)\right)\right| + \left|\log\left(\Pi_\delta\left(\tilde{\eta}(x)\right)\right) - \log\left(\eta(x)\right)\right| \\ &\quad + \left|\tilde{l}\left(1-\Pi_\delta\left(\tilde{\eta}(x)\right)\right) - \log\left(1-\Pi_\delta\left(\tilde{\eta}(x)\right)\right)\right| + \left|\log\left(1-\Pi_\delta\left(\tilde{\eta}(x)\right)\right) - \log\left(1-\eta(x)\right)\right| \\ &\le \delta + \sup_{t\in[\Pi_\delta(\tilde{\eta}(x))\wedge\eta(x),\infty)}\left|\log'(t)\right| \cdot \left|\Pi_\delta\left(\tilde{\eta}(x)\right) - \eta(x)\right| \\ &\quad + \delta + \sup_{t\in[\min\{1-\Pi_\delta(\tilde{\eta}(x)),1-\eta(x)\},\infty)}\left|\log'(t)\right| \cdot \left|\Pi_\delta\left(\tilde{\eta}(x)\right) - \eta(x)\right| \\ &\le \delta + \sup_{t\in[7\eta(x)/8,\infty)}\left|\log'(t)\right| \cdot \left|\Pi_\delta\left(\tilde{\eta}(x)\right) - \eta(x)\right| \\ &\quad + \delta + \sup_{t\in\left[\frac{7-7\eta(x)}{8},\infty\right)}\left|\log'(t)\right| \cdot \left|\Pi_\delta\left(\tilde{\eta}(x)\right) - \eta(x)\right| \\ &\le 2\delta + \frac{8}{7\eta(x)} \cdot \frac{\delta}{8} + \frac{8}{7-7\eta(x)} \cdot \frac{\delta}{8}, \ \forall\, x \in \Omega_1,\end{aligned}$$

meaning that

$$\begin{aligned}&\left|\tilde{f}(x) - \log\frac{\eta(x)}{1-\eta(x)}\right| \le 2\delta + \frac{8}{7\eta(x)} \cdot \frac{\delta}{8} + \frac{8}{7-7\eta(x)} \cdot \frac{\delta}{8} \\ &= 2\delta + \frac{\delta}{7\eta(x)(1-\eta(x))} \le \frac{2}{3} + \frac{2}{7} < 1, \ \forall\, x \in \Omega_1.\end{aligned} \tag{C.81}$$

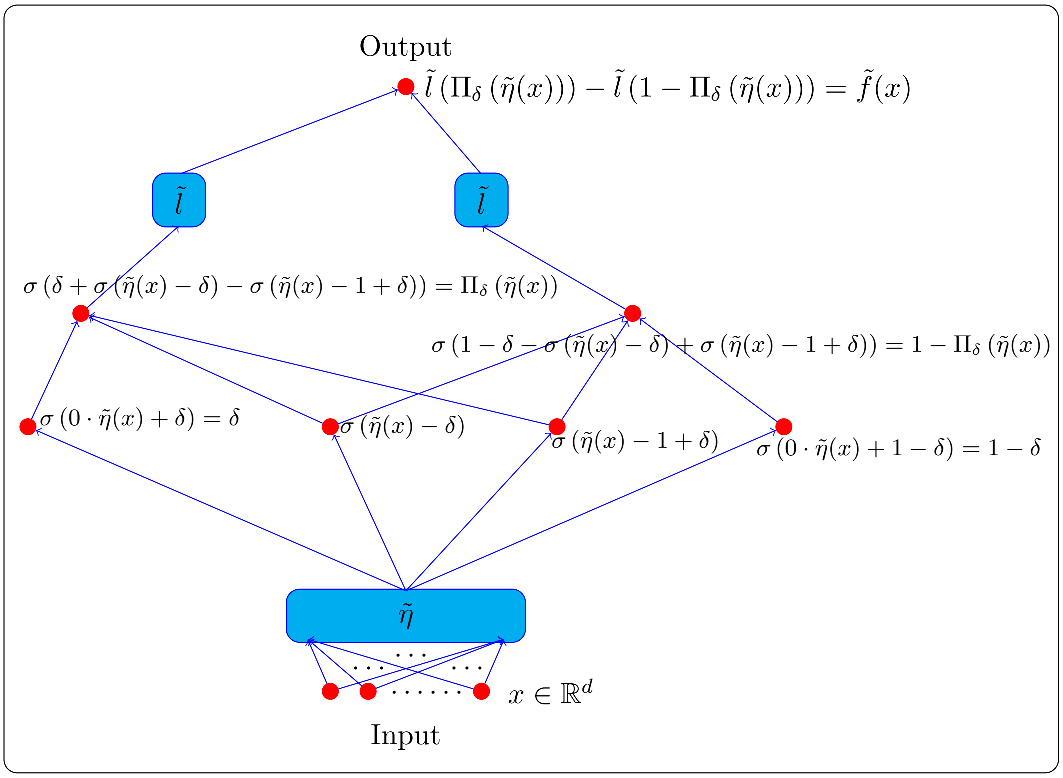

Figure C.7: The network representing the function $\tilde{f}$.

Besides, note that

$$\begin{aligned}
&x\in\Omega_2\Rightarrow\tilde{\eta}(x)\in[-\xi_1,\delta+\xi_1]\Rightarrow\Pi_\delta\left(\tilde{\eta}(x)\right)\in[\delta,\delta+\xi_1]\\
&\Rightarrow\tilde{l}\left(\Pi_\delta\left(\tilde{\eta}(x)\right)\right)\in\left[\log\delta,\delta+\log\left(\delta+\xi_1\right)\right]\\
&\quad\text{as well as }\tilde{l}\left(1-\Pi_\delta\left(\tilde{\eta}(x)\right)\right)\in\left[-\delta+\log(1-\delta-\xi_1),\log(1-\delta)\right]\\
&\Rightarrow\tilde{f}(x)\le2\delta+\log\frac{\xi_1+\delta}{1-\xi_1-\delta}\le\log2+\log\frac{2\delta}{1-2\delta}=\log\frac{4\delta}{1-2\delta}.
\end{aligned}$$

Therefore, by (C.79) and the definition of $\tilde{f}$, we have

$$\log\frac{\delta}{1-\delta}\le\tilde{f}(x)\le\log\frac{4\delta}{1-2\delta}=-\log\frac{1-2\delta}{4\delta},\ \forall\,x\in\Omega_2. \tag{C.82}$$

Similarly, we can show that

$$\log\frac{1-2\delta}{4\delta}\le\tilde{f}(x)\le\log\frac{1-\delta}{\delta},\ \forall\,x\in\Omega_3. \tag{C.83}$$

Then it follows from (C.81), (C.82), (C.83) and Lemma C.8 that

$$\inf\left\{\mathcal{E}_P^\phi\left(f\right)\,\middle|\,\begin{aligned}&f\in\mathcal{F}_d^{\mathbf{FNN}}\left(C_{d_\star,d_*,\beta,r,q}\log\frac{1}{\delta},KC_{d_\star,d_*,\beta,r,q}\delta^{-\frac{d_*}{\beta\cdot(1\wedge\beta)^q}},\right.\\&\left.KC_{d_\star,d_*,\beta,r,q}\delta^{-\frac{d_*}{\beta\cdot(1\wedge\beta)^q}}\log\frac{1}{\delta},1,\log\frac{1-\delta}{\delta}\right)\end{aligned}\right\}$$

$$
\begin{aligned}
&\le \mathcal{E}_P^{\phi}\left(\tilde{f}\right) \le \phi\left(\log\frac{1-2\delta}{4\delta}\right)P_X(\Omega_2)+\phi\left(\log\frac{1-2\delta}{4\delta}\right)P_X(\Omega_3)\\
&\quad+\int_{\Omega_1}\sup\left\{\frac{\left|\tilde{f}(x)-\log\frac{\eta(x)}{1-\eta(x)}\right|^2}{2(2+\mathrm{e}^t+\mathrm{e}^{-t})}\middle| t\in\left[\tilde{f}(x)\wedge\log\frac{\eta(x)}{1-\eta(x)},\tilde{f}(x)\vee\log\frac{\eta(x)}{1-\eta(x)}\right]\right\}\mathrm{d}P_X(x)\\
&\le \int_{\Omega_1}\sup\left\{\frac{\left|\tilde{f}(x)-\log\frac{\eta(x)}{1-\eta(x)}\right|^2}{2(2+\mathrm{e}^t+\mathrm{e}^{-t})}\middle| t\in\left[-1+\log\frac{\eta(x)}{1-\eta(x)},1+\log\frac{\eta(x)}{1-\eta(x)}\right]\right\}\mathrm{d}P_X(x)\\
&\quad+P_X(\Omega_2\cup\Omega_3)\cdot\log\frac{1+2\delta}{1-2\delta}\\
&\le \int_{\Omega_1}\left|\tilde{f}\,(x)-\log\frac{\eta(x)}{1-\eta(x)}\right|^2\cdot 2\cdot(1-\eta(x))\eta(x)\mathrm{d}P_X(x)+6\delta\\
&\le \int_{\Omega_1}\left|2\delta+\frac{\delta}{7\eta(x)(1-\eta(x))}\right|^2\cdot 2\cdot(1-\eta(x))\eta(x)\mathrm{d}P_X(x)+6\delta\\
&\le \int_{\Omega_1}\frac{\delta^2}{(1-\eta(x))\eta(x)}\mathrm{d}P_X(x)+6\delta\le\frac{\delta^2}{\delta(1-\delta)}+6\delta<8\delta,
\end{aligned}
$$

which proves this lemma. ■

Now we are in the position to prove Theorem 2.2 and Theorem 2.3.

Proof [Proof of Theorem 2.2 and Theorem 2.3] We first prove Theorem 2.3. According to Lemma C.15, there exist $(D_1,D_2,D_3)\in(0,\infty)^3$ only depending on $(d_\star,d_*,\beta,r,q)$ such that (C.74) holds for any $\delta\in(0,1/3)$ and any $P\in\mathcal{H}^{d,\beta,r}_{4,q,K,d_\star,d_*}$. Take $E_1=1+D_1$, then $E_1>0$ only depends on $(d_\star,d_*,\beta,r,q)$. We next show that for any constants $\boldsymbol{a}:=(a_2,a_3)\in(0,\infty)^2$ and $\boldsymbol{b}:=(b_1,b_2,b_3,b_4,b_5)\in(0,\infty)^5$, there exist constants $E_2\in(3,\infty)$ only depends on $(\boldsymbol{a},d_\star,d_*,\beta,r,q,K)$ and $E_3\in(0,\infty)$ only depending on $(\boldsymbol{a},\boldsymbol{b},\nu,d,d_\star,d_*,\beta,r,q,K)$ such that when $n\ge E_2$, the ϕ-ERM $\hat{f}^{\mathbf{FNN}}_n$ defined by (2.14) with

$$
\begin{aligned}
&E_1\cdot\log n\le G\le b_1\cdot\log n,\\
&a_2\cdot\left(\frac{(\log n)^5}{n}\right)^{\frac{-d_*}{d_*+\beta\cdot(1\wedge\beta)^q}}\le N\le b_2\cdot\left(\frac{(\log n)^5}{n}\right)^{\frac{-d_*}{d_*+\beta\cdot(1\wedge\beta)^q}},\\
&a_3\cdot\left(\frac{(\log n)^5}{n}\right)^{\frac{-d_*}{d_*+\beta\cdot(1\wedge\beta)^q}}\cdot\log n\le S\le b_3\cdot\left(\frac{(\log n)^5}{n}\right)^{\frac{-d_*}{d_*+\beta\cdot(1\wedge\beta)^q}}\cdot\log n,\\
&\frac{\beta\cdot(1\wedge\beta)^q}{d_*+\beta\cdot(1\wedge\beta)^q}\cdot\log n\le F\le b_4\log n,\ \text{ and } 1\le B\le b_5\cdot n^{\nu}
\end{aligned}
\tag{C.84}
$$

must satisfy

$$
\begin{aligned}
&\sup_{P\in\mathcal{H}^{d,\beta,r}_{4,q,K,d_\star,d_*}}\boldsymbol{E}_{P^{\otimes n}}\left[\mathcal{E}_P^{\phi}\left(\hat{f}^{\mathbf{FNN}}_n\right)\right]\le E_3\cdot\left(\frac{(\log n)^5}{n}\right)^{\frac{\beta\cdot(1\wedge\beta)^q}{d_*+\beta\cdot(1\wedge\beta)^q}}\\
&\text{and}\quad\sup_{P\in\mathcal{H}^{d,\beta,r}_{4,q,K,d_\star,d_*}}\boldsymbol{E}_{P^{\otimes n}}\left[\mathcal{E}_P\left(\hat{f}^{\mathbf{FNN}}_n\right)\right]\le E_3\cdot\left(\frac{(\log n)^5}{n}\right)^{\frac{\beta\cdot(1\wedge\beta)^q}{2d_*+2\beta\cdot(1\wedge\beta)^q}},
\end{aligned}
\tag{C.85}
$$

which will lead to the results of Theorem 2.3.

Let $\boldsymbol{a} := (a_2, a_3) \in (0,\infty)^2$ and $\boldsymbol{b} := (b_1, b_2, b_3, b_4, b_5) \in (0,\infty)^5$ be arbitrary and fixed. Take

$$D_4 = 1 \vee \left(\frac{D_2 K}{a_2}\right)^{\frac{\beta\cdot(1\wedge\beta)^q}{d_*}} \vee \left(\frac{D_3 E_1 K}{D_1 a_3}\right)^{\frac{\beta\cdot(1\wedge\beta)^q}{d_*}},$$

then $D_4 > 0$ only depends on $(\boldsymbol{a}, d_\star, d_*, \beta, r, q, K)$. Hence there exists $E_2 \in (3,\infty)$ only depending on $(\boldsymbol{a}, d_\star, d_*, \beta, r, q, K)$ such that

$$\begin{aligned} 0 < \frac{(\log t)^5}{t} < D_4 \cdot \left(\frac{(\log t)^5}{t}\right)^{\frac{\beta\cdot(1\wedge\beta)^q}{d_*+\beta\cdot(1\wedge\beta)^q}} < 1/4 \\ < 1 < \log t,\ \forall\, t \in [E_2, \infty). \end{aligned} \tag{C.86}$$

From now on we assume that $n \geq E_2$, and (C.84) holds. We have to show that there exists $E_3 \in (0,\infty)$ only depending on $(\boldsymbol{a}, \boldsymbol{b}, \nu, d, d_\star, d_*, \beta, r, q, K)$ such that (C.85) holds.

Let P be an arbitrary probability in $\mathcal{H}^{d,\beta,r}_{4,q,K,d_\star,d_*}$. Denote by η the conditional probability function $x \mapsto P(\{1\}\,|x)$ of P. Then there exists an $\hat\eta \in \mathcal{G}_d^{\mathbf{CHOM}}(q, K, d_\star, d_*, \beta, r)$ such that $\hat\eta = \eta$, P_X-a.s.. Define

$$\zeta := D_4 \cdot \left(\frac{(\log n)^5}{n}\right)^{\frac{\beta\cdot(1\wedge\beta)^q}{d_*+\beta\cdot(1\wedge\beta)^q}}. \tag{C.87}$$

By (C.86), $0 < n^{\frac{-\beta\cdot(1\wedge\beta)^q}{d_*+\beta\cdot(1\wedge\beta)^q}} \leq \zeta < \frac{1}{4}$ and there hold inequalities

$$\log 2 < \log \frac{1-\zeta}{\zeta} \leq \log \frac{1}{\zeta} \leq \log\left(n^{\frac{\beta\cdot(1\wedge\beta)^q}{d_*+\beta\cdot(1\wedge\beta)^q}}\right) \leq F, \tag{C.88}$$

$$\begin{aligned} &D_1 \log \frac{1}{\zeta} \leq D_1 \log\left(n^{\frac{\beta\cdot(1\wedge\beta)^q}{d_*+\beta\cdot(1\wedge\beta)^q}}\right) \\ &\leq D_1 \log n \leq \max\{1, D_1 \log n\} \leq E_1 \log n \leq G, \end{aligned} \tag{C.89}$$

and

$$\begin{aligned} &K D_2 \zeta^{\frac{-d_*/\beta}{(1\wedge\beta)^q}} = K D_2 \cdot D_4^{\frac{-d_*/\beta}{(1\wedge\beta)^q}} \cdot \left(\frac{(\log n)^5}{n}\right)^{\frac{-d_*}{d_*+\beta\cdot(1\wedge\beta)^q}} \\ &\leq K D_2 \cdot \left|\left(\frac{D_2 K}{a_2}\right)^{\frac{\beta\cdot(1\wedge\beta)^q}{d_*}}\right|^{\frac{-d_*/\beta}{(1\wedge\beta)^q}} \cdot \left(\frac{(\log n)^5}{n}\right)^{\frac{-d_*}{d_*+\beta\cdot(1\wedge\beta)^q}} \\ &= a_2 \cdot \left(\frac{(\log n)^5}{n}\right)^{\frac{-d_*}{d_*+\beta\cdot(1\wedge\beta)^q}} \leq N. \end{aligned} \tag{C.90}$$

Consequently,

$$\begin{aligned}
&KD_3\zeta^{\frac{-d_*/\beta}{(1\wedge\beta)^q}}\cdot\log\frac{1}{\zeta}=KD_3\cdot D_4^{\frac{-d_*/\beta}{(1\wedge\beta)^q}}\cdot\left(\frac{(\log n)^5}{n}\right)^{\frac{-d_*}{d_*+\beta\cdot(1\wedge\beta)^q}}\cdot\log\frac{1}{\zeta}\\
&\leq KD_3\cdot\left|\left(\frac{D_3E_1K}{D_1a_3}\right)^{\frac{\beta\cdot(1\wedge\beta)^q}{d_*}}\right|^{\frac{-d_*/\beta}{(1\wedge\beta)^q}}\cdot\left(\frac{(\log n)^5}{n}\right)^{\frac{-d_*}{d_*+\beta\cdot(1\wedge\beta)^q}}\cdot\log\frac{1}{\zeta} \qquad (C.91)\\
&=a_3\cdot\left(\frac{(\log n)^5}{n}\right)^{\frac{-d_*}{d_*+\beta\cdot(1\wedge\beta)^q}}\cdot\frac{D_1\cdot\log\frac{1}{\zeta}}{E_1}\leq a_3\cdot\left(\frac{(\log n)^5}{n}\right)^{\frac{-d_*}{d_*+\beta\cdot(1\wedge\beta)^q}}\cdot\log n\leq S.
\end{aligned}$$

Then it follows from (C.74), (C.87), (C.89), (C.88), (C.90), and (C.91) that

$$\begin{aligned}
&\inf\left\{\mathcal{E}_P^{\phi}(f)\,\middle|\,f\in\mathcal{F}_d^{\mathbf{FNN}}(G,N,S,B,F)\right\}\\
&\leq\inf\left\{\mathcal{E}_P^{\phi}(f)\,\middle|\,f\in\mathcal{F}_d^{\mathbf{FNN}}\left(D_1\log\frac{1}{\zeta},\frac{KD_2}{\zeta^{\frac{d_*/\beta}{(1\wedge\beta)^q}}},\frac{KD_3}{\zeta^{\frac{d_*/\beta}{(1\wedge\beta)^q}}}\cdot\log\frac{1}{\zeta},1,\log\frac{1-\zeta}{\zeta}\right)\right\} \qquad (C.92)\\
&\leq 8\zeta=8D_4\cdot\left(\frac{(\log n)^5}{n}\right)^{\frac{\beta\cdot(1\wedge\beta)^q}{d_*+\beta\cdot(1\wedge\beta)^q}}.
\end{aligned}$$

Besides, from (C.88) we know $\mathrm{e}^F>2$. Hence by taking $\delta_0=\frac{1}{\mathrm{e}^F+1}$ in Lemma C.10, we obtain immediately that there exists

$$\psi:[0,1]^d\times\{-1,1\}\to\left[0,\log\left((10\mathrm{e}^F+10)\cdot\log\left(\mathrm{e}^F+1\right)\right)\right], \qquad (C.93)$$

such that

$$\int_{[0,1]^d\times\{-1,1\}}\psi(x,y)\mathrm{d}P(x,y)=\inf\left\{\mathcal{R}_P^{\phi}(f)\mid f:[0,1]^d\to\mathbb{R}\text{ is measurable}\right\}, \qquad (C.94)$$

and for any measurable $f:[0,1]^d\to[-F,F]$,

$$\begin{aligned}
&\int_{[0,1]^d\times\{-1,1\}}(\phi(yf(x))-\psi(x,y))^2\mathrm{d}P(x,y)\\
&\leq 125000\left|\log\left(1+\mathrm{e}^F\right)\right|^2\cdot\int_{[0,1]^d\times\{-1,1\}}(\phi(yf(x))-\psi(x,y))\,\mathrm{d}P(x,y) \qquad (C.95)\\
&\leq 500000F^2\cdot\int_{[0,1]^d\times\{-1,1\}}(\phi(yf(x))-\psi(x,y))\,\mathrm{d}P(x,y).
\end{aligned}$$

Moreover, it follows from Corollary A.1 with $\gamma=\frac{1}{n}$ that

$$\begin{aligned}
&\log W\leq(S+Gd+1)(2G+5)\log\left((\max\{N,d\}+1)(2nG+2n)B\right)\\
&\leq C_{\boldsymbol{b},d}\cdot(\log n)^2\cdot\left(\frac{(\log n)^5}{n}\right)^{\frac{-d_*}{d_*+\beta\cdot(1\wedge\beta)^q}}\cdot\log\left((\max\{N,d\}+1)(2nG+2n)b_5n^{\nu}\right) \qquad (C.96)\\
&\leq C_{\boldsymbol{b},d,\nu}\cdot(\log n)^3\cdot\left(\frac{(\log n)^5}{n}\right)^{\frac{-d_*}{d_*+\beta\cdot(1\wedge\beta)^q}}=E_4\cdot(\log n)^3\cdot\left(\frac{(\log n)^5}{n}\right)^{\frac{-d_*}{d_*+\beta\cdot(1\wedge\beta)^q}}
\end{aligned}$$

for some constant $E_4 \in (0,\infty)$ only depending on$(\boldsymbol{b}, d, \nu)$, where

$$W = 3 \vee \mathcal{N}\left(\left\{ f|_{[0,1]^d} \middle| f \in \mathcal{F}_d^{\mathbf{FNN}}(G,N,S,B,F)\right\}, \frac{1}{n}\right).$$

Also, note that

$$\sup_{t\in[-F,F]} \phi(t) = \log\left(1+\mathrm{e}^F\right) \le \log\left((10\mathrm{e}^F+10)\cdot\log\left(\mathrm{e}^F+1\right)\right) \le 7F. \tag{C.97}$$

Therefore, by taking $\epsilon = \frac{1}{2}$, $\gamma = \frac{1}{n}$, $\Gamma = 500000F^2$, $M = 7F$, and

$$\mathcal{F} = \left\{ f|_{[0,1]^d} \middle| f \in \mathcal{F}_d^{\mathbf{FNN}}(G,N,S,B,F)\right\}$$

in Theorem 2.1 and combining (C.93), (C.94), (C.95), (C.96), (C.92), we obtain

$$\begin{aligned}
&\boldsymbol{E}_{P^{\otimes n}}\left[\mathcal{E}_P^{\phi}\left(\hat{f}_n^{\mathbf{FNN}}\right)\right] = \boldsymbol{E}_{P^{\otimes n}}\left[\mathcal{R}_P^{\phi}\left(\hat{f}_n^{\mathbf{FNN}}\right) - \int_{[0,1]^d\times\{-1,1\}} \psi(x,y)\mathrm{d}P(x,y)\right]\\
&\le 360\cdot\frac{\Gamma\log W}{n} + \frac{4}{n} + \frac{30M\log W}{n} + 30\cdot\sqrt{\frac{\Gamma\log W}{n^2}} + 2\inf_{f\in\mathcal{F}}\left(\mathcal{R}_P^{\phi}(f) - \int\psi\mathrm{d}P\right)\\
&\le \frac{360\Gamma\log W}{n} + \frac{\Gamma\log W}{n} + \frac{\Gamma\log W}{n} + \frac{\Gamma\log W}{n} + 2\inf_{f\in\mathcal{F}}\mathcal{E}_P^{\phi}(f)\\
&\le \frac{2\cdot 10^8\cdot F^2\cdot\log W}{n} + 16D_4\cdot\left(\frac{(\log n)^5}{n}\right)^{\frac{\beta\cdot(1\wedge\beta)^q}{d_*+\beta\cdot(1\wedge\beta)^q}}\\
&\le \frac{10^9\cdot|b_4\log n|^2\cdot E_4\cdot(\log n)^3\cdot\left(\frac{(\log n)^5}{n}\right)^{\frac{-d_*}{d_*+\beta\cdot(1\wedge\beta)^q}}}{n} + 16D_4\cdot\left(\frac{(\log n)^5}{n}\right)^{\frac{\beta\cdot(1\wedge\beta)^q}{d_*+\beta\cdot(1\wedge\beta)^q}}\\
&= \left(16D_4 + 10^9\cdot|b_4|^2\cdot E_4\right)\cdot\left(\frac{(\log n)^5}{n}\right)^{\frac{\beta\cdot(1\wedge\beta)^q}{d_*+\beta\cdot(1\wedge\beta)^q}} \le E_3\cdot\left(\frac{(\log n)^5}{n}\right)^{\frac{\beta\cdot(1\wedge\beta)^q}{d_*+\beta\cdot(1\wedge\beta)^q}}
\end{aligned}$$

with

$$E_3 := 4\cdot\left(16D_4 + 10^9\cdot|b_4|^2\cdot E_4\right) + 4$$

only depending on $(\boldsymbol{a}, \boldsymbol{b}, \nu, d, d_\star, d_*, \beta, r, q, K)$. We then apply the calibration inequality (2.21) and conclude that

$$\begin{aligned}
&\boldsymbol{E}_{P^{\otimes n}}\left[\mathcal{E}_P\left(\hat{f}_n^{\mathbf{FNN}}\right)\right] \le 2\sqrt{2}\cdot\boldsymbol{E}_{P^{\otimes n}}\left[\sqrt{\mathcal{E}_P^{\phi}\left(\hat{f}_n^{\mathbf{FNN}}\right)}\right] \le 4\cdot\sqrt{\boldsymbol{E}_{P^{\otimes n}}\left[\mathcal{E}_P^{\phi}\left(\hat{f}_n^{\mathbf{FNN}}\right)\right]}\\
&\le 4\cdot\sqrt{\left(16D_4 + 10^9\cdot|b_4|^2\cdot E_4\right)\cdot\left(\frac{(\log n)^5}{n}\right)^{\frac{\beta\cdot(1\wedge\beta)^q}{d_*+\beta\cdot(1\wedge\beta)^q}}}\\
&\le E_3\cdot\left(\frac{(\log n)^5}{n}\right)^{\frac{\beta\cdot(1\wedge\beta)^q}{2d_*+2\beta\cdot(1\wedge\beta)^q}}.
\end{aligned} \tag{C.98}$$

Since P is arbitrary, the desired bound (C.85) follows. Setting $\mathrm{c} = E_1$ completes the proof of Theorem 2.3.

Now it remains to show Theorem 2.2. Indeed, it follows from (2.33) that

$$\mathcal{H}_1^{d,\beta,r} \subset \mathcal{H}_{4,0,1,1,d}^{d,\beta,r}.$$

Then by taking $q = 0$, $d_* = d$ and $d_\star = K = 1$ in Theorem 2.3, we obtain that there exists a constant $\mathrm{c} \in (0, \infty)$ only depending on (d, β, r) such that the estimator $\hat{f}_n^{\mathbf{FNN}}$ defined by (2.14) with

$$\begin{aligned}
&\mathrm{c}\log n \le G \lesssim \log n,\ N \asymp \left(\frac{(\log n)^5}{n}\right)^{\frac{-d}{d+\beta\cdot(1\wedge\beta)^0}} = \left(\frac{(\log n)^5}{n}\right)^{\frac{-d}{d+\beta}},\\
&S \asymp \left(\frac{(\log n)^5}{n}\right)^{\frac{-d}{d+\beta\cdot(1\wedge\beta)^0}}\cdot\log n = \left(\frac{(\log n)^5}{n}\right)^{\frac{-d}{d+\beta}}\cdot\log n,\\
&1 \le B \lesssim n^\nu,\ \text{and}\ \ \frac{\beta}{d+\beta}\cdot\log n = \frac{\beta\cdot(1\wedge\beta)^0}{d+\beta\cdot(1\wedge\beta)^0}\cdot\log n \le F \lesssim \log n
\end{aligned}$$

must satisfy

$$\begin{aligned}
&\sup_{P\in\mathcal{H}_1^{d,\beta,r}} \boldsymbol{E}_{P^{\otimes n}}\left[\mathcal{E}_P^\phi\left(\hat{f}_n^{\mathbf{FNN}}\right)\right] \le \sup_{P\in\mathcal{H}_{4,0,1,1,d}^{d,\beta,r}} \boldsymbol{E}_{P^{\otimes n}}\left[\mathcal{E}_P^\phi\left(\hat{f}_n^{\mathbf{FNN}}\right)\right]\\
&\lesssim \left(\frac{(\log n)^5}{n}\right)^{\frac{\beta\cdot(1\wedge\beta)^0}{d+\beta\cdot(1\wedge\beta)^0}} = \left(\frac{(\log n)^5}{n}\right)^{\frac{\beta}{d+\beta}}
\end{aligned}$$

and

$$\begin{aligned}
&\sup_{P\in\mathcal{H}_1^{d,\beta,r}} \boldsymbol{E}_{P^{\otimes n}}\left[\mathcal{E}_P\left(\hat{f}_n^{\mathbf{FNN}}\right)\right] \le \sup_{P\in\mathcal{H}_{4,0,1,1,d}^{d,\beta,r}} \boldsymbol{E}_{P^{\otimes n}}\left[\mathcal{E}_P\left(\hat{f}_n^{\mathbf{FNN}}\right)\right]\\
&\lesssim \left(\frac{(\log n)^5}{n}\right)^{\frac{\beta\cdot(1\wedge\beta)^0}{2d+2\beta\cdot(1\wedge\beta)^0}} = \left(\frac{(\log n)^5}{n}\right)^{\frac{\beta}{2d+2\beta}}.
\end{aligned}$$

This completes the proof of Theorem 2.2. ■

C.5 Proof of Theorem 2.5

Appendix C.5 is devoted to the proof of Theorem 2.5. To this end, we need the following lemmas. Note that the logistic loss is given by $\phi(t) = \log(1 + \mathrm{e}^{-t})$ with $\phi'(t) = -\frac{1}{1+\mathrm{e}^t} \in (-1, 0)$ and $\phi''(t) = \frac{\mathrm{e}^t}{(1+\mathrm{e}^t)^2} = \frac{1}{\mathrm{e}^t+\mathrm{e}^{-t}+2} \in (0, \frac{1}{4}]$ for all $t \in \mathbb{R}$.

Lemma C.16 *Let $\eta_0 \in (0,1)$, $F_0 \in \left(0, \log\frac{1+\eta_0}{1-\eta_0}\right)$, $a \in [-F_0, F_0]$, $\phi(t) = \log(1 + \mathrm{e}^{-t})$ be the logistic loss, $d \in \mathbb{N}$, and P be a Borel probability measure on $[0,1]^d \times \{-1, 1\}$ of which the conditional probability function $[0,1]^d \ni z \mapsto P(\{1\}\,|z) \in [0,1]$ is denoted by η. Then for* *any* $x \in [0,1]^d$ *such that* $|2\eta(x)-1| > \eta_0$*, there holds*

$$\begin{aligned}0 &\le |a - F_0\mathrm{sgn}(2\eta(x)-1)| \cdot \left(\frac{1-\eta_0}{2}\phi'(-F_0) - \frac{\eta_0+1}{2}\phi'(F_0)\right)\\ &\le |a - F_0\mathrm{sgn}(2\eta(x)-1)| \cdot \left(\frac{1-\eta_0}{2}\phi'(-F_0) - \frac{\eta_0+1}{2}\phi'(F_0)\right)\\ &\quad + \frac{1}{2\left(\mathrm{e}^{-F_0}+\mathrm{e}^{F_0}+2\right)}\left|a - F_0\mathrm{sgn}(2\eta(x)-1)\right|^2\\ &\le \int_{\{-1,1\}} \left(\phi(ya) - \phi(yF_0\mathrm{sgn}(2\eta(x)-1))\right)\mathrm{d}P(y|x)\\ &\le |a - F_0\mathrm{sgn}(2\eta(x)-1)| + F_0^2.\end{aligned} \tag{C.99}$$

Proof Given $x \in [0,1]^d$, recall the function V_x defined in the proof of Lemma C.7. By Taylor expansion, there exists ξ between a and $F_0\mathrm{sgn}(2\eta(x)-1)$ such that

$$\begin{aligned}&\int_{\{-1,1\}} \left(\phi(ya) - \phi(yF_0\mathrm{sgn}(2\eta(x)-1))\right)\mathrm{d}P(y|x)\\ &= V_x(a) - V_x(F_0\mathrm{sgn}(2\eta(x)-1))\\ &= (a - F_0\mathrm{sgn}(2\eta(x)-1)) \cdot V_x'(F_0\mathrm{sgn}(2\eta(x)-1)) + \frac{1}{2}\left|a - F_0\mathrm{sgn}(2\eta(x)-1)\right|^2 \cdot V_x''(\xi).\end{aligned} \tag{C.100}$$

Since $\xi \in [-F_0, F_0]$, we have

$$\begin{aligned}0 &\le \frac{1}{\mathrm{e}^{-F_0}+\mathrm{e}^{F_0}+2} = \inf\left\{\phi''(t) \mid t \in [-F_0, F_0]\right\}\\ &\le V_x''(\xi) = \eta(x)\phi''(\xi) + (1-\eta(x))\phi''(-\xi) \le \frac{1}{4}\end{aligned}$$

and then

$$\begin{aligned}0 &\le \frac{1}{2}\left|a - F_0\mathrm{sgn}(2\eta(x)-1)\right|^2 \cdot \frac{1}{\mathrm{e}^{-F_0}+\mathrm{e}^{F_0}+2}\\ &\le \frac{1}{2}\left|a - F_0\mathrm{sgn}(2\eta(x)-1)\right|^2 \cdot V_x''(\xi)\\ &\le \frac{1}{2}\left(|a|+F_0\right)^2 \cdot \frac{1}{4} \le \frac{1}{2}F_0^2.\end{aligned} \tag{C.101}$$

On the other hand, if $2\eta(x)-1 > \eta_0$, then

$$\begin{aligned}&(a - F_0\mathrm{sgn}(2\eta(x)-1)) \cdot V_x'(F_0\mathrm{sgn}(2\eta(x)-1))\\ &= (a - F_0)\left(\eta(x)\phi'(F_0) - (1-\eta(x))\phi'(-F_0)\right)\\ &= |a - F_0|\left((1-\eta(x))\phi'(-F_0) - \eta(x)\phi'(F_0))\right)\\ &\ge |a - F_0|\left(\left(1 - \frac{1+\eta_0}{2}\right)\phi'(-F_0) - \frac{1+\eta_0}{2}\phi'(F_0)\right)\\ &= |a - F_0\mathrm{sgn}(2\eta(x)-1)| \cdot \left(\frac{1-\eta_0}{2}\phi'(-F_0) - \frac{1+\eta_0}{2}\phi'(F_0)\right).\end{aligned}$$

Similarly, if $2\eta(x) - 1 < -\eta_0$, then

$$\begin{aligned}
&(a - F_0 \mathrm{sgn}(2\eta(x) - 1)) \cdot V_x'(F_0 \mathrm{sgn}(2\eta(x) - 1)) \\
&= (a + F_0)\left(\eta(x)\phi'(-F_0) - (1 - \eta(x))\phi'(F_0)\right) \\
&= |a + F_0| \left(\eta(x)\phi'(-F_0) - (1 - \eta(x))\phi'(F_0))\right) \\
&\geq |a + F_0| \left(\frac{1 - \eta_0}{2}\phi'(-F_0) - \left(1 - \frac{1 - \eta_0}{2}\right)\phi'(F_0)\right) \\
&= |a - F_0 \mathrm{sgn}(2\eta(x) - 1)| \cdot \left(\frac{1 - \eta_0}{2}\phi'(-F_0) - \frac{1 + \eta_0}{2}\phi'(F_0)\right).
\end{aligned}$$

Therefore, for given $x \in [0,1]^d$ satisfying $|2\eta(x) - 1| > \eta_0$, there always holds

$$\begin{aligned}
&(a - F_0 \mathrm{sgn}(2\eta(x) - 1)) \cdot V_x'(F_0 \mathrm{sgn}(2\eta(x) - 1)) \\
&\quad \geq |a - F_0 \mathrm{sgn}(2\eta(x) - 1)| \cdot \left(\frac{1 - \eta_0}{2}\phi'(-F_0) - \frac{1 + \eta_0}{2}\phi'(F_0)\right).
\end{aligned} \tag{C.102}$$

We next show that $\frac{1-\eta_0}{2}\phi'(-F_0) - \frac{1+\eta_0}{2}\phi'(F_0) > 0$. Indeed, let $g(t) = \frac{1-\eta_0}{2}\phi'(-t) - \frac{1+\eta_0}{2}\phi'(t)$. Then $g'(t) = -\frac{1-\eta_0}{2}\phi''(-t) - \frac{1+\eta_0}{2}\phi''(t) < 0$, i.e., g is strictly decreasing, and thus

$$\frac{1 - \eta_0}{2}\phi'(-F_0) - \frac{1 + \eta_0}{2}\phi'(F_0) = g(F_0) > g\left(\log\frac{1 + \eta_0}{1 - \eta_0}\right) = 0. \tag{C.103}$$

Moreover, we also have

$$\begin{aligned}
&(a - F_0 \mathrm{sgn}(2\eta(x) - 1)) \cdot V_x'(F_0 \mathrm{sgn}(2\eta(x) - 1)) \\
&\leq |a - F_0 \mathrm{sgn}(2\eta(x) - 1)| \cdot |V_x'(F_0 \mathrm{sgn}(2\eta(x) - 1))| \\
&= |a - F_0 \mathrm{sgn}(2\eta(x) - 1)| \cdot |\eta(x)\phi'(F_0 \mathrm{sgn}(2\eta(x) - 1)) - (1 - \eta(x))\phi'(-F_0 \mathrm{sgn}(2\eta(x) - 1))| \\
&\leq |a - F_0 \mathrm{sgn}(2\eta(x) - 1)| \, |\eta(x) + (1 - \eta(x))| = |a - F_0 \mathrm{sgn}(2\eta(x) - 1)|.
\end{aligned} \tag{C.104}$$

Then the first inequality of (C.99) is from (C.103), the third inequality of (C.99) is due to (C.100), (C.101) and (C.102), and the last inequality of (C.99) is from (C.100), (C.101) and (C.104). Thus we complete the proof. ∎

Lemma C.17 *Let* $\eta_0 \in (0,1)$, $F_0 \in \left(0, \log\frac{1+\eta_0}{1-\eta_0}\right)$, $d \in \mathbb{N}$, *and* P *be a Borel probability measure on* $[0,1]^d \times \{-1,1\}$ *of which the conditional probability function* $[0,1]^d \ni z \mapsto P(\{1\}|z) \in [0,1]$ *is denoted by* η. *Define*

$$\begin{aligned}
&\psi : [0,1]^d \times \{-1,1\} \to \mathbb{R}, \\
&\qquad (x,y) \mapsto \begin{cases} \phi\left(yF_0 \mathrm{sgn}(2\eta(x) - 1)\right), & \text{if } |2\eta(x) - 1| > \eta_0, \\ \phi\left(y \log\dfrac{\eta(x)}{1 - \eta(x)}\right), & \text{if } |2\eta(x) - 1| \leq \eta_0. \end{cases}
\end{aligned} \tag{C.105}$$

Then there hold

$$\begin{aligned}&\int_{[0,1]^d\times\{-1,1\}}\left(\phi\left(yf(x)\right)-\psi(x,y)\right)^2\mathrm{d}P(x,y)\\&\leq\frac{8}{1-\eta_0^2}\cdot\int_{[0,1]^d\times\{-1,1\}}\left(\phi\left(yf(x)\right)-\psi(x,y)\right)\mathrm{d}P(x,y)\end{aligned}\tag{C.106}$$

for any measurable $f:[0,1]^d\to[-F_0,F_0]$ *, and*

$$0\leq\psi(x,y)\leq\log\frac{2}{1-\eta_0},\quad\forall(x,y)\in[0,1]^d\times\{-1,1\}.\tag{C.107}$$

Proof Recall that given $x\in[0,1]^d$, $V_x(t)=\eta(x)\phi(t)+(1-\eta(x))\phi(-t),\forall t\in\mathbb{R}$. Due to inequality (C.99) and Lemma C.7, for any measurable $f:[0,1]^d\to[-F_0,F_0]$, we have

$$\begin{aligned}&\int_{[0,1]^d\times\{-1,1\}}\left(\phi\left(yf(x)\right)-\psi(x,y)\right)\mathrm{d}P(x,y)\\&=\int_{|2\eta(x)-1|>\eta_0}\int_{\{-1,1\}}\phi\left(yf(x)\right)-\phi\left(yF_0\mathrm{sgn}(2\eta(x)-1)\right)\mathrm{d}P(y|x)\mathrm{d}P_X(x)\\&\quad+\int_{|2\eta(x)-1|\leq\eta_0}\int_{\{-1,1\}}\phi\left(yf(x)\right)-\phi\left(y\log\frac{\eta(x)}{1-\eta(x)}\right)\mathrm{d}P(y|x)\mathrm{d}P_X(x)\\&\geq\int_{|2\eta(x)-1|>\eta_0}\frac{1}{2\left(\mathrm{e}^{F_0}+\mathrm{e}^{-F_0}+2\right)}\left|f(x)-F_0\mathrm{sgn}(2\eta(x)-1)\right|^2\mathrm{d}P_X(x)\\&\quad+\int_{|2\eta(x)-1|\leq\eta_0}\left[\inf_{t\in\left[\log\frac{1-\eta_0}{1+\eta_0},\log\frac{1+\eta_0}{1-\eta_0}\right]}\frac{1}{2(\mathrm{e}^{t}+\mathrm{e}^{-t}+2)}\right]\left|f(x)-\log\frac{\eta(x)}{1-\eta(x)}\right|^2\mathrm{d}P_X(x)\\&\geq\frac{1}{2}\frac{1}{\frac{1+\eta_0}{1-\eta_0}+\frac{1-\eta_0}{1+\eta_0}+2}\int_{\{|2\eta(x)-1|>\eta_0\}\times\{-1,1\}}\left|\phi\left(yf(x)\right)-\phi\left(yF_0\mathrm{sgn}(2\eta(x)-1)\right)\right|^2\mathrm{d}P(x,y)\\&\quad+\frac{1}{2}\frac{1}{\frac{1+\eta_0}{1-\eta_0}+\frac{1-\eta_0}{1+\eta_0}+2}\int_{\{|2\eta(x)-1|\leq\eta_0\}\times\{-1,1\}}\left|\phi\left(yf(x)\right)-\phi\left(y\log\frac{\eta(x)}{1-\eta(x)}\right)\right|^2\mathrm{d}P(x,y)\\&=\frac{1-\eta_0^2}{8}\cdot\int_{[0,1]^d\times\{-1,1\}}\left(\phi\left(yf(x)\right)-\psi(x,y)\right)^2\mathrm{d}P(x,y),\end{aligned}$$

where the second inequality is from (C.16) and the fact that $F_0\in\left(0,\log\frac{1+\eta_0}{1-\eta_0}\right)$. Thus we have proved the inequality (C.106).

On the other hand, from the definition of ψ as well as $F_0\in\left(0,\log\frac{1+\eta_0}{1-\eta_0}\right)$, we also have

$$0\leq\psi(x,y)\leq\max\left\{\phi(-F_0),\phi\left(-\log\frac{1+\eta_0}{1-\eta_0}\right)\right\}\leq\phi\left(-\log\frac{1+\eta_0}{1-\eta_0}\right)=\log\frac{2}{1-\eta_0},$$

which gives the inequality (C.107). The proof is completed. ■

Now we are in the position to prove Theorem 2.5.

Proof [Proof of Theorem 2.5] Let $\eta_0\in(0,1)\cap[0,t_1]$, $F_0\in(0,\log\frac{1+\eta_0}{1-\eta_0})\cap[0,1]$, $\xi\in(0,\frac{1}{2}\wedge t_2]$ and $P\in\mathcal{H}^{d,\beta,r,I,\Theta,s_1,s_2}_{6,t_1,c_1,t_2,c_2}$ be arbitrary. Denote by η the conditional probability function $P(\{1\}\,|\cdot)$ of P. By definition, there exists a classifier $\mathtt{C}\in\mathcal{C}^{d,\beta,r,I,\Theta}$ such that (2.24), (2.50) and (2.51) hold. According to Proposition A.4 and the proof of Theorem 3.4 in Kim et al. (2021), there exist positive constants G_0,N_0,S_0,B_0 only depending on d,β,r,I,Θ and $\tilde{f}_0\in\mathcal{F}^{\mathbf{FNN}}_d(G_\xi,N_\xi,S_\xi,B_\xi,1)$ such that $\tilde{f}_0(x)=\mathtt{C}(x)$ for $x\in[0,1]^d$ with $\Delta_{\mathtt{C}}(x)>\xi$, where

$$G_\xi=G_0\log\frac{1}{\xi},\ N_\xi=N_0\left(\frac{1}{\xi}\right)^{\frac{d-1}{\beta}},\ S_\xi=S_0\left(\frac{1}{\xi}\right)^{\frac{d-1}{\beta}}\log\left(\frac{1}{\xi}\right),\ B_\xi=\left(\frac{B_0}{\xi}\right). \tag{C.108}$$

Define $\psi:[0,1]^d\times\{-1,1\}\to\mathbb{R}$ by (C.105). Then for any measurable function $f:[0,1]^d\to[-F_0,F_0]$, there holds

$$\begin{aligned}
&\mathcal{E}_P(f)=\mathcal{E}_P\left(\frac{f}{F_0}\right)\le\int_{[0,1]^d}\left|\frac{f(x)}{F_0}-\operatorname{sgn}(2\eta(x)-1)\right||2\eta(x)-1|\,\mathrm{d}P_X(x)\\
&\le 2P\left(|2\eta(x)-1|\le\eta_0\right)+\int_{|2\eta(x)-1|>\eta_0}\left|\frac{f(x)}{F_0}-\operatorname{sgn}(2\eta(x)-1)\right|\mathrm{d}P_X(x)\\
&\le 2c_1\eta_0^{s_1}+\frac{1}{F_0}\int_{|2\eta(x)-1|>\eta_0}|f(x)-F_0\operatorname{sgn}(2\eta(x)-1)|\,\mathrm{d}P_X(x)\\
&\le 2c_1\eta_0^{s_1}+\int_{|2\eta(x)-1|>\eta_0}\frac{\int(\phi(yf(x))-\phi(yF_0\operatorname{sgn}(2\eta(x)-1)))\,\mathrm{d}P(y|x)}{F_0\cdot\left(\frac{1-\eta_0}{2}\phi'(-F_0)-\frac{\eta_0+1}{2}\phi'(F_0)\right)}\mathrm{d}P_X(x)\\
&\le 2c_1\eta_0^{s_1}+\frac{\int_{[0,1]^d\times\{-1,1\}}(\phi(yf(x))-\psi(x,y))\,\mathrm{d}P(x,y)}{F_0\cdot\left(\frac{1-\eta_0}{2}\phi'(-F_0)-\frac{\eta_0+1}{2}\phi'(F_0)\right)},
\end{aligned}\tag{C.109}$$

where the first inequality is from Theorem 2.31 of Steinwart and Christmann (2008), the third inequality is due to the noise condition (2.24), and the fourth inequality is from (C.99) in Lemma C.16.

Let $\mathcal{F}=\mathcal{F}^{\mathbf{FNN}}_d(G_\xi,N_\xi,S_\xi,B_\xi,F_0)$ with $(G_\xi,N_\xi,S_\xi,B_\xi)$ given by (C.108), $\Gamma=\frac{8}{1-\eta_0^2}$ and $M=\frac{2}{1-\eta_0}$ in Theorem 2.1. Then we will use this theorem to derive the desired generalization bounds for the ϕ-ERM $\hat{f}_n:=\hat{f}^{\mathbf{FNN}}_n$ over $\mathcal{F}^{\mathbf{FNN}}_d(G_\xi,N_\xi,S_\xi,B_\xi,F_0)$. Indeed, Lemma C.17 guarantees that the conditions (2.3), (2.4) and (2.5) of Theorem 2.1 are satisfied. Moreover, take $\gamma=\frac{1}{n}$. Then $W=\max\{3,\ \mathcal{N}(\mathcal{F},\gamma)\}$ satisfies

$$\log W\le C_{d,\beta,r,I,\Theta}\xi^{-\frac{d-1}{\beta}}\left(\log\frac{1}{\xi}\right)^2\left(\log\frac{1}{\xi}+\log n\right).$$

Thus the expectation of $\int_{[0,1]^d\times\{-1,1\}}\left(\phi(y\hat{f}_n(x))-\psi(x,y)\right)\mathrm{d}P(x,y)$ can be bounded by inequality (2.6) in Theorem 2.1 as

$$\begin{aligned}&\boldsymbol{E}_{P^{\otimes n}}\left[\int_{[0,1]^d\times\{-1,1\}}\left(\phi(y\hat{f}_n(x))-\psi(x,y)\right)\mathrm{d}P(x,y)\right]\\&\leq\frac{4000C_{d,\beta,r,I,\Theta}\xi^{-\frac{d-1}{\beta}}\left(\log\frac{1}{\xi}\right)^2\left(\log\frac{1}{\xi}+\log n\right)}{n(1-\eta_0^2)}\\&\qquad+2\inf_{f\in\mathcal{F}}\left(\mathcal{R}_P^\phi(f)-\int_{[0,1]^d\times\{-1,1\}}\psi(x,y)\mathrm{d}P(x,y)\right).\end{aligned}\tag{C.110}$$

We next estimate the approximation error, i.e., the second term on the right hand side of (C.110). Take $f_0=F_0\tilde{f}_0\in\mathcal{F}$ where $\tilde{f}_0\in\mathcal{F}_d^{\mathbf{FNN}}(G_\xi,N_\xi,S_\xi,B_\xi,1)$ satisfying $\tilde{f}_0(x)=\mathtt{C}(x)$ for $x\in[0,1]^d$ with $\Delta_{\mathtt{C}}(x)>\xi$. Then one can bound the approximation error as

$$\begin{aligned}&\inf_{f\in\mathcal{F}}\left(\mathcal{R}_P^\phi(f)-\int_{[0,1]^d\times\{-1,1\}}\psi(x,y)\mathrm{d}P(x,y)\right)\\&\leq\mathcal{R}_P^\phi(f_0)-\int_{[0,1]^d\times\{-1,1\}}\psi(x,y)\mathrm{d}P(x,y)=I_1+I_2+I_3,\end{aligned}\tag{C.111}$$

where

$$\begin{aligned}I_1&:=\int_{\{|2\eta(x)-1|>\eta_0,\Delta_{\mathtt{C}}(x)>\xi\}\times\{-1,1\}}\phi(yf_0(x))-\phi(yF_0\mathrm{sgn}(2\eta(x)-1))\mathrm{d}P(x,y),\\I_2&:=\int_{\{|2\eta(x)-1|\leq\eta_0\}\times\{-1,1\}}\phi(yf_0(x))-\phi\left(y\log\frac{\eta(x)}{1-\eta(x)}\right)\mathrm{d}P(x,y),\\I_3&:=\int_{\{|2\eta(x)-1|>\eta_0,\Delta_{\mathtt{C}}(x)\leq\xi\}\times\{-1,1\}}\phi(yf_0(x))-\phi(yF_0\mathrm{sgn}(2\eta(x)-1))\mathrm{d}P(x,y).\end{aligned}$$

Note that $f_0(x)=F_0\tilde{f}_0(x)=F_0\mathtt{C}(x)=F_0\mathrm{sgn}(2\eta(x)-1)$ for P_X-almost all $x\in[0,1]^d$ with $\Delta_{\mathtt{C}}(x)>\xi$. Thus it follows that $I_1=0$. On the other hand, from Lemma C.7 and the noise condition (2.24), we see that

$$\begin{aligned}I_2&\leq\int_{\{|2\eta(x)-1|\leq\eta_0\}\times\{-1,1\}}\left|f_0(x)-\log\frac{\eta(x)}{1-\eta(x)}\right|^2\mathrm{d}P(x,y)\\&\leq\int_{\left\{|2\eta_(x)-1|\leq\eta_0\right\}\times\{-1,1\}}\left(F_0+\log\frac{1+\eta_0}{1-\eta_0}\right)^2\mathrm{d}P(x,y)\leq4\left(\log\frac{1+\eta_0}{1-\eta_0}\right)^2c_1\cdot\eta_0^{s_1}.\end{aligned}\tag{C.112}$$

Moreover, due to Lemma C.16 and the margin condition (2.51), we have

$$\begin{aligned}I_3&\leq\int_{\{|2\eta(x)-1|>\eta_0,\Delta_{\mathtt{C}}(x)\leq\xi\}}\left(2F_0+F_0^2\right)\mathrm{d}P_X(x)\\&\leq3F_0\cdot P_X(\left\{x\in[0,1]^d\middle|\Delta_{\mathtt{C}}(x)\leq\xi\right\})\leq3F_0\cdot c_2\cdot\xi^{s_2}.\end{aligned}\tag{C.113}$$

The estimates above together with (C.109) and (C.110) give

$$\begin{aligned}
&\boldsymbol{E}_{P^{\otimes n}}\left[\mathcal{E}_P(\hat{f}_n)\right]\\
&\leq 2c_1\eta_0^{s_1}+\frac{1}{F_0}\cdot\frac{\boldsymbol{E}_{P^{\otimes n}}\left[\int_{[0,1]^d\times\{-1,1\}}\left(\phi(y\hat{f}_n(x))-\psi(x,y)\right)\mathrm{d}P(x,y)\right]}{\frac{1-\eta_0}{2}\phi'(-F_0)-\frac{\eta_0+1}{2}\phi'(F_0)}\\
&\leq 2c_1\eta_0^{s_1}+\frac{8\left|\log\frac{1+\eta_0}{1-\eta_0}\right|^2c_1\eta_0^{s_1}+6F_0c_2\xi^{s_2}+\frac{4000C_{d,\beta,r,I,\Theta}\xi^{-\frac{d-1}{\beta}}\left(\log\frac{1}{\xi}\right)^2\left(\log\frac{1}{\xi}+\log n\right)}{n(1-\eta_0^2)}}{F_0\cdot\left(\frac{1-\eta_0}{2}\phi'(-F_0)-\frac{\eta_0+1}{2}\phi'(F_0)\right)}.
\end{aligned}\tag{C.114}$$

Since P is arbitrary, we can take the supremum over all $P\in\mathcal{H}_{6,t_1,c_1,t_2,c_2}^{d,\beta,r,I,\Theta,s_1,s_2}$ to obtain from (C.114) that

$$\begin{aligned}
&\sup_{P\in\mathcal{H}_{6,t_1,c_1,t_2,c_2}^{d,\beta,r,I,\Theta,s_1,s_2}}\boldsymbol{E}_{P^{\otimes n}}\left[\mathcal{E}_P(\hat{f}_n^{\mathbf{FNN}})\right]\\
&\leq 2c_1\eta_0^{s_1}+\frac{8\left|\log\frac{1+\eta_0}{1-\eta_0}\right|^2c_1\eta_0^{s_1}+6F_0c_2\xi^{s_2}+\frac{4000C_{d,\beta,r,I,\Theta}\xi^{-\frac{d-1}{\beta}}\left(\log\frac{1}{\xi}\right)^2\left(\log\frac{1}{\xi}+\log n\right)}{n(1-\eta_0^2)}}{F_0\cdot\left(\frac{1-\eta_0}{2}\phi'(-F_0)-\frac{\eta_0+1}{2}\phi'(F_0)\right)}.
\end{aligned}\tag{C.115}$$

(C.115) holds for all $\eta_0\in(0,1)\cap[0,t_1]$, $F_0\in(0,\log\frac{1+\eta_0}{1-\eta_0})\cap[0,1]$, $\xi\in(0,\frac{1}{2}\wedge t_2]$. We then take suitable η_0, F_0, and ξ in (C.115) to derive the convergence rates stated in Theorem 2.5.

$$\begin{aligned}
&\sup_{P\in\mathcal{H}_{6,t_1,c_1,t_2,c_2}^{d,\beta,r,I,\Theta,s_1,s_2}}\boldsymbol{E}_{P^{\otimes n}}\left[\mathcal{E}_P(\hat{f}_n^{\mathbf{FNN}})\right]\\
&\leq 2c_1\eta_0^{s_1}+\frac{8\left|\log\frac{1+\eta_0}{1-\eta_0}\right|^2c_1\cdot\eta_0^{s_1}+6F_0c_2\xi^{s_2}+\frac{4000C_{d,\beta,r,I,\Theta}\xi^{\frac{d-1}{\beta}}\left(\log\frac{1}{\xi}\right)^2\left(\log\frac{1}{\xi}+\log n\right)}{n(1-\eta_0^2)}}{F_0\cdot\left(\frac{1-\eta_0}{2}\phi'(-F_0)-\frac{\eta_0+1}{2}\phi'(F_0)\right)}.
\end{aligned}\tag{C.116}$$

Case I. When $s_1=s_2=\infty$, taking $\eta_0=F_0=t_1\wedge\frac{1}{2}$ and $\xi=t_2\wedge\frac{1}{2}$ in (C.115) yields

$$\sup_{P\in\mathcal{H}_{6,t_1,c_1,t_2,c_2}^{d,\beta,r,I,\Theta,s_1,s_2}}\boldsymbol{E}_{P^{\otimes n}}\left[\mathcal{E}_P\left(\hat{f}_n^{\mathbf{FNN}}\right)\right]\lesssim\frac{\log n}{n}.$$

Case II. When $s_1=\infty$ and $s_2<\infty$, taking $\eta_0=F_0=t_1\wedge\frac{1}{2}$ and $\xi\asymp\left(\frac{(\log n)^3}{n}\right)^{\frac{1}{s_2+\frac{d-1}{\beta}}}$ in (C.115) yields

$$\sup_{P\in\mathcal{H}_{6,t_1,c_1,t_2,c_2}^{d,\beta,r,I,\Theta,s_1,s_2}}\boldsymbol{E}_{P^{\otimes n}}\left[\mathcal{E}_P\left(\hat{f}_n^{\mathbf{FNN}}\right)\right]\lesssim\left(\frac{(\log n)^3}{n}\right)^{\frac{1}{1+\frac{d-1}{\beta s_2}}}.$$

Case III. When $s_1<\infty$ and $s_2=\infty$, take $\eta_0=F_0\asymp\left(\frac{\log n}{n}\right)^{\frac{1}{s_1+2}}$ and $\xi=t_2\wedge\frac{1}{2}$ in (C.115). From the fact that $\frac{\eta_0}{4}\leq\frac{1-\eta_0}{2}\phi'(-\eta_0)-\frac{\eta_0+1}{2}\phi'(\eta_0)\leq\eta_0,\forall 0\leq\eta_0\leq 1$, the item in the denominator of the second term on the right hand side of (C.115) is larger than $\frac{1}{4}\eta_0^2$. Then we have

$$\sup_{P\in\mathcal{H}_{6,t_1,c_1,t_2,c_2}^{d,\beta,r,I,\Theta,s_1,s_2}} \boldsymbol{E}_{P^{\otimes n}}\left[\mathcal{E}_P\left(\hat{f}_n^{\mathbf{FNN}}\right)\right] \lesssim \left(\frac{\log n}{n}\right)^{\frac{s_1}{s_1+2}}.$$

Case IV. When $s_1 < \infty$ and $s_2 < \infty$, taking

$$\eta_0 = F_0 \asymp \left(\frac{(\log n)^3}{n}\right)^{\frac{s_2}{s_2+(s_1+1)\left(s_2+\frac{d-1}{\beta}\right)}} \text{ and } \xi \asymp \left(\frac{(\log n)^3}{n}\right)^{\frac{s_1+1}{s_2+(s_1+1)\left(s_2+\frac{d-1}{\beta}\right)}}$$

in (C.115) yields

$$\sup_{P\in\mathcal{H}_{6,t_1,c_1,t_2,c_2}^{d,\beta,r,I,\Theta,s_1,s_2}} \boldsymbol{E}_{P^{\otimes n}}\left[\mathcal{E}_P\left(\hat{f}_n^{\mathbf{FNN}}\right)\right] \lesssim \left(\frac{(\log n)^3}{n}\right)^{\frac{s_1}{1+(s_1+1)\left(1+\frac{d-1}{\beta s_2}\right)}}.$$

Combining above cases, we obtain the desired results. The proof of Theorem 2.5 is completed. ∎

C.6 Proof of Theorem 2.6 and Corollary 2.1

In Appendix C.6, we provide the proof of Theorem 2.6 and Corollary 2.1. Hereinafter, for $a \in \mathbb{R}^d$ and $R \in \mathbb{R}$, we define $\mathscr{B}(a, R) := \left\{x \in \mathbb{R}^d \middle| \|x-a\|_2 \le R\right\}$.

Lemma C.18 *Let $d \in \mathbb{N}$, $\beta \in (0,\infty)$, $r \in (0,\infty)$, $Q \in \mathbb{N}\cap(1,\infty)$,*

$$G_{Q,d} := \left\{ \left(\frac{k_1}{2Q},\dots,\frac{k_d}{2Q}\right)^\top \middle| k_1,\dots,k_d \text{ are odd integers}\right\} \cap [0,1]^d,$$

and $T : G_{Q,d} \to \{-1, 1\}$ be a map. Then there exist a constant $\mathrm{c}_1 \in (0, \frac{1}{9999})$ only depending on (d, β, r), and an $f \in \mathcal{B}_r^\beta\left([0,1]^d\right)$ depending on (d,β,r,Q,T), such that $\|f\|_{[0,1]^d} = \frac{\mathrm{c}_1}{Q^\beta}$, and

$$f(x) = \|f\|_{[0,1]^d} \cdot T(a) = \frac{\mathrm{c}_1}{Q^\beta}\cdot T(a),\ \forall\, a \in G_{Q,d},\ x \in \mathscr{B}(a, \frac{1}{5Q}) \cap [0,1]^d.$$

Proof Let

$$\kappa : \mathbb{R} \to [0,1], t \mapsto \frac{\int_t^\infty \exp\left(-1/(x-1/9)\right)\cdot\exp\left(-1/(1/8-x)\right)\cdot \mathbb{1}_{(1/9,1/8)}(x)\mathrm{d}x}{\int_{1/9}^{1/8}\exp\left(-1/(x-1/9)\right)\cdot\exp\left(-1/(1/8-x)\right)\mathrm{d}x}$$

be a well defined infinitely differentiable decreasing function on $\mathbb{R}$ with $\kappa(t) = 1$ for $t \le 1/9$ and $\kappa(t) = 0$ for $t \ge 1/8$. Then define $b := \lceil\beta\rceil - 1$, $\lambda := \beta - b$,

$$u : \mathbb{R}^d \to [0,1], x \mapsto \kappa(\|x\|_2^2),$$

and $\mathrm{c}_2 := \left\|u|_{[-2,2]^d}\right\|_{\mathcal{C}^{b,\lambda}([-2,2]^d)}$. Obviously, u only depends on d, and c_2 only depends on (d,β). Since u is infinitely differentiable and supported in $\mathscr{B}(\mathbf{0}, \sqrt{\frac{1}{8}})$, we have $0 < \mathrm{c}_2 < \infty$. Take $\mathrm{c}_1 := \frac{r}{4\mathrm{c}_2} \wedge \frac{1}{10000}$. Then c_1 only depends on (d,β,r), and $0 < \mathrm{c}_1 < \frac{1}{9999}$. Define

$$f : [0,1]^d \to \mathbb{R}, x \mapsto \sum_{a\in G_{Q,d}} T(a)\cdot\frac{\mathrm{c}_1}{Q^\beta}\cdot u(Q\cdot(x-a)).$$

We then show that these c_1 and f defined above have the desired properties.

For any $\boldsymbol{m} \in (\mathbb{N} \cup \{0\})^d$, we write $u_{\boldsymbol{m}}$ for $\mathrm{D}^{\boldsymbol{m}} u$, i.e., the partial derivative of u with respect to the multi-index $\boldsymbol{m}$. An elementary calculation yields

$$\mathrm{D}^{\boldsymbol{m}} f(x) = \sum_{a \in G_{Q,d}} T(a) \cdot \frac{c_1}{Q^{\beta - \|\boldsymbol{m}\|_1}} \cdot u_{\boldsymbol{m}}(Q \cdot (x-a)),\ \forall\, \boldsymbol{m} \in (\mathbb{N} \cup \{0\})^d,\ x \in [0,1]^d. \tag{C.117}$$

Note that the supports of the functions $T(a) \cdot \frac{c_1}{Q^{\beta - \|\boldsymbol{m}\|_1}} \cdot u_{\boldsymbol{m}}(Q \cdot (x-a))$ $(a \in G_{Q,d})$ in (C.117) are disjoint. Indeed, we have

$$\begin{aligned}
&\left\{ x \in \mathbb{R}^d \,\middle|\, T(a) \cdot \frac{c_1}{Q^{\beta - \|\boldsymbol{m}\|_1}} \cdot u_{\boldsymbol{m}}(Q \cdot (x-a)) \neq 0 \right\} \\
&\subset \mathscr{B}(a, \frac{\sqrt{1/8}}{Q}) \subset \left\{ a + v \,\middle|\, v \in (\frac{-1}{2Q}, \frac{1}{2Q})^d \right\} \\
&\subset [0,1]^d \setminus \left\{ z + v \,\middle|\, v \in [\frac{-1}{2Q}, \frac{1}{2Q}]^d \right\},\ \forall\, \boldsymbol{m} \in (\mathbb{N} \cup \{0\})^d,\ a \in G_{Q,d},\ z \in G_{Q,d} \setminus \{a\},
\end{aligned} \tag{C.118}$$

and sets $\mathscr{B}(a, \frac{\sqrt{1/8}}{Q})$ $(a \in G_{Q,d})$ are disjoint. Therefore,

$$\begin{aligned}
&\|\mathrm{D}^{\boldsymbol{m}} f\|_{[0,1]^d} = \sup_{a \in G_{Q,d}} \sup_{x \in [0,1]^d} \left| T(a) \cdot \frac{c_1}{Q^{\beta - \|\boldsymbol{m}\|_1}} \cdot u_{\boldsymbol{m}}(Q \cdot (x-a)) \right| \\
&= \sup_{a \in G_{Q,d}} \sup_{x \in \mathscr{B}(a, \frac{\sqrt{1/8}}{Q})} \left| T(a) \cdot \frac{c_1}{Q^{\beta - \|\boldsymbol{m}\|_1}} \cdot u_{\boldsymbol{m}}(Q \cdot (x-a)) \right| \\
&= \sup_{a \in G_{Q,d}} \sup_{x \in \mathscr{B}(\mathbf{0}, \sqrt{1/8})} \left| \frac{c_1}{Q^{\beta - \|\boldsymbol{m}\|_1}} \cdot u_{\boldsymbol{m}}(x) \right| \leq \sup_{x \in [-2,2]^d} \left| \frac{c_1}{Q^{\beta - \|\boldsymbol{m}\|_1}} \cdot u_{\boldsymbol{m}}(x) \right| \\
&\leq \sup_{x \in [-2,2]^d} |c_1 \cdot u_{\boldsymbol{m}}(x)| \leq c_1 c_2,\ \forall\, \boldsymbol{m} \in (\mathbb{N} \cup \{0\})^d \text{ with } \|\boldsymbol{m}\|_1 \leq b.
\end{aligned} \tag{C.119}$$

In particular, we have that

$$\|f\|_{[0,1]^d} = \sup_{a \in G_{Q,d}} \sup_{x \in \mathscr{B}(\mathbf{0}, \sqrt{1/8})} \left| \frac{c_1}{Q^{\beta}} \cdot u(x) \right| = \frac{c_1}{Q^{\beta}}. \tag{C.120}$$

Besides, for any $a \in G_{Q,d}$, any $x \in \mathscr{B}(a, \frac{1}{5Q}) \cap [0,1]^d$, and any $z \in G_{Q,d} \setminus \{a\}$, we have

$$\|Q \cdot (x-z)\|_2 \geq Q \|a-z\|_2 - Q \|x-a\|_2 \geq 1 - \frac{1}{5} > \sqrt{1/8} > \sqrt{1/9} > \|Q \cdot (x-a)\|_2,$$

which means that $u(Q \cdot (x-z)) = 0$ and $u(Q \cdot (x-a)) = 1$. Thus

$$\begin{aligned}
&f(x) = T(a) \cdot \frac{c_1}{Q^{\beta}} \cdot u(Q \cdot (x-a)) + \sum_{z \in G_{Q,d} \setminus \{a\}} T(z) \cdot \frac{c_1}{Q^{\beta}} \cdot u(Q \cdot (x-z)) \\
&= T(a) \cdot \frac{c_1}{Q^{\beta}} \cdot 1 + \sum_{z \in G_{Q,d} \setminus \{a\}} T(z) \cdot \frac{c_1}{Q^{\beta}} \cdot 0 \\
&= T(a) \cdot \frac{c_1}{Q^{\beta}},\ \forall\, a \in G_{Q,d},\ x \in \mathscr{B}(a, \frac{1}{5Q}) \cap [0,1]^d.
\end{aligned} \tag{C.121}$$

Now it remains to show that $f \in \mathcal{B}_r^\beta\left([0,1]^d\right)$. Let $\boldsymbol{m} \in (\mathbb{N}\cup\{0\})^d$ be an arbitrary multi-index with $\|\boldsymbol{m}\|_1 = b$, and x, y be arbitrary points in $\bigcup_{a\in G_{Q,d}}\left\{a+v\,\middle|\, v\in(-\frac{1}{2Q},\frac{1}{2Q})^d\right\}$. Then there exist $a_x, a_y \in G_{Q,d}$, such that $x - a_x \in (-\frac{1}{2Q},\frac{1}{2Q})^d$ and $y - a_y \in (-\frac{1}{2Q},\frac{1}{2Q})^d$. If $a_x = a_y$, then it follows from (C.118) that

$$u_{\boldsymbol{m}}(Q\cdot(x-z)) = u_{\boldsymbol{m}}(Q\cdot(y-z)) = 0,\ \forall\, z \in G_{Q,d}\setminus\{a_x\},$$

which, together with the fact that $\{Q\cdot(x-a_x), Q\cdot(y-a_y)\} \subset (-\frac{1}{2},\frac{1}{2})^d$, yields

$$\begin{aligned}
&|\mathrm{D}^{\boldsymbol{m}} f(x) - \mathrm{D}^{\boldsymbol{m}} f(y)|\\
&= \left|T(a_x)\cdot\frac{\mathrm{c}_1}{Q^{\beta-\|\boldsymbol{m}\|_1}}\cdot u_{\boldsymbol{m}}(Q\cdot(x-a_x)) - T(a_y)\cdot\frac{\mathrm{c}_1}{Q^{\beta-\|\boldsymbol{m}\|_1}}\cdot u_{\boldsymbol{m}}(Q\cdot(y-a_y))\right|\\
&= \mathrm{c}_1\cdot\left|\frac{u_{\boldsymbol{m}}(Q\cdot(x-a_x)) - u_{\boldsymbol{m}}(Q\cdot(y-a_y))}{Q^\lambda}\right|\\
&\le \frac{\mathrm{c}_1}{Q^\lambda}\cdot\|Q\cdot(x-a_x) - Q\cdot(y-a_y)\|_2^\lambda\cdot\sup_{z,z'\in(-\frac{1}{2},\frac{1}{2})^d, z\ne z',}\left|\frac{u_{\boldsymbol{m}}(z)-u_{\boldsymbol{m}}(z')}{\|z-z'\|_2^\lambda}\right|\\
&\le \frac{\mathrm{c}_1}{Q^\lambda}\cdot\|Q\cdot(x-a_x) - Q\cdot(y-a_y)\|_2^\lambda\cdot \mathrm{c}_2 = \mathrm{c}_1\mathrm{c}_2\cdot\|x-y\|_2^\lambda.
\end{aligned}$$

If, otherwise, $a_x \ne a_y$, then it is easy to show that

$$\begin{aligned}
&\{t\cdot x + (1-t)\cdot y | t\in[0,1]\} \cap \left\{a_x + v\,\middle|\, v\in[-\frac{1}{2Q},\frac{1}{2Q}]^d\setminus(-\frac{1}{2Q},\frac{1}{2Q})^d\right\} \ne \varnothing,\\
&\{t\cdot x + (1-t)\cdot y | t\in[0,1]\} \cap \left\{a_y + v\,\middle|\, v\in[-\frac{1}{2Q},\frac{1}{2Q}]^d\setminus(-\frac{1}{2Q},\frac{1}{2Q})^d\right\} \ne \varnothing.
\end{aligned}$$

In other words, the line segment joining points x and y intersects boundaries of rectangles $\left\{a_x+v\,\middle|\, v\in(-\frac{1}{2Q},\frac{1}{2Q})^d\right\}$ and $\left\{a_y+v\,\middle|\, v\in(-\frac{1}{2Q},\frac{1}{2Q})^d\right\}$. Take

$$x' \in \{t\cdot x + (1-t)\cdot y | t\in[0,1]\} \cap \left\{a_x + v\,\middle|\, v\in[-\frac{1}{2Q},\frac{1}{2Q}]^d\setminus(-\frac{1}{2Q},\frac{1}{2Q})^d\right\}$$

and

$$y' \in \{t\cdot x + (1-t)\cdot y | t\in[0,1]\} \cap \left\{a_y + v\,\middle|\, v\in[-\frac{1}{2Q},\frac{1}{2Q}]^d\setminus(-\frac{1}{2Q},\frac{1}{2Q})^d\right\}$$

(cf. Figure C.8).

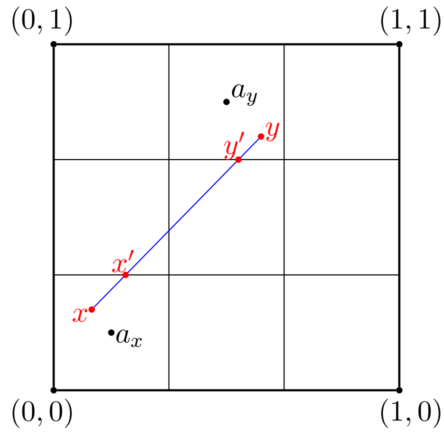

Figure C.8: Illustration of the points x, y, a_x, a_y, x', y' when $Q = 3$ and $d = 2$.

Obviously, we have that

$$\{Q\cdot(x-a_x), Q\cdot(x'-a_x), Q\cdot(y-a_y), Q\cdot(y'-a_y)\}\subset[-\frac{1}{2},\frac{1}{2}]^d.$$

By (C.118), we have that

$$\begin{aligned}&u_{\boldsymbol{m}}(Q\cdot(x-z))\cdot(1-\mathbb{1}_{\{a_x\}}(z)) = u_{\boldsymbol{m}}(Q\cdot(x'-z))\\ &= u_{\boldsymbol{m}}(Q\cdot(y'-z)) = u_{\boldsymbol{m}}(Q\cdot(y-z))\cdot(1-\mathbb{1}_{\{a_y\}}(z)) = 0,\ \forall\, z\in G_{Q,d}.\end{aligned}$$

Consequently,

$$\begin{aligned}&\left|\mathrm{D}^{\boldsymbol{m}}f(x)-\mathrm{D}^{\boldsymbol{m}}f(y)\right|\le\left|\mathrm{D}^{\boldsymbol{m}}f(x)\right|+\left|\mathrm{D}^{\boldsymbol{m}}f(y)\right|\\
&=\left|T(a_x)\cdot\frac{\mathrm{c}_1}{Q^{\beta-\|\boldsymbol{m}\|_1}}\cdot u_{\boldsymbol{m}}(Q\cdot(x-a_x))\right|+\left|T(a_y)\cdot\frac{\mathrm{c}_1}{Q^{\beta-\|\boldsymbol{m}\|_1}}\cdot u_{\boldsymbol{m}}(Q\cdot(y-a_y))\right|\\
&=\frac{\mathrm{c}_1}{Q^\lambda}\cdot\left|u_{\boldsymbol{m}}(Q\cdot(x-a_x))\right|+\frac{\mathrm{c}_1}{Q^\lambda}\cdot\left|u_{\boldsymbol{m}}(Q\cdot(y-a_y))\right|\\
&=\frac{\mathrm{c}_1}{Q^\lambda}\cdot\left|u_{\boldsymbol{m}}(Q\cdot(x-a_x))-u_{\boldsymbol{m}}(Q\cdot(x'-a_x))\right|\\
&\quad+\frac{\mathrm{c}_1}{Q^\lambda}\cdot\left|u_{\boldsymbol{m}}(Q\cdot(y-a_y))-u_{\boldsymbol{m}}(Q\cdot(y'-a_y))\right|\\
&\le\frac{\mathrm{c}_1}{Q^\lambda}\cdot\left\|Q\cdot(x-a_x)-Q\cdot(x'-a_x)\right\|_2^\lambda\cdot\sup_{z,z'\in[-\frac{1}{2},\frac{1}{2}]^d,z\neq z',}\left|\frac{u_{\boldsymbol{m}}(z)-u_{\boldsymbol{m}}(z')}{\|z-z'\|_2^\lambda}\right|\\
&\quad+\frac{\mathrm{c}_1}{Q^\lambda}\cdot\left\|Q\cdot(y-a_y)-Q\cdot(y'-a_y)\right\|_2^\lambda\cdot\sup_{z,z'\in[-\frac{1}{2},\frac{1}{2}]^d,z\neq z',}\left|\frac{u_{\boldsymbol{m}}(z)-u_{\boldsymbol{m}}(z')}{\|z-z'\|_2^\lambda}\right|\\
&\le\frac{\mathrm{c}_1}{Q^\lambda}\cdot\left|\left\|Q\cdot(x-a_x)-Q\cdot(x'-a_x)\right\|_2^\lambda+\left\|Q\cdot(y-a_y)-Q\cdot(y'-a_y)\right\|_2^\lambda\right|\cdot\mathrm{c}_2\\
&=\mathrm{c}_1\mathrm{c}_2\cdot\left|\left\|x-x'\right\|_2^\lambda+\left\|y-y'\right\|_2^\lambda\right|\le2\mathrm{c}_1\mathrm{c}_2\cdot\left\|x-y\right\|_2^\lambda.\end{aligned}$$

Therefore, no matter whether $a_x = a_y$ or not, we always have that

$$|\mathrm{D}^{\boldsymbol{m}} f(x) - \mathrm{D}^{\boldsymbol{m}} f(y)| \leq 2\mathrm{c}_1\mathrm{c}_2 \cdot \|x-y\|_2^\lambda .$$

Since $\boldsymbol{m}$, x, y are arbitrary, we deduce that

$$|\mathrm{D}^{\boldsymbol{m}} f(x) - \mathrm{D}^{\boldsymbol{m}} f(y)| \leq 2\mathrm{c}_1\mathrm{c}_2 \cdot \|x-y\|_2^\lambda$$

for any $\boldsymbol{m} \in (\mathbb{N}\cup\{0\})^d$ with $\|\boldsymbol{m}\|_1 = b$ and any $x, y \in \bigcup_{a\in G_{Q,d}} \left\{a+v \,\middle|\, v \in (-\frac{1}{2Q}, \frac{1}{2Q})^d\right\}$. Note that $\bigcup_{a\in G_{Q,d}} \left\{a+v \,\middle|\, v \in (-\frac{1}{2Q}, \frac{1}{2Q})^d\right\}$ is dense in $[0,1]^d$. Hence, by taking limit, we obtain

$$\begin{aligned}&|\mathrm{D}^{\boldsymbol{m}} f(x) - \mathrm{D}^{\boldsymbol{m}} f(y)| \\ &\leq 2\mathrm{c}_1\mathrm{c}_2 \cdot \|x-y\|_2^\lambda ,\ \forall\, \boldsymbol{m} \in (\mathbb{N}\cup\{0\})^d \text{ with } \|\boldsymbol{m}\|_1 = b,\ \forall\, x,y \in [0,1]^d.\end{aligned} \tag{C.122}$$

Combining (C.119) and (C.122), we conclude that $\|f\|_{\mathcal{C}^{b,\lambda}([0,1]^d)} \leq \mathrm{c}_1\mathrm{c}_2 + 2\mathrm{c}_1\mathrm{c}_2 < r$. Thus $f \in \mathcal{B}_r^\beta\left([0,1]^d\right)$. Then the proof of this lemma is completed. ∎

Let $\mathscr{P}$ and $\mathscr{Q}$ be two arbitrary probability measures which have the same domain. We write $\mathscr{P} << \mathscr{Q}$ if $\mathscr{P}$ is absolutely continuous with respect to $\mathscr{Q}$. The *Kullback-Leibler divergence* (*KL divergence*) from $\mathscr{Q}$ to $\mathscr{P}$ is given by

$$\mathrm{KL}(\mathscr{P}\|\mathscr{Q}) := \begin{cases} \int \log\left(\frac{\mathrm{d}\mathscr{P}}{\mathrm{d}\mathscr{Q}}\right)\mathrm{d}\mathscr{P}, & \text{if } \mathscr{P} << \mathscr{Q}, \\ +\infty, & \text{otherwise,}\end{cases}$$

where $\frac{\mathrm{d}\mathscr{P}}{\mathrm{d}\mathscr{Q}}$ is the *Radon-Nikodym derivative* of $\mathscr{P}$ with respect to $\mathscr{Q}$ (cf. Definition 2.5 of Tsybakov (2009)).

Lemma C.19 *Suppose $\eta_1 : [0,1]^d \to [0,1]$ and $\eta_2 : [0,1]^d \to (0,1)$ are two Borel measurable functions, and $\mathscr{Q}$ is a Borel probability measure on $[0,1]^d$. Then $P_{\eta_1,\mathscr{Q}} << P_{\eta_2,\mathscr{Q}}$, and*

$$\frac{\mathrm{d}P_{\eta_1,\mathscr{Q}}}{\mathrm{d}P_{\eta_2,\mathscr{Q}}}(x,y) = \begin{cases}\frac{\eta_1(x)}{\eta_2(x)}, & \text{if } y=1, \\ \frac{1-\eta_1(x)}{1-\eta_2(x)}, & \text{if } y=-1.\end{cases}$$

Proof Let $f : [0,1]^d \times \{-1,1\} \to [0,\infty),\ (x,y) \mapsto \begin{cases}\frac{\eta_1(x)}{\eta_2(x)}, & \text{if } y=1, \\ \frac{1-\eta_1(x)}{1-\eta_2(x)}, & \text{if } y=-1.\end{cases}$ Then we have that f is well defined and measurable. For any Borel subset S of $[0,1]^d \times \{-1,1\}$, let $S_1 := \left\{x \in [0,1]^d \,\middle|\, (x,1) \in S\right\}$, and $S_2 := \left\{x \in [0,1]^d \,\middle|\, (x,-1) \in S\right\}$. Obvioulsy, $S_1 \times \{1\}$ and $S_2 \times \{-1\}$ are measurable and disjoint. Besides, it is easy to verify that $S = (S_1 \times \{1\}) \cup (S_2 \times \{-1\})$. Therefore,

$$\begin{aligned}&\int_S f(x,y)\mathrm{d}P_{\eta_2,\mathscr{Q}}(x,y) \\ &= \int_{S_1}\int_{\{1\}} f(x,y)\mathrm{d}\mathscr{M}_{\eta_2(x)}(y)\mathrm{d}\mathscr{Q}(x) + \int_{S_2}\int_{\{-1\}} f(x,y)\mathrm{d}\mathscr{M}_{\eta_2(x)}(y)\mathrm{d}\mathscr{Q}(x)\end{aligned}$$

$$
\begin{aligned}
&= \int_{S_1} \eta_2(x) f(x,1) \mathrm{d}\mathscr{Q}(x) + \int_{S_2} (1-\eta_2(x)) f(x,-1) \mathrm{d}\mathscr{Q}(x) \\
&= \int_{S_1} \eta_1(x) \mathrm{d}\mathscr{Q}(x) + \int_{S_2} (1-\eta_1(x)) \mathrm{d}\mathscr{Q}(x) \\
&= \int_{S_1} \int_{\{1\}} \mathrm{d}\mathscr{M}_{\eta_1(x)}(y) \mathrm{d}\mathscr{Q}(x) + \int_{S_2} \int_{\{-1\}} \mathrm{d}\mathscr{M}_{\eta_1(x)}(y) \mathrm{d}\mathscr{Q}(x) \\
&= P_{\eta_1,\mathscr{Q}}(S_1 \times \{1\}) + P_{\eta_1,\mathscr{Q}}(S_2 \times \{-1\}) = P_{\eta_1,\mathscr{Q}}(S).
\end{aligned}
$$

Since S is arbitrary, we deduce that $P_{\eta_1,\mathscr{Q}} << P_{\eta_2,\mathscr{Q}}$, and $\frac{\mathrm{d}P_{\eta_1,\mathscr{Q}}}{\mathrm{d}P_{\eta_2,\mathscr{Q}}} = f$. This completes the proof. ■

Lemma C.20 *Let $\varepsilon \in (0, \frac{1}{5}]$, $\mathscr{Q}$ be a Borel probability on $[0,1]^d$, and $\eta_1 : [0,1]^d \to [\varepsilon, 3\varepsilon]$, $\eta_2 : [0,1]^d \to [\varepsilon, 3\varepsilon]$ be two measurable functions. Then*

$$
\mathrm{KL}(P_{\eta_1,\mathscr{Q}} || P_{\eta_2,\mathscr{Q}}) \leq 9\varepsilon.
$$

Proof By Lemma C.19,

$$
\begin{aligned}
&\mathrm{KL}(P_{\eta_1,\mathscr{Q}} || P_{\eta_2,\mathscr{Q}}) \\
&= \int_{[0,1]^d \times \{-1,1\}} \log\left(\frac{\eta_1(x)}{\eta_2(x)} \cdot \mathbb{1}_{\{1\}}(y) + \frac{1-\eta_1(x)}{1-\eta_2(x)} \cdot \mathbb{1}_{\{-1\}}(y) \right) \mathrm{d}P_{\eta_1,\mathscr{Q}}(x,y) \\
&= \int_{[0,1]^d} \left(\eta_1(x) \log\left(\frac{\eta_1(x)}{\eta_2(x)} \right) + (1-\eta_1(x)) \log\left(\frac{1-\eta_1(x)}{1-\eta_2(x)} \right) \right) \mathrm{d}\mathscr{Q}(x) \\
&\leq \int_{[0,1]^d} \left(3\varepsilon \cdot \left| \log\left(\frac{\eta_1(x)}{\eta_2(x)} \right) \right| + \left| \log\left(\frac{1-\eta_1(x)}{1-\eta_2(x)} \right) \right| \right) \mathrm{d}\mathscr{Q}(x) \\
&\leq \int_{[0,1]^d} \left(3\varepsilon \cdot \log\left(\frac{3\varepsilon}{\varepsilon} \right) + \log\left(\frac{1-\varepsilon}{1-3\varepsilon} \right) \right) \mathrm{d}\mathscr{Q}(x) \\
&= \log\left(1 + \frac{2\varepsilon}{1-3\varepsilon} \right) + 3\varepsilon \cdot \log 3 \leq \frac{2\varepsilon}{1-3\varepsilon} + 4\varepsilon \leq 9\varepsilon.
\end{aligned}
$$

■

Lemma C.21 *Let $m \in \mathbb{N} \cap (1, \infty)$, Ω be a set with $\#(\Omega) = m$, and $\{0,1\}^\Omega$ be the set of all functions mapping from Ω to $\{0,1\}$. Then there exists a subset E of $\{0,1\}^\Omega$, such that $\#(E) \geq 1 + 2^{m/8}$, and*

$$
\#\left(\{ x \in \Omega |\, f(x) \neq g(x) \} \right) \geq \frac{m}{8}, \; \forall\, f \in E, \; \forall\, g \in E \setminus \{f\}.
$$

Proof If $m \leq 8$, then $E = \{0,1\}^\Omega$ have the desired properties. The proof for the case $m > 8$ can be found in Lemma 2.9 of Tsybakov (2009). ■

Lemma C.22 *Let ϕ be the logistic loss,*

$$\begin{aligned}&\mathcal{J}:(0,1)^2\to\mathbb{R}\\&\quad(x,y)\mapsto(x+y)\log\frac{2}{x+y}+(2-x-y)\log\frac{2}{2-x-y}\\&\qquad\qquad-\left(x\log\frac{1}{x}+(1-x)\log\frac{1}{1-x}+y\log\frac{1}{y}+(1-y)\log\frac{1}{1-y}\right),\end{aligned}\tag{C.123}$$

$\mathscr{Q}$ be a Borel probability measure on $[0,1]^d$, and $\eta_1:[0,1]^d\to(0,1)$, $\eta_2:[0,1]^d\to(0,1)$ be two measurable functions. Then there hold

$$\mathcal{J}(x,y)=\mathcal{J}(y,x)\geq 0,\ \forall\, x\in(0,1),\ y\in(0,1),\tag{C.124}$$

$$\frac{\varepsilon}{4}<\mathcal{J}(\varepsilon,3\varepsilon)=\mathcal{J}(3\varepsilon,\varepsilon)<\varepsilon,\ \forall\,\varepsilon\in(0,\frac{1}{6}],\tag{C.125}$$

and

$$\int_{[0,1]^d}\mathcal{J}(\eta_1(x),\eta_2(x))\mathrm{d}\mathscr{Q}(x)\leq\inf_{f\in\mathcal{F}_d}\left|\mathcal{E}^{\phi}_{P_{\eta_1,\mathscr{Q}}}(f)+\mathcal{E}^{\phi}_{P_{\eta_2,\mathscr{Q}}}(f)\right|.\tag{C.126}$$

Proof Let $g:(0,1)\to(0,\infty),x\mapsto x\log\frac{1}{x}+(1-x)\log\frac{1}{1-x}$. Then it is easy to verify that g is concave (i.e., $-g$ is convex), and

$$\mathcal{J}(x,y)=2g(\frac{x+y}{2})-g(x)-g(y),\ \forall\, x\in(0,1),\ y\in(0,1).$$

This yields (C.124).

An elementary calculation gives

$$\begin{aligned}&\mathcal{J}(\varepsilon,3\varepsilon)=\mathcal{J}(3\varepsilon,\varepsilon)\\&=\varepsilon\log\frac{27}{16}-\log\left(\frac{(1-2\varepsilon)^2}{(1-\varepsilon)(1-3\varepsilon)}\right)+4\varepsilon\log(1-2\varepsilon)-\varepsilon\log(1-\varepsilon)-3\varepsilon\log(1-3\varepsilon)\\&\xlongequal{\text{Taylor expansion}}\varepsilon\log\frac{27}{16}+\sum_{k=2}^{\infty}\frac{3^k+1-2\cdot 2^k}{k\cdot(k-1)}\cdot\varepsilon^k,\ \forall\,\varepsilon\in(0,1/3).\end{aligned}$$

Therefore,

$$\begin{aligned}&\frac{\varepsilon}{4}<\varepsilon\log\frac{27}{16}\leq\varepsilon\log\frac{27}{16}+\sum_{k=2}^{\infty}\frac{1+\left(\left(\frac{3}{2}\right)^k-2\right)\cdot 2^k}{k\cdot(k-1)}\cdot\varepsilon^k=\varepsilon\log\frac{27}{16}+\sum_{k=2}^{\infty}\frac{3^k+1-2\cdot 2^k}{k\cdot(k-1)}\cdot\varepsilon^k\\&=\mathcal{J}(\varepsilon,3\varepsilon)=\mathcal{J}(3\varepsilon,\varepsilon)=\varepsilon\log\frac{27}{16}+\sum_{k=2}^{\infty}\frac{3^k+1-2\cdot 2^k}{k\cdot(k-1)}\cdot\varepsilon^k\leq\varepsilon\log\frac{27}{16}+\sum_{k=2}^{\infty}\frac{3^k-7}{k\cdot(k-1)}\cdot\varepsilon^k\\&=\varepsilon\log\frac{27}{16}+\varepsilon^2+\varepsilon\cdot\sum_{k=3}^{\infty}\frac{3^k-7}{k\cdot(k-1)}\cdot\varepsilon^{k-1}\leq\varepsilon\log\frac{27}{16}+\varepsilon/6+\varepsilon\cdot\sum_{k=3}^{\infty}\frac{3^k}{3\cdot(3-1)}\cdot\left(\frac{1}{6}\right)^{k-1}\\&=\left(\frac{1}{6}+\frac{1}{4}+\log\frac{27}{16}\right)\cdot\varepsilon<\varepsilon,\ \forall\,\varepsilon\in(0,1/6],\end{aligned}$$

which proves (C.125).

Define $f_1 : [0,1]^d \to \mathbb{R}, x \mapsto \log \frac{\eta_1(x)}{1-\eta_1(x)}$ and $f_2 : [0,1]^d \to \mathbb{R}, x \mapsto \log \frac{\eta_2(x)}{1-\eta_2(x)}$. Then it is easy to verify that

$$\mathcal{R}^{\phi}_{P_{\eta_i},\mathscr{Q}}(f_i) = \int_{[0,1]^d} g(\eta_i(x)) \mathrm{d}\mathscr{Q}(x) \in (0,\infty),\ \forall\, i \in \{1,2\},$$

and

$$\inf\left\{a\phi(t) + (1-a)\phi(-t) \middle| t \in \mathbb{R}\right\} = g(a),\ \forall\, a \in (0,1).$$

Consequently, for any measurable function $f : [0,1]^d \to \mathbb{R}$, there holds

$$\begin{aligned}
&\mathcal{E}^{\phi}_{P_{\eta_1},\mathscr{Q}}(f) + \mathcal{E}^{\phi}_{P_{\eta_2},\mathscr{Q}}(f) \geq \mathcal{R}^{\phi}_{P_{\eta_1},\mathscr{Q}}(f) - \mathcal{R}^{\phi}_{P_{\eta_1},\mathscr{Q}}(f_1) + \mathcal{R}^{\phi}_{P_{\eta_2},\mathscr{Q}}(f) - \mathcal{R}^{\phi}_{P_{\eta_2},\mathscr{Q}}(f_2) \\
&= \int_{[0,1]^d} \left((\eta_1(x)+\eta_2(x))\phi(f(x)) + (2-\eta_1(x)-\eta_2(x))\phi(-f(x)\right) \mathrm{d}\mathscr{Q}(x) \\
&\qquad\qquad - \mathcal{R}^{\phi}_{P_{\eta_1},\mathscr{Q}}(f_1) - \mathcal{R}^{\phi}_{P_{\eta_2},\mathscr{Q}}(f_2) \\
&\geq \int_{[0,1]^d} 2\cdot \inf\left\{ \frac{\eta_1(x)+\eta_2(x)}{2}\phi(t) + (1-\frac{\eta_1(x)+\eta_2(x)}{2})\phi(-t) \middle| t \in \mathbb{R}\right\} \mathrm{d}\mathscr{Q}(x) \\
&\qquad\qquad - \mathcal{R}^{\phi}_{P_{\eta_1},\mathscr{Q}}(f_1) - \mathcal{R}^{\phi}_{P_{\eta_2},\mathscr{Q}}(f_2) \\
&= \int_{[0,1]^d} 2g(\frac{\eta_1(x)+\eta_2(x)}{2}) \mathrm{d}\mathscr{Q}(x) - \mathcal{R}^{\phi}_{P_{\eta_1},\mathscr{Q}}(f_1) - \mathcal{R}^{\phi}_{P_{\eta_2},\mathscr{Q}}(f_2) \\
&= \int_{[0,1]^d} \left(2g(\frac{\eta_1(x)+\eta_2(x)}{2}) - g(\eta_1(x)) - g(\eta_2(x))\right) \mathrm{d}\mathscr{Q}(x) \\
&= \int_{[0,1]^d} \mathcal{J}(\eta_1(x),\eta_2(x)) \mathrm{d}\mathscr{Q}(x).
\end{aligned}$$

This proves (C.126). ■

Proof [Proof of Theorem 2.6 and Corollary 2.1] We first prove Theorem 2.6. Let n be an arbitrary integer greater than $\left|\frac{7}{1-A}\right|^{\frac{d_*+\beta\cdot(1\wedge\beta)^q}{\beta\cdot(1\wedge\beta)^q}}$. Take $b := \lceil\beta\rceil - 1$, $\lambda := \beta + 1 - \lceil\beta\rceil$, $Q := \left\lfloor n^{\frac{1}{d_*+\beta\cdot(1\wedge\beta)^q}}\right\rfloor + 1$, $\mathtt{M} := \left\lceil 2^{Q^{d_*}/8}\right\rceil$,

$$G_{Q,d_*} := \left\{ (\frac{k_1}{2Q},\ldots,\frac{k_{d_*}}{2Q})^\top \middle| k_1,\ldots,k_{d_*} \text{ are odd integers}\right\} \cap [0,1]^{d_*},$$

and $\mathcal{J}$ to be the function defined in (C.123). Note that $\#\left(G_{Q,d_*}\right) = Q^{d_*}$. Thus it follows from Lemma C.21 that there exist functions $T_j : G_{Q,d_*} \to \{-1,1\}$, $j = 0,1,2,\ldots,\mathtt{M}$, such that

$$\#\left(\{a \in G_{Q,d_*} | T_i(a) \neq T_j(a)\}\right) \geq \frac{Q^{d_*}}{8},\ \forall\, 0 \leq i < j \leq \mathtt{M}. \tag{C.127}$$

According to Lemma C.18, for each $j \in \{0,1,\ldots,\mathtt{M}\}$, there exists an $f_j \in \mathcal{B}^{\beta}_{\frac{r\wedge 1}{777}}\left([0,1]^{d_*}\right)$, such that

$$\frac{\mathrm{c}_1}{Q^\beta} = \|f_j\|_{[0,1]^{d_*}} \leq \|f_j\|_{\mathcal{C}^{b,\lambda}([0,1]^{d_*})} \leq \frac{1\wedge r}{777}, \tag{C.128}$$

and

$$f_j(x) = \frac{c_1}{Q^\beta} \cdot T_j(a),\ \forall\ a \in G_{Q,d_*},\ x \in \mathscr{B}(a, \frac{1}{5Q}) \cap [0,1]^{d_*}, \tag{C.129}$$

where $c_1 \in (0, \frac{1}{9999})$ only depends on (d_*, β, r). Define

$$g_j : [0,1]^{d_*} \to \mathbb{R}, x \mapsto \frac{c_1}{Q^\beta} + f_j(x).$$

It follows from (C.128) that

$$\mathbf{ran}(g_j) \subset \left[0, \frac{2c_1}{Q^\beta}\right] \subset \left[0, 2 \cdot \frac{1 \wedge r}{777}\right] \subset [0,1] \tag{C.130}$$

and

$$\frac{c_1}{Q^\beta} + \|g_j\|_{\mathcal{C}^{b,\lambda}([0,1]^{d_*})} \le \frac{2c_1}{Q^\beta} + \|f_j\|_{\mathcal{C}^{b,\lambda}([0,1]^{d_*})} \le 2 \cdot \frac{1 \wedge r}{777} + \frac{1 \wedge r}{777} < r, \tag{C.131}$$

meaning that

$$g_j \in \mathcal{B}_r^\beta\left([0,1]^{d_*}\right) \text{ and } g_j + \frac{c_1}{Q^\beta} \in \mathcal{B}_r^\beta\left([0,1]^{d_*}\right). \tag{C.132}$$

We then define

$$h_{0,j} : [0,1]^d \to [0,1],\ (x_1, \ldots, x_d)^\top \mapsto g_j(x_1, \ldots, x_{d_*})$$

if $q = 0$, and define

$$h_{0,j} : [0,1]^d \to [0,1]^K,\ (x_1, \ldots, x_d)^\top \mapsto (g_j(x_1, \ldots, x_{d_*}), 0, 0, \ldots, 0)^\top$$

if $q > 0$. Note that $h_{0,j}$ is well defined because $d_* \le d$ and $\mathbf{ran}(g_j) \subset [0,1]$. Take

$$\varepsilon = \frac{1}{2} \cdot \left|\frac{1 \wedge r}{777}\right|^{\sum_{k=0}^{q-1}(1\wedge\beta)^k} \cdot \left|\frac{2c_1}{Q^\beta}\right|^{(1\wedge\beta)^q}.$$

From (C.128) we see that

$$0 < \varepsilon \le \frac{1 \wedge r}{777}. \tag{C.133}$$

For all real number t, define the function

$$u_t : [0,1]^{d_*} \to \mathbb{R},\ (x_1, \ldots, x_{d_*})^\top \mapsto t + \frac{1 \wedge r}{777} \cdot |x_1|^{(1\wedge\beta)}.$$

Then it follows from (C.133) and the elementary inequality

$$||z_1|^w - |z_2|^w| \le |z_1 - z_2|^w,\ \forall\ z_1 \in \mathbb{R}, z_2 \in \mathbb{R}, w \in (0,1]$$

that

$$\begin{aligned} &\max\left\{\|u_\varepsilon\|_{[0,1]^{d_*}}, \|u_0\|_{[0,1]^{d_*}}\right\} \le \max\left\{\|u_\varepsilon\|_{\mathcal{C}^{b,\lambda}([0,1]^{d_*})}, \|u_0\|_{\mathcal{C}^{b,\lambda}([0,1]^{d_*})}\right\} \\ &\le \|u_0\|_{\mathcal{C}^{b,\lambda}([0,1]^{d_*})} + \varepsilon \le \frac{1 \wedge r}{777} \cdot 2 + \varepsilon \le \frac{1 \wedge r}{777} \cdot 2 + \frac{1 \wedge r}{777} < r \wedge 1, \end{aligned} \tag{C.134}$$

which means that

$$\mathbf{ran}(u_0)\cup\mathbf{ran}(u_\varepsilon)\subset[0,1], \tag{C.135}$$

and

$$\{u_0,u_\varepsilon\}\subset\mathcal{B}_r^\beta\left([0,1]^{d_*}\right). \tag{C.136}$$

Next, for each $i\in\mathbb{N}$, define

$$\begin{aligned}h_i:[0,1]^K&\to\mathbb{R},\\ (x_1,\ldots,x_K)^\top&\mapsto u_0(x_1,\ldots,x_{d_*})\end{aligned}$$

if $i=q>0$, and define

$$h_i:[0,1]^K\to\mathbb{R}^K,\ (x_1,\ldots,x_K)^\top\mapsto(u_0(x_1,\ldots,x_{d_*}),0,0,\ldots,0)^\top$$

otherwise. It follows from (C.135) that $\mathbf{ran}(h_i)\subset[0,1]$ if $i=q>0$, and $\mathbf{ran}(h_i)\subset[0,1]^K$ otherwise. Thus, for each $j\in\{0,1,\ldots,\mathtt{M}\}$, we can well define

$$\eta_j:[0,1]^d\to\mathbb{R},\ x\mapsto\varepsilon+h_q\circ h_{q-1}\circ\cdots\circ h_3\circ h_2\circ h_1\circ h_{0,j}(x).$$

We then deduce from (C.132) and (C.136) that

$$\eta_j\in\mathcal{G}_d^{\mathbf{CH}}(q,K,d_*,\beta,r),\ \forall\, j\in\{0,1,\ldots,\mathtt{M}\}. \tag{C.137}$$

Moreover, an elementary calculation gives

$$\begin{aligned}&\left|\frac{1\wedge r}{777}\right|^{\sum_{k=0}^{q-1}(1\wedge\beta)^k}\cdot|g_j(x_1,\ldots,x_{d_*})|^{(1\wedge\beta)^q}+\varepsilon\\&=\eta_j(x_1,\ldots,x_d),\ \forall\,(x_1,\ldots,x_d)\in[0,1]^d,\ \forall\, j\in\{0,1,\ldots,\mathtt{M}\},\end{aligned} \tag{C.138}$$

which, together with (C.130), yields

$$\begin{aligned}&0<\varepsilon\le\eta_j(x_1,\ldots,x_d)\le\left|\frac{1\wedge r}{777}\right|^{\sum_{k=0}^{q-1}(1\wedge\beta)^k}\cdot\left|\frac{2\mathrm{c}_1}{Q^\beta}\right|^{(1\wedge\beta)^q}+\varepsilon=2\varepsilon+\varepsilon\\&=3\varepsilon\le\left|\frac{3\mathrm{c}_1}{Q^\beta}\right|^{(1\wedge\beta)^q}<\frac{1}{Q^{\beta\cdot(1\wedge\beta)^q}}\le\frac{1}{n^{\frac{\beta\cdot(1\wedge\beta)^q}{d_*+\beta\cdot(1\wedge\beta)^q}}}\le\frac{1-A}{7}<\frac{1-A}{2}\\&<1,\ \forall\,(x_1,\ldots,x_d)\in[0,1]^d,\ \forall\, j\in\{0,1,\ldots,\mathtt{M}\}.\end{aligned}$$

Consequently,

$$\mathbf{ran}(\eta_j)\subset[\varepsilon,3\varepsilon]\subset(0,1),\ \forall\, j\in\{0,1,\ldots,\mathtt{M}\}, \tag{C.139}$$

and

$$\left\{x\in[0,1]^d\middle|\,|2\eta_j(x)-1|\le A\right\}=\varnothing,\ \forall\, j\in\{0,1,\ldots,\mathtt{M}\}. \tag{C.140}$$

Combining (C.137), (C.139), and (C.140), we obtain

$$P_j:=P_{\eta_j}\in\mathcal{H}^{d,\beta,r}_{5,A,q,K,d_*},\ \forall\, j\in\{0,1,2,\ldots,\mathtt{M}\}. \tag{C.141}$$

By (C.129) and (C.138), for any $0 \le i < j \le \mathtt{M}$, any $a \in G_{Q,d_*}$ with $T_i(a) \ne T_j(a)$, and any $x \in [0,1]^d$ with $(x)_{\{1,2,\dots,d_*\}} \in \mathscr{B}(a, \frac{1}{5Q})$, there holds

$$
\begin{aligned}
&\mathcal{J}(\eta_i(x), \eta_j(x)) \\
&= \mathcal{J}\left(\left|\frac{1\wedge r}{777}\right|^{\sum_{k=0}^{q-1}(1\wedge\beta)^k} \cdot \left|\frac{\mathrm{c}_1}{Q^\beta} + T_i(a)\cdot\frac{\mathrm{c}_1}{Q^\beta}\right|^{(1\wedge\beta)^q} + \varepsilon, \right. \\
&\qquad\qquad \left. \left|\frac{1\wedge r}{777}\right|^{\sum_{k=0}^{q-1}(1\wedge\beta)^k} \cdot \left|\frac{\mathrm{c}_1}{Q^\beta} + T_j(a)\cdot\frac{\mathrm{c}_1}{Q^\beta}\right|^{(1\wedge\beta)^q} + \varepsilon \right) \\
&= \mathcal{J}\left(\left|\frac{1\wedge r}{777}\right|^{\sum_{k=0}^{q-1}(1\wedge\beta)^k} \cdot \left|\frac{2\mathrm{c}_1}{Q^\beta}\right|^{(1\wedge\beta)^q} + \varepsilon, \left|\frac{1\wedge r}{777}\right|^{\sum_{k=0}^{q-1}(1\wedge\beta)^k} \cdot |0|^{(1\wedge\beta)^q} + \varepsilon \right) \\
&= \mathcal{J}(2\varepsilon + \varepsilon, \varepsilon) = \mathcal{J}(\varepsilon, 3\varepsilon).
\end{aligned}
$$

Thus it follows from Lemma C.22 and (C.127) that

$$
\begin{aligned}
&\inf_{f\in\mathcal{F}_d} \left(\mathcal{E}^\phi_{P_j}(f) + \mathcal{E}^\phi_{P_i}(f)\right) \ge \int_{[0,1]^d} \mathcal{J}(\eta_i(x), \eta_j(x))\mathrm{d}x \\
&\ge \sum_{a\in G_{Q,d_*}:\, T_j(a)\ne T_i(a)} \int_{[0,1]^d} \mathcal{J}(\eta_i(x), \eta_j(x)) \cdot \mathbb{1}_{\mathscr{B}(a,\frac{1}{5Q})}\big((x)_{\{1,\dots,d_*\}}\big)\mathrm{d}x \\
&= \sum_{a\in G_{Q,d_*}:\, T_j(a)\ne T_i(a)} \int_{[0,1]^d} \mathcal{J}(\varepsilon, 3\varepsilon) \cdot \mathbb{1}_{\mathscr{B}(a,\frac{1}{5Q})}\big((x)_{\{1,\dots,d_*\}}\big)\mathrm{d}x \\
&\ge \sum_{a\in G_{Q,d_*}:\, T_j(a)\ne T_i(a)} \int_{[0,1]^d} \frac{\varepsilon}{4} \cdot \mathbb{1}_{\mathscr{B}(a,\frac{1}{5Q})}\big((x)_{\{1,\dots,d_*\}}\big)\mathrm{d}x \qquad\qquad (C.142) \\
&= \frac{\#\left(\left\{a\in G_{Q,d_*} \middle| T_j(a)\ne T_i(a)\right\}\right)}{Q^{d_*}} \cdot \int_{\mathscr{B}(\mathbf{0},\frac{1}{5})} \frac{\varepsilon}{4}\mathrm{d}x_1\mathrm{d}x_2\cdots\mathrm{d}x_{d_*} \\
&\ge \frac{1}{8}\cdot \int_{\mathscr{B}(\mathbf{0},\frac{1}{5})} \frac{\varepsilon}{4}\mathrm{d}x_1\mathrm{d}x_2\cdots\mathrm{d}x_{d_*} \ge \frac{1}{8}\cdot\int_{[-\frac{1}{\sqrt{25d_*}},\frac{1}{\sqrt{25d_*}}]^{d_*}} \frac{\varepsilon}{4}\mathrm{d}x_1\mathrm{d}x_2\cdots\mathrm{d}_{x_{d_*}} \\
&\ge \left|\frac{2}{\sqrt{25d_*}}\right|^{d_*} \cdot \frac{\varepsilon}{32} =: s,\ \forall\, 0\le i<j\le \mathtt{M}.
\end{aligned}
$$

Let $\hat{f}_n$ be an arbitrary $\mathcal{F}_d$-valued statistic on $([0,1]^d \times \{-1,1\})^n$ from the sample $\{(X_i,Y_i)\}_{i=1}^n$, and let $\mathcal{T} : ([0,1]^d\times\{-1,1\})^n \to \mathcal{F}_d$ be the map associated with $\hat{f}_n$, i.e., $\hat{f}_n = \mathcal{T}(X_1,Y_1,\dots,X_n,Y_n)$. Take

$$
\mathcal{T}_0 : \mathcal{F}_d \to \{0,1,\dots,\mathtt{M}\}, f \mapsto \inf \mathop{\arg\min}_{j\in\{0,1,\dots,\mathtt{M}\}} \mathcal{E}^\phi_{P_j}(f),
$$

i.e., $\mathcal{T}_0(f)$ is the smallest integer $j \in \{0,\dots,\mathtt{M}\}$ such that $\mathcal{E}^\phi_{P_j}(f) \le \mathcal{E}^\phi_{P_i}(f)$ for any $i \in \{0,\dots,\mathtt{M}\}$. Define $g_* = \mathcal{T}_0\circ\mathcal{T}$. Note that, for any $j\in\{0,1,\dots,\mathtt{M}\}$ and any $f\in\mathcal{F}_d$ there holds

$$
\mathcal{T}_0(f) \ne j \overset{(C.142)}{\Rightarrow} \mathcal{E}^\phi_{P_{\mathcal{T}_0(f)}}(f) + \mathcal{E}^\phi_{P_j}(f) \ge s \Rightarrow \mathcal{E}^\phi_{P_j}(f) + \mathcal{E}^\phi_{P_j}(f) \ge s \Rightarrow \mathcal{E}^\phi_{P_j}(f) \ge \frac{s}{2},
$$

which, together with the fact that the range of $\mathcal{T}$ is contained in $\mathcal{F}_d$, yields

$$\begin{aligned}&\mathbb{1}_{\mathbb{R}\setminus\{j\}}(g_*(z)) = \mathbb{1}_{\mathbb{R}\setminus\{j\}}(\mathcal{T}_0(\mathcal{T}((z))))\\ &\leq \mathbb{1}_{[\frac{s}{2},\infty]}(\mathcal{E}^{\phi}_{P_j}(\mathcal{T}(z))),\ \forall\, z\in([0,1]^d\times\{-1,1\})^n,\ \forall\, j\in\{0,1,\ldots,\mathtt{M}\}\,.\end{aligned} \tag{C.143}$$

Consequently,

$$\begin{aligned}&\sup_{P\in\mathcal{H}^{d,\beta,r}_{5,A,q,K,d_*}} \boldsymbol{E}_{P^{\otimes n}}\left[\mathcal{E}^{\phi}_{P}(\hat{f}_n)\right] \geq \sup_{j\in\{0,1,\ldots,\mathtt{M}\}} \boldsymbol{E}_{P_j^{\otimes n}}\left[\mathcal{E}^{\phi}_{P_j}(\hat{f}_n)\right]\\ &= \sup_{j\in\{0,1,\ldots,\mathtt{M}\}} \int \mathcal{E}^{\phi}_{P_j}(\mathcal{T}(z))\mathrm{d}P_j^{\otimes n}(z) \geq \sup_{j\in\{0,1,\ldots,\mathtt{M}\}} \int \frac{\mathbb{1}_{[\frac{s}{2},\infty]}(\mathcal{E}^{\phi}_{P_j}(\mathcal{T}(z)))}{2/s}\mathrm{d}P_j^{\otimes n}(z)\\ &\geq \sup_{j\in\{0,1,\ldots,\mathtt{M}\}} \int \frac{\mathbb{1}_{\mathbb{R}\setminus\{j\}}(g_*(z))}{2/s}\mathrm{d}P_j^{\otimes n}(z) = \sup_{j\in\{0,1,\ldots,\mathtt{M}\}} \frac{P_j^{\otimes n}\left(g_*\neq j\right)}{2/s}\\ &\geq \frac{s}{2}\cdot\inf\left\{\sup_{j\in\{0,1,\ldots,\mathtt{M}\}} P_j^{\otimes n}\left(g\neq j\right)\middle|\begin{array}{l} g \text{ is a measurable function from}\\ ([0,1]^d\times\{-1,1\})^n \text{ to } \{0,1,\ldots,\mathtt{M}\}\end{array}\right\},\end{aligned} \tag{C.144}$$

where the first inequality follows from (C.141) and the third inequality follows from (C.143).

We then use Proposition 2.3 of Tsybakov (2009) to bound the right hand side of (C.144). By Lemma C.20, we have that

$$\frac{1}{\mathtt{M}}\cdot\sum_{j=1}^{\mathtt{M}}\mathrm{KL}(P_j^{\otimes n}||P_0^{\otimes n}) = \frac{n}{\mathtt{M}}\cdot\sum_{j=1}^{\mathtt{M}}\mathrm{KL}(P_j||P_0) \leq \frac{n}{\mathtt{M}}\cdot\sum_{j=1}^{\mathtt{M}} 9\varepsilon = 9n\varepsilon,$$

which, together with Proposition 2.3 of Tsybakov (2009), yields

$$\begin{aligned}&\inf\left\{\sup_{j\in\{0,1,\ldots,\mathtt{M}\}} P_j^{\otimes n}\left(g\neq j\right)\middle|\begin{array}{l} g \text{ is a measurable function from}\\ ([0,1]^d\times\{-1,1\})^n \text{ to } \{0,1,\ldots,\mathtt{M}\}\end{array}\right\}\\ &\geq \sup_{\tau\in(0,1)}\left(\frac{\tau\mathtt{M}}{1+\tau\mathtt{M}}\cdot\left(1+\frac{9n\varepsilon+\sqrt{\frac{9n\varepsilon}{2}}}{\log\tau}\right)\right) \geq \frac{\sqrt{\mathtt{M}}}{1+\sqrt{\mathtt{M}}}\cdot\left(1+\frac{9n\varepsilon+\sqrt{\frac{9n\varepsilon}{2}}}{\log\frac{1}{\sqrt{\mathtt{M}}}}\right)\\ &\geq \frac{\sqrt{\mathtt{M}}}{1+\sqrt{\mathtt{M}}}\cdot\left(1-\left|\frac{9n\varepsilon+\sqrt{\frac{9n\varepsilon}{2}}}{\log\frac{1}{\sqrt{\mathtt{M}}}}\right|\right) \geq \frac{\sqrt{\mathtt{M}}}{1+\sqrt{\mathtt{M}}}\cdot\left(1-\left|\frac{9n\varepsilon+\frac{1}{10}+12n\varepsilon}{\log\sqrt{\mathtt{M}}}\right|\right)\\ &\geq \frac{\sqrt{\mathtt{M}}}{1+\sqrt{\mathtt{M}}}\cdot\left(1-\left|\frac{21n\varepsilon}{\frac{1}{2}\log\left(2^{Q^{d_*}/8}\right)}\right|-\frac{1/10}{\log\sqrt{2}}\right)\\ &= \frac{\sqrt{\mathtt{M}}}{1+\sqrt{\mathtt{M}}}\cdot\left(1-\left|\frac{336n}{Q^{d_*}\log 2}\cdot\frac{1}{2}\cdot\left|\frac{1\wedge r}{777}\right|^{\sum_{k=0}^{q-1}(1\wedge\beta)^k}\cdot\left|\frac{2\mathrm{c}_1}{Q^{\beta}}\right|^{(1\wedge\beta)^q}\right|-\frac{1/10}{\log\sqrt{2}}\right)\\ &\geq \frac{\sqrt{\mathtt{M}}}{1+\sqrt{\mathtt{M}}}\cdot\left(1-\left|\frac{336n}{Q^{d_*}\log 2}\cdot\frac{1}{2}\cdot\frac{1}{777}\cdot\left|\frac{1}{Q^{\beta}}\right|^{(1\wedge\beta)^q}\right|-\frac{1/10}{\log\sqrt{2}}\right)\\ &\geq \frac{\sqrt{\mathtt{M}}}{1+\sqrt{\mathtt{M}}}\cdot\left(1-\left|\frac{336}{\log 2}\cdot\frac{1}{2}\cdot\frac{1}{777}\right|-\frac{1/10}{\log\sqrt{2}}\right) \geq \frac{\sqrt{\mathtt{M}}}{1+\sqrt{\mathtt{M}}}\cdot\frac{1}{3}\geq\frac{1}{6}.\end{aligned}$$

Combining this with (C.144), we obtain that

$$\begin{aligned}
&\sup_{P\in\mathcal{H}^{d,\beta,r}_{5,A,q,K,d_*}} \boldsymbol{E}_{P^{\otimes n}}\left[\mathcal{E}^{\phi}_P(\hat{f}_n)\right] \geq \frac{s}{2}\cdot\frac{1}{6} \\
&= \left|\frac{2}{\sqrt{25d_*}}\right|^{d_*}\cdot\frac{|2\mathrm{c}_1|^{(1\wedge\beta)^q}}{768}\cdot\left|\frac{1\wedge r}{777}\right|^{\sum_{k=0}^{q-1}(1\wedge\beta)^k}\cdot\left|\frac{1}{Q^\beta}\right|^{(1\wedge\beta)^q} \\
&\geq \left|\frac{2}{\sqrt{25d_*}}\right|^{d_*}\cdot\frac{|2\mathrm{c}_1|^{(1\wedge\beta)^q}}{768}\cdot\left|\frac{1\wedge r}{777}\right|^{\sum_{k=0}^{q-1}(1\wedge\beta)^k}\cdot\frac{1}{2^{\beta\cdot(1\wedge\beta)^q}}\cdot\frac{1}{n^{\frac{\beta\cdot(1\wedge\beta)^q}{d_*+\beta\cdot(1\wedge\beta)^q}}}.
\end{aligned}$$

Since $\hat{f}_n$ is arbitrary, we deduce that

$$\inf_{\hat{f}_n}\sup_{P\in\mathcal{H}^{d,\beta,r}_{5,A,q,K,d_*}} \boldsymbol{E}_{P^{\otimes n}}\left[\mathcal{E}^{\phi}_P(\hat{f}_n)\right] \geq \mathrm{c}_0 n^{-\frac{\beta\cdot(1\wedge\beta)^q}{d_*+\beta\cdot(1\wedge\beta)^q}}$$

with $\mathrm{c}_0 := \left|\frac{2}{\sqrt{25d_*}}\right|^{d_*}\cdot\frac{|2\mathrm{c}_1|^{(1\wedge\beta)^q}}{768}\cdot\left|\frac{1\wedge r}{777}\right|^{\sum_{k=0}^{q-1}(1\wedge\beta)^k}\cdot\frac{1}{2^{\beta\cdot(1\wedge\beta)^q}}$ only depending on (d_*,β,r,q). Thus we complete the proof of Theorem 2.6.

Now it remains to prove Corollary 2.1. Indeed, it follows from (2.33) that

$$\mathcal{H}^{d,\beta,r}_{3,A} = \mathcal{H}^{d,\beta,r}_{5,A,0,1,d}.$$

Taking $q=0$, $K=1$ and $d_*=d$ in Theorem 2.6, we obtain that there exists an constant $\mathrm{c}_0\in(0,\infty)$ only depending on (d,β,r), such that

$$\begin{aligned}
&\inf_{\hat{f}_n}\sup_{P\in\mathcal{H}^{d,\beta,r}_{3,A}} \boldsymbol{E}_{P^{\otimes n}}\left[\mathcal{E}^{\phi}_P(\hat{f}_n)\right] = \inf_{\hat{f}_n}\sup_{P\in\mathcal{H}^{d,\beta,r}_{5,A,0,1,d}} \boldsymbol{E}_{P^{\otimes n}}\left[\mathcal{E}^{\phi}_P(\hat{f}_n)\right] \geq \mathrm{c}_0 n^{-\frac{\beta\cdot(1\wedge\beta)^0}{d+\beta\cdot(1\wedge\beta)^0}} \\
&= \mathrm{c}_0 n^{-\frac{\beta}{d+\beta}} \text{ provided that } n > \left|\frac{7}{1-A}\right|^{\frac{d+\beta\cdot(1\wedge\beta)^0}{\beta\cdot(1\wedge\beta)^0}} = \left|\frac{7}{1-A}\right|^{\frac{d+\beta}{\beta}}.
\end{aligned}$$

This proves Corollary 2.1. ■

C.7 Proof of (3.7)

Appendix C.7 is devoted to the proof of the bound (3.7).

Proof [Proof of (3.7)]Fix $\nu\in[0,\infty)$ and $\mu\in[1,\infty)$. Let P be an arbitrary probability in $\mathcal{H}^{d,\beta}_7$. Denote by η the conditional probability function $P(\{1\}|\cdot)$ of P. According to Lemma C.2 and the definition of $\mathcal{H}^{d,\beta}_7$, there exists a function $f^*\in\mathcal{B}^\beta_1\left([0,1]^d\right)$ such that

$$f^*_{\phi,P} \xlongequal{P_X\text{-a.s.}} \log\frac{\eta}{1-\eta} \xlongequal{P_X\text{-a.s.}} f^*. \tag{C.145}$$

Thus there exists a measurable set Ω contained in $[0,1]^d$ such that $P_X(\Omega)=1$ and

$$\log\frac{\eta(x)}{1-\eta(x)} = f^*(x),\ \forall\, x\in\Omega. \tag{C.146}$$

Let δ be an arbitrary number in $(0,1/3)$. Then it follows from Corollary B.1 that there exists

$$\tilde{g} \in \mathcal{F}_d^{\mathbf{FNN}}\left(C_{d,\beta}\log\frac{1}{\delta}, C_{d,\beta}\delta^{-d/\beta}, C_{d,\beta}\delta^{-d/\beta}\log\frac{1}{\delta}, 1, \infty\right) \tag{C.147}$$

such that $\sup_{x\in[0,1]^d}|f^*(x)-\tilde{g}(x)| \leq \delta$. Let $T:\mathbb{R}\to[-1,1], z\mapsto \min\{\max\{z,-1\},1\}$ and

$$\tilde{f}:\mathbb{R}\to[-1,1],\ x\mapsto T(\tilde{g}(x)) = \begin{cases} -1, & \text{if } \tilde{g}(x)<-1,\\ \tilde{g}(x), & \text{if } -1\leq\tilde{g}(x)\leq 1,\\ 1, & \text{if } \tilde{g}(x)>1.\end{cases}$$

Obviously, $|T(z)-T(w)|\leq|z-w|$ for any real numbers z and w, and

$$\begin{aligned}&\sup_{x\in[0,1]^d}\left|f^*(x)-\tilde{f}(x)\right| \xlongequal{\because \|f^*\|_{[0,1]^d}\leq 1} \sup_{x\in[0,1]^d}|T(f^*(x))-T(\tilde{g}(x))|\\ &\leq \sup_{x\in[0,1]^d}|f^*(x)-\tilde{g}(x)|\leq\delta.\end{aligned} \tag{C.148}$$

Besides, it is easy to verify that

$$\tilde{f}(x)=\sigma(\tilde{g}(x)+1)-\sigma(\tilde{g}(x)-1)-1,\ \forall\, x\in\mathbb{R}^d,$$

which, together with (C.147), yields

$$\begin{aligned}\tilde{f} &\in \mathcal{F}_d^{\mathbf{FNN}}\left(1+C_{d,\beta}\log\frac{1}{\delta}, 1+C_{d,\beta}\delta^{-d/\beta}, 4+C_{d,\beta}\delta^{-d/\beta}\log\frac{1}{\delta}, 1, 1\right)\\ &\subset \mathcal{F}_d^{\mathbf{FNN}}\left(C_{d,\beta}\log\frac{1}{\delta}, C_{d,\beta}\delta^{-d/\beta}, C_{d,\beta}\delta^{-d/\beta}\log\frac{1}{\delta}, 1, 1\right).\end{aligned}$$

In addition, it follows from Lemma C.7 that

$$\begin{aligned}&\frac{1}{2(\mathrm{e}^{\mu}+\mathrm{e}^{-\mu}+2)}|f(x)-f^*(x)|^2 \leq \int_{\{-1,1\}}(\phi(yf(x))-\phi(yf^*(x)))\,\mathrm{d}P(y|x)\\ &\leq\frac{1}{4}|f(x)-f^*(x)|^2,\ \forall \text{ measurable } f:[0,1]^d\to[-\mu,\mu],\ \forall\, x\in\Omega.\end{aligned} \tag{C.149}$$

Take $\widetilde{C}:=2(\mathrm{e}^{\mu}+\mathrm{e}^{-\mu}+2)$. Integrating both side with respect to x in (C.149) and using (C.148), we obtain

$$\begin{aligned}&\int_{[0,1]^d\times\{-1,1\}}(\phi(yf(x))-\phi(yf^*(x)))^2\mathrm{d}P(x,y)\\ &\leq\int_{[0,1]^d\times\{-1,1\}}(f(x)-f^*(x))^2\mathrm{d}P(x,y)=\int_{[0,1]^d}|f(x)-f^*(x)|^2\mathrm{d}P_X(x)\\ &\xlongequal{\because P_X(\Omega)=1}\int_{\Omega}\frac{\widetilde{C}}{2(\mathrm{e}^{\mu}+\mathrm{e}^{-\mu}+2)}|f(x)-f^*(x)|^2\mathrm{d}P_X(x)\\ &\leq\widetilde{C}\int_{\Omega}\int_{\{-1,1\}}(\phi(yf(x))-\phi(yf^*(x)))\,\mathrm{d}P(y|x)\mathrm{d}P_X(x)\\ &\xlongequal{\because P_X(\Omega)=1}\widetilde{C}\int_{[0,1]^d\times\{-1,1\}}(\phi(yf(x))-\phi(yf^*(x)))\,\mathrm{d}P(x,y)\\ &\xlongequal{\text{by Lemma C.3}}\widetilde{C}\mathcal{E}_P^{\phi}(f),\quad\forall \text{ measurale } f:[0,1]^d\to[-\mu,\mu],\end{aligned} \tag{C.150}$$

and

$$\begin{aligned}
&\inf\left\{\mathcal{E}_P^\phi(f)\middle| f\in\mathcal{F}_d^{\mathbf{FNN}}\left(C_{d,\beta}\log\frac{1}{\delta}, C_{d,\beta}\delta^{-\frac{d}{\beta}}, C_{d,\beta}\delta^{-\frac{d}{\beta}}\log\frac{1}{\delta},1,1\right)\right\}\\
&\leq \mathcal{E}_P^\phi(\tilde{f}) \xlongequal{\text{by Lemma C.3}} \int_{[0,1]^d}\int_{\{-1,1\}}\Big(\phi(y\tilde{f}(x))-\phi(yf^*(x))\Big)\mathrm{d}P(y|x)\mathrm{d}P_X(x)\\
&\xlongequal{\because P_X(\Omega)=1} \int_\Omega\int_{\{-1,1\}}\Big(\phi(y\tilde{f}(x))-\phi(yf^*(x))\Big)\mathrm{d}P(y|x)\mathrm{d}P_X(x)\\
&\leq \int_\Omega \frac{1}{4}\left|\tilde{f}(x)-f^*(x)\right|^2\mathrm{d}P_X(x) \leq \int_{[0,1]^d}\left|\tilde{f}(x)-f^*(x)\right|^2\mathrm{d}P_X(x)\leq\delta^2.
\end{aligned}\tag{C.151}$$

Take c to be the maximum of the three constants $C_{d,\beta}$ in (C.151). Hence $\mathrm{c}\in(0,\infty)$ only depends on (d,β). Now suppose (3.8) holds. Then it follows that there exists $l\in(0,\infty)$ not depending on n and P such that $N\cdot\left(\frac{\log^3 n}{n}\right)^{\frac{d}{d+2\beta}}>l$ and $\frac{S}{\log n}\cdot\left(\frac{\log^3 n}{n}\right)^{\frac{d}{d+2\beta}}>l$ for any $n>1/l$. We then take $\delta=\delta_n:=\left(\frac{\mathrm{c}}{l}\right)^{\frac{\beta}{d}}\cdot\left(\frac{(\log n)^3}{n}\right)^{\frac{1}{2+d/\beta}}\asymp\left(\frac{(\log n)^3}{n}\right)^{\frac{1}{2+d/\beta}}$. Thus $\lim_{n\to\infty}\frac{1}{n\cdot\delta_n}=0=\lim_{n\to\infty}\delta_n$, which means that $\frac{1}{n}\leq\delta_n<1/3$ for $n>C_{l,\mathrm{c},d,\beta}$. We then deduce from (C.151) that

$$\begin{aligned}
&\inf\left\{\mathcal{E}_P^\phi(f)\,\middle|\, f\in\mathcal{F}_d^{\mathbf{FNN}}(G,N,S,B,F)\right\}\\
&\leq\inf\left\{\mathcal{E}_P^\phi(f)\middle| f\in\mathcal{F}_d^{\mathbf{FNN}}\left(\mathrm{c}\log n, l\left|\frac{(\log n)^3}{n}\right|^{\frac{-d}{2\beta+d}}, l\left|\frac{(\log n)^3}{n}\right|^{\frac{-d}{2\beta+d}}\log n, B, F\right)\right\}\\
&=\inf\left\{\mathcal{E}_P^\phi(f)\middle| f\in\mathcal{F}_d^{\mathbf{FNN}}\left(\mathrm{c}\log n,\mathrm{c}\delta_n^{-\frac{d}{\beta}},\mathrm{c}\delta_n^{-\frac{d}{\beta}}\log n,B,F\right)\right\}\\
&\leq\inf\left\{\mathcal{E}_P^\phi(f)\middle| f\in\mathcal{F}_d^{\mathbf{FNN}}\left(\mathrm{c}\log\frac{1}{\delta_n},\mathrm{c}\delta_n^{-\frac{d}{\beta}},\mathrm{c}\delta_n^{-\frac{d}{\beta}}\log\frac{1}{\delta_n},B,F\right)\right\}\\
&\leq\inf\left\{\mathcal{E}_P^\phi(f)\middle| f\in\mathcal{F}_d^{\mathbf{FNN}}\left(C_{d,\beta}\log\frac{1}{\delta_n},C_{d,\beta}\delta_n^{-\frac{d}{\beta}},C_{d,\beta}\delta_n^{-\frac{d}{\beta}}\log\frac{1}{\delta_n},1,1\right)\right\}\\
&\leq\delta_n^2,\ \forall\, n>C_{l,\mathrm{c},d,\beta},
\end{aligned}\tag{C.152}$$

where we use the fact the infimum taken over a larger set is smaller. Define $W=3\vee\mathcal{N}\left(\mathcal{F}_d^{\mathbf{FNN}}(G,N,S,B,F),\frac{1}{n}\right)$. Then by taking $\mathcal{F}=\left\{f|_{[0,1]^d}\middle| f\in\mathcal{F}_d^{\mathbf{FNN}}(G,N,S,B,F)\right\}$, $\psi(x,y)=\phi(yf^*(x))$, $\Gamma=\widetilde{C}$, $M=2$, $\gamma=\frac{1}{n}$ in Theorem 2.1, and using (C.145), (C.150), (C.152), we deduce that

$$\begin{aligned}
&\boldsymbol{E}_{P^{\otimes n}}\left[\left\|\hat{f}_n^{\mathbf{FNN}}-f^*_{\phi,P}\right\|^2_{\mathcal{L}^2_{P_X}}\right]=\boldsymbol{E}_{P^{\otimes n}}\left[\left\|\hat{f}_n^{\mathbf{FNN}}-f^*\right\|^2_{\mathcal{L}^2_{P_X}}\right]\leq\widetilde{C}\boldsymbol{E}_{P^{\otimes n}}\left[\mathcal{E}_P^\phi(\hat{f}_n^{\mathbf{FNN}})\right]\\
&\xlongequal{\text{by Lemma C.3}}\widetilde{C}\boldsymbol{E}_{P^{\otimes n}}\left[\mathcal{R}_P^\phi\left(\hat{f}_n^{\mathbf{FNN}}\right)-\int_{[0,1]^d\times\{-1,1\}}\psi(x,y)\mathrm{d}P(x,y)\right]\\
&\leq\frac{500\cdot\left|\widetilde{C}\right|^2\cdot\log W}{n}+2\widetilde{C}\inf_{f\in\mathcal{F}}\left(\mathcal{R}_P^\phi(f)-\int_{[0,1]^d\times\{-1,1\}}\psi(x,y)\mathrm{d}P(x,y)\right)
\end{aligned}$$

$$\xlongequal{\text{by Lemma C.3}} \frac{500 \cdot \left|\widetilde{C}\right|^2 \cdot \log W}{n} + 2\widetilde{C} \inf_{f \in \mathcal{F}} \mathcal{E}_P^{\phi}(f) \leq \frac{500 \cdot \left|\widetilde{C}\right|^2 \cdot \log W}{n} + 2\widetilde{C}\delta_n^2$$

for $n > C_{l,\mathrm{c},d,\beta}$. Taking the supremum, we obtain,

$$\sup_{P \in \mathcal{H}_7^{d,\beta}} \boldsymbol{E}_{P^{\otimes n}} \left[\left\| \hat{f}_n^{\mathbf{FNN}} - f_{\phi,P}^* \right\|_{\mathcal{L}^2_{P_X}}^2 \right] \leq \frac{500 \cdot \left|\widetilde{C}\right|^2 \cdot \log W}{n} + 2\widetilde{C}\delta_n^2, \ \forall\, n > C_{l,\mathrm{c},d,\beta}. \tag{C.153}$$

Moreover, it follows from (3.8) and Corollary A.1 that

$$\begin{aligned} &\log W \leq (S + Gd + 1)(2G + 5)\log\left((\max\{N, d\} + 1) \cdot B \cdot (2nG + 2n)\right) \lesssim (G + S)G \log n \\ &\lesssim \left(\left(\frac{n}{\log^3 n}\right)^{\frac{d}{d+2\beta}} \log n + \log n\right) \cdot (\log n) \cdot (\log n) \lesssim n \cdot \left(\frac{(\log n)^3}{n}\right)^{\frac{2\beta}{d+2\beta}}. \end{aligned}$$

Plugging this into (C.153), we obtain

$$\begin{aligned} &\sup_{P \in \mathcal{H}_7^{d,\beta}} \boldsymbol{E}_{P^{\otimes n}} \left[\left\| \hat{f}_n^{\mathbf{FNN}} - f_{\phi,P}^* \right\|_{\mathcal{L}^2_{P_X}}^2 \right] \lesssim \frac{\log W}{n} + \delta_n^2 \\ &\lesssim \left|\frac{(\log n)^3}{n}\right|^{\frac{2\beta}{d+2\beta}} + \left|\left(\frac{(\log n)^3}{n}\right)^{\frac{1}{2+d/\beta}}\right|^2 \lesssim \left|\frac{(\log n)^3}{n}\right|^{\frac{2\beta}{d+2\beta}}, \end{aligned}$$

which proves the desired result. ■

References

Martin Anthony and Peter L Bartlett. *Neural Network Learning: Theoretical Foundations*. Cambridge University Press, New York, NY, 2009.

Kendall Atkinson and Weimin Han. *Theoretical Numerical Analysis: A Functional Analysis Framework*. Springer, New York, NY, third edition, 2009.

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

Peter L. Bartlett, Olivier Bousquet, and Shahar Mendelson. Local Rademacher complexities. *The Annals of Statistics*, 33(4):1497–1537, 2005.

Peter L Bartlett, Nick Harvey, Christopher Liaw, and Abbas Mehrabian. Nearly-tight VC-dimension and pseudodimension bounds for piecewise linear neural networks. *The Journal of Machine Learning Research*, 20(1):2285–2301, 2019.

Felipe Cucker and Ding Xuan Zhou. *Learning Theory: An Approximation Theory Viewpoint*. Cambridge University Press, New York, NY, 2007.

Lawrence C Evans. *Partial Differential Equations.* American Mathematical Society, Providence, RI, second edition, 2010.

Zhiying Fang, Han Feng, Shuo Huang, and Ding-Xuan Zhou. Theory of deep convolutional neural networks II: Spherical analysis. *Neural Networks*, 131:154–162, 2020.

Max H Farrell, Tengyuan Liang, and Sanjog Misra. Deep neural networks for estimation and inference. *Econometrica*, 89(1):181–213, 2021.

Han Feng, Shuo Huang, and Ding-Xuan Zhou. Generalization analysis of CNNs for classification on spheres. *IEEE Transactions on Neural Networks and Learning Systems*, 34(9): 6200–6213, 2023.

Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning.* MIT press, Cambridge, MA, 2016.

Xin Guo, Lexin Li, and Qiang Wu. Modeling interactive components by coordinate kernel polynomial models. *Mathematical Foundations of Computing*, 3(4):263–277, 2020.

Zheng-Chu Guo, Dao-Hong Xiang, Xin Guo, and Ding-Xuan Zhou. Thresholded spectral algorithms for sparse approximations. *Analysis and Applications*, 15(3):433–455, 2017.

László Györfi, Michael Kohler, Adam Krzyżak, and Harro Walk. *A Distribution-free Theory of Nonparametric Regression.* Springer, New York, NY, 2002.

Song Han, Jeff Pool, John Tran, and William J Dally. Learning both weights and connections for efficient neural networks. In *Proceedings of the Advances in Neural Information Processing Systems 28*, pages 1135–1143, Montreal, Canada, 2015.

Juncai He, Lin Li, and Jinchao Xu. Approximation properties of deep ReLU CNNs. *Research in the Mathematical Sciences*, 9(3):1–24, 2022.

Geoffrey Hinton, Li Deng, Dong Yu, George E Dahl, Abdel-rahman Mohamed, Navdeep Jaitly, Andrew Senior, Vincent Vanhoucke, Patrick Nguyen, Tara N Sainath, et al. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *IEEE Signal Processing Magazine*, 29(6):82–97, 2012.

Tianyang Hu, Ruiqi Liu, Zuofeng Shang, and Guang Cheng. Minimax optimal deep neural network classifiers under smooth decision boundary. *arXiv preprint arXiv:2207.01602*, 2022a.

Tianyang Hu, Jun Wang, Wenjia Wang, and Zhenguo Li. Understanding square loss in training overparametrized neural network classifiers. In *Proceedings of the Advances in Neural Information Processing Systems 35*, pages 16495–16508, New Orleans, LA, United States, 2022b.

Like Hui and Mikhail Belkin. Evaluation of neural architectures trained with square loss vs cross-entropy in classification tasks. In *Proceedings of the Ninth International Conference on Learning Representations*, pages 1–17, Virtual Event, 2021.

Masaaki Imaizumi and Kenji Fukumizu. Deep neural networks learn non-smooth functions effectively. In *Proceedings of the Twenty-second International Conference on Artificial Intelligence and Statistics*, pages 869–878, Naha, Okinawa, Japan, 2019.

Mohit Iyyer, Varun Manjunatha, Jordan Boyd-Graber, and Hal Daumé III. Deep unordered composition rivals syntactic methods for text classification. In *Proceedings of the Fifty-third Annual Meeting of the Association for Computational Linguistics and the Seventh International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1681–1691, Beijing, China, 2015.

Katarzyna Janocha and Wojciech Czarnecki. On loss functions for deep neural networks in classification. *Schedae Informaticae*, 25:49–59, 2016.

Ziwei Ji, Justin Li, and Matus Telgarsky. Early-stopped neural networks are consistent. In *Proceedings of the Advances in Neural Information Processing Systems 34*, pages 1805–1817, Virtual Event, 2021.

Iain M. Johnstone. Oracle inequalities and nonparametric function estimation. In *Proceedings of the Twenty-third International Congress of Mathematicians (Volume III)*, pages 267–278, Berlin, Germany, 1998.

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

Diederik P Kingma and Jimmy Ba. ADAM: A method for stochastic optimization. In *Proceedings of the Third International Conference on Learning Representations*, pages 1–15, San Diego, CA, United States, 2015.

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, Adam Krzyżak, and Benjamin Walter. On the rate of convergence of image classifiers based on convolutional neural networks. *Annals of the Institute of Statistical Mathematics*, 74:1085–1108, 2022.

Vladimir Koltchinskii. Local Rademacher complexities and oracle inequalities in risk minimization. *The Annals of Statistics*, 34(6):2593–2656, 2006.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Proceedings of the Advances in Neural Information Processing Systems 25*, pages 1097–1105, Lake Tahoe, NV, United States, 2012.

Shao-Bo Lin and Ding-Xuan Zhou. Distributed kernel-based gradient descent algorithms. *Constructive Approximation*, 47(2):249–276, 2018.

Yufeng Liu, Hao Helen Zhang, and Yichao Wu. Hard or soft classification? Large-margin unified machines. *Journal of the American Statistical Association*, 106(493):166–177, 2011.

Tong Mao, Zhongjie Shi, and Ding-Xuan Zhou. Theory of deep convolutional neural networks III: Approximating radial functions. *Neural Networks*, 144:778–790, 2021.

Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. *Foundations of Machine Learning*. MIT press, Cambridge, MA, second edition, 2018.

Kevin P Murphy. *Machine Learning: A Probabilistic Perspective*. MIT press, Cambridge, MA, 2012.

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

Tomaso Poggio, Fabio Anselmi, and Lorenzo Rosasco. I-theory on depth vs width: Hierarchical function composition. Technical report, Center for Brains, Minds and Machines, MIT, Cambridge, MA, 2015.

Tomaso Poggio, Hrushikesh Mhaskar, Lorenzo Rosasco, Brando Miranda, and Qianli Liao. Why and when can deep–but not shallow–networks avoid the curse of dimensionality: A review. *arXiv preprint arXiv:1611.00740*, 2016.

Tomaso Poggio, Hrushikesh Mhaskar, Lorenzo Rosasco, Brando Miranda, and Qianli Liao. Why and when can deep-but not shallow-networks avoid the curse of dimensionality: A review. *International Journal of Automation and Computing*, 14(5):503–519, 2017.

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

Guohao Shen, Yuling Jiao, Yuanyuan Lin, and Jian Huang. Non-asymptotic excess risk bounds for classification with deep convolutional neural networks. *arXiv preprint arXiv:2105.00292*, 2021.

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *Proceedings of the Third International Conference on Learning Representations*, pages 1–14, San Diego, CA, United States, 2015.

Ingo Steinwart and Andreas Christmann. *Support Vector Machines*. Springer, New York, NY, 2008.

Charles J. Stone. Optimal global rates of convergence for nonparametric regression. *The Annals of Statistics*, 10(4):1040–1053, 1982.

Taiji Suzuki. Adaptivity of deep ReLU network for learning in Besov and mixed smooth Besov spaces: Optimal rate and curse of dimensionality. In *Proceedings of the Seventh International Conference on Learning Representations*, pages 1–25, New Orleans, LA, United States, 2019.

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

Alexandre B Tsybakov. *Introduction to Nonparametric Estimation*. Springer, New York, NY, 2009.

Martin J Wainwright. *High-dimensional Statistics: A Non-asymptotic Viewpoint*. Cambridge University Press, Cambridge, 2019.

Dao-Hong Xiang. Classification with Gaussians and convex loss II: Improving error bounds by noise conditions. *Science China Mathematics*, 54(1):165–171, 2011.

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

Zihan Zhang, Lei Shi, and Ding-Xuan Zhou. Convolutional neural networks for spherical data classification. *preprint*, 2024.

Ding-Xuan Zhou. Deep distributed convolutional neural networks: Universality. *Analysis and Applications*, 16(6):895–919, 2018.

Ding-Xuan Zhou. Theory of deep convolutional neural networks: Downsampling. *Neural Networks*, 124:319–327, 2020a.

Ding-Xuan Zhou. Universality of deep convolutional neural networks. *Applied and Computational Harmonic Analysis*, 48(2):787–794, 2020b.