Electronic Journal of Statistics
Vol. 17 (2023) 3613–3659
ISSN: 1935-7524
https://doi.org/10.1214/23-EJS2187

# Optimal convergence rates of deep neural networks in a classification setting

**Joseph T. Meyer**

*Institute for Applied Mathematics, Heidelberg University, Im Neuenheimer Feld 205, 69120 Heidelberg, Germany*
*e-mail:* joseph-theo.meyer@uni-heidelberg.de

**Abstract:** We establish optimal convergence rates up to a log factor for a class of deep neural networks in a classification setting under a restraint sometimes referred to as the Tsybakov noise condition. We construct classifiers based on empirical risk minimization in a general setting where the boundary of the Bayes rule can be approximated well by neural networks. Corresponding rates of convergence are proven with respect to the misclassification error using an additional condition that acts as a requirement for the "correct noise exponent". It is then shown that these rates are optimal in the minimax sense. For other estimation procedures, similar convergence rates have been established. Our first main contribution is to prove that the rates are optimal under the additional condition. Secondly, our main theorem establishes almost optimal rates in a generalized setting. We use this to show optimal rates which circumvent the curse of dimensionality.

**MSC2020 subject classifications:** 62C20, 62G05.
**Keywords and phrases:** Tsybakov noise condition, classification, deep neural networks.

Received June 2022.

## Contents

## 1. Introduction

We consider i.i.d. data $(Y_i, X_i)_{i=1}^n$ with $Y_i \in \{0,1\}$ and $X_i \in \mathbb{R}^d$. Our goal is to provide an estimator of the form $\widehat{Y} = \mathbb{1}(X \in \widehat{G})$ where $\widehat{G}$ is constructed with a neural network which approximates $Y$ well with respect to the misclassification error. We show optimal convergence rates up to a log factor under the following two conditions. First, the underlying distribution $\mathbb{Q}$ satisfies a noise condition as in [24] described below. Second, the boundary of the set

$$G^*_{\mathbb{Q}} := \left\{ x \;\middle|\; f_{\mathbb{Q}}(x) \geq \frac{1}{2} \right\}$$

with $f_{\mathbb{Q}}(x) := \mathbb{Q}(Y = 1 | X = x)$ satisfies certain regularity conditions.

Neural Networks have shown outstanding results in many classification tasks such as image recognition [7], language recognition [4], cancer recognition [10], and other disease detection [17]. Our work follows current approaches in the statistical literature to explain the success of neural networks, e.g. the impactful contributions [12, 22]. The objective is to fill a gap in the literature by proving optimal convergence rates (up to a log factor) in a specific setting which was also considered in [11]. We wish to obtain the same optimal convergence rates as in [18, 24] and focus on deep feedforward neural networks with ReLU-activation functions. Deep networks have been considered in many theoretical articles [12, 13, 19, 20] and have proven useful in many applications [16, 21]. Intuitively we wish to approximate the set $G^*_{\mathbb{Q}}$ directly instead of approximating the regression function $f_{\mathbb{Q}}$. The classification setting we consider is similar to the setting given in [18, 24]. In particular, we assume that $\mathbb{Q}$ satisfies a noise condition which can be described as follows. For $\mathbb{Q}$-measurable sets $G_1, G_2$ define

$$\begin{aligned} d_{f_{\mathbb{Q}}}(G_1, G_2) &:= \int_{G_1 \Delta G_2} |2 f_{\mathbb{Q}}(x) - 1| \; \mathbb{Q}_X(\mathrm{d}x), \\ d_{\Delta}(G_1, G_2) &:= \mathbb{Q}_X(G_1 \Delta G_2), \end{aligned}$$

where $G_1 \Delta G_2 := (G_1 \backslash G_2) \cup (G_2 \backslash G_1)$ is the symmetric difference. The condition then states that there exists a constant $\kappa \geq 1$ such that

$$d_{f_{\mathbb{Q}}}(G, G^*_{\mathbb{Q}}) \geq c_1 d^{\kappa}_{\Delta}(G, G^*_{\mathbb{Q}}) \tag{1.1}$$

for some constant $c_1 > 0$ and all $G$. This requirement is sometimes referred to as the Tsybakov noise condition. It can be interpreted as a restraint on the probability distribution regarding regions where $f_{\mathbb{Q}}(x)$ takes on values close to $\frac{1}{2}$. Roughly speaking it forces the mass to decay at a certain rate when one approaches the boundary $f_{\mathbb{Q}}(x) = \frac{1}{2}$. Using this one can achieve rates approaching $n^{-1}$ for small $\kappa$, i.e. if there is not much mass where $f_{\mathbb{Q}}(x) \approx \frac{1}{2}$. The condition has been used in many statistical articles considering classification such as [1] and [25] who analyze support vector machines. Similarly to [18], we show optimal convergence rates in the case where the boundary of $G^*_{\mathbb{Q}}$ satisfies certain regularity conditions, i.e. is similar to an element of a Dudley class [6]. More precisely, we consider sets that are slightly more general than the sets given in [19]. While many other approximation results using neural networks exist, see [5] using sigmoid activation functions or [26] using piecewise linear functions among others, the methods used in [19] inspired us to obtain the results for our setting. The sets they consider have been used in many articles such as [20]. As an estimator, we use a risk minimizer of the empirical version of the misclassification error. Precisely calculating this estimator involves finding a global minimum of a highly non-convex loss with respect to the parameters of a neural network. Typically, such calculations are not feasible in practice. Thus the results we provide are theoretical in nature and do not have direct useful applications, as is typical for results of this kind [13, 22]. From our point of view, the main value of current contributions is to show results such as consistency in situations that are typical for statisticians using relatively simple classes of neural networks. In time, the techniques developed may be used to show claims in cases that are closer to those encountered in reality and using classes of neural networks that are closer to those used in practice.

A lot of work has been done regarding consistency of feedforward deep neural networks. [22] prove optimal convergence rates up to a log factor with respect to the uniform norm in a regression setting. Among others similar results were given by [9] for non-continuous regression functions with respect to the $L_2$-norm, [14] who did not use a sparsity constraint, and [2]. Regarding results for classification, [20] show convergence rates considering the misclassification error in a noiseless setting. Consistency results which include condition (1.1) in the assumptions are given by [3, 8, 13]. In contrast to our approach, the previously mentioned articles attempt to estimate the regression function $f_{\mathbb{Q}}$ instead of directly estimating the set $G^*_{\mathbb{Q}}$. Additionally, while some obtain optimal convergence rates, the settings do not correspond to the setting given in [24]. In particular, the (optimal) convergence rates differ from ours in these papers.

A very interesting contribution was made by [11] who consider estimators based on the hinge loss instead of empirical risk minimization. In Theorem 3.2, they show almost optimal rates as in Corollary 3.8 in a similar situation using a similar additional condition. We include the corollary since it differs from their result in the following manner. (1) They prove consistency for a different estimator. The underlying proofs differ significantly. (2) The sets they use for estimation are generalizations of boundary fragments. This simplifies the proofs.

Examples of sets not covered by their definition are given in Appendix A.2. (3) They do not prove lower bounds under the additional condition.

### 1.1. Contribution

Our contribution includes the following.

- Theorem 3.5 establishes convergence rates in a general setting where the boundary of the set $G^*_{\mathbb{Q}}$ can be well approximated by neural networks. It utilizes (1.1) as well as a condition which forces $\kappa$ to be the "correct parameter" for the distribution $\mathbb{Q}$. This enables us to prove rates in a variety of settings.
- Corollary 3.8 together with Theorem 4.1 prove optimal convergence rates up to a log factor in the minimax sense for the setting described above. In particular, Theorem 4.1 shows that the rates from [11, 18, 24] remain optimal when including the additional constraint.
- We then use Theorem 3.5 to prove convergence rates up to a log factor under an additional constraint which circumvents the curse of dimensionality (Corollary 3.11) in the sense that the rates do not decrease exponentially in the dimension $d$. Theorem 4.2 shows that the rates obtained are optimal up to a log factor.

### 1.2. Outline

After introducing some notation we rigorously introduce the problem at hand in Section 2. We provide some convergence results considering empirical risk minimizers with respect to arbitrary sets. These results are then used to prove our main consistency theorems regarding neural networks in Section 3. Section 4 includes the corresponding lower bounds followed by some concluding remarks in Section 5.

### 1.3. Notation

We introduce some general notation which is used throughout this article.

For $x \in \mathbb{R}$ let $\lfloor x \rfloor := \max\{k \in \mathbb{Z} \mid k \leq x\}$ and $\lceil x \rceil := \min\{k \in \mathbb{Z} \mid k \geq x\}$. Let $\lambda$ be the Lebesgue measure. For a function $g : \Omega \subseteq \mathbb{R}^s \to \mathbb{R}$ and $k \in \mathbb{N}$ denote by

$$\|g\|_\infty := \sup_{x \in \Omega} |g(x)|, \quad \|g\|_{L^k} := \left( \int |g|^k \, \lambda(\mathrm{d}x) \right)^{\frac{1}{k}}$$

the uniform norm and the $L^k$-norm respectively. Note that we omit the dependence on $\Omega$ in the notation. For $x \in \mathbb{R}^s$, let $\|x\|_2$ and $\|x\|_\infty$ be the euclidean-norm and the uniform-norm respectively. For $j \in \{1, \dots, s\}$ let

$$x_{-j} := (x_1, \dots, x_{j-1}, x_{j+1}, \dots, x_s).$$

Additionally, let

$$\mathcal{B}_r(x) := \{y \in \mathbb{R}^s \mid \|x-y\|_\infty \le r\}, \quad \mathcal{B}_r^\circ(x) := \{y \in \mathbb{R}^s \mid \|x-y\|_\infty < r\}.$$

For $a \in \mathbb{N}^s$ let $|a| := \sum_{i=1}^s a_i$.

Now, let $\beta \in (0,\infty)$. Define $m := \max\{k \in \mathbb{N} \mid k < \beta\}$ and $\omega := \beta - m > 0$. For $f \in \mathcal{C}\big([0,1]^s, \mathbb{R}\big)$ let

$$\|f\|_{\mathcal{C}^\beta} := \sum_{|\alpha| \le m} \big\|\partial^\alpha f\big\|_\infty + \sum_{|\alpha|=m} \sup_{x \ne y} \frac{|\partial^\alpha f(x) - \partial^\alpha f(y)|}{\|x-y\|_\infty^\omega}$$

be the Hölder-norm. For $B > 0$, define the class of Hölder-continuous functions by

$$\mathcal{F}_{\beta,B,s} := \Big\{ f \in \mathcal{C}\big([0,1]^s, \mathbb{R}\big) \;\Big|\; \|f\|_{\mathcal{C}^\beta} \le B \Big\}.$$

Let $G_1, G_2 \subseteq \Omega$ be two subsets. We write

$$G_1 \Delta G_2 := (G_1 \backslash G_2) \cup (G_2 \backslash G_1), \qquad \mathbb{1}(x \in G_1) := \begin{cases} 1, & \text{for } x \in G_1, \\ 0, & \text{otherwise} \end{cases}$$

for their symmetric difference and the indicator function corresponding to $G_1$ respectively.

## 2. General convergence results

In this section, we state our results in a relatively general setting. The results on neural networks in the next section only consider the case where $\mathbb{Q}_X$ has a bounded density with respect to the Lebesgue measure. Our setup is similar to the binary classification setup of [24].

### *2.1. Classification setup*

Let $(X_i, Y_i)_{i=1}^n$ be *i.i.d.* observations distributed according to some probability measure $\mathbb{Q}$, where $X_i \in \mathbb{R}^d$ and $Y_i \in \{0,1\}$. Denote by $\mathbb{Q}_X$ the marginal probability distribution with respect to $X \in \mathbb{R}^d$. The goal is to predict $Y \in \{0,1\}$ when observing $X \in \mathbb{R}^d$, where $(X,Y)$ is distributed according to $\mathbb{Q}$ independently of $(X_i, Y_i)_{i=1}^n$ using classifiers of the form

$$\widehat{Y} := \mathbb{1}(X \in \widehat{G})$$

for some $\mathbb{Q}$-measurable set $\widehat{G} \subseteq \mathbb{R}^d$. Note that a classifier is uniquely determined by $\widehat{G}$. Performance is measured by the misclassification error

$$R(\widehat{G}) := \mathbb{P}\big(Y \ne \widehat{Y}\big) = \mathbb{E}\Big[\big(Y - 1(X \in \widehat{G})\big)^2\Big].$$

For $f_{\mathbb{Q}}(x) := \mathbb{E}[Y|X = x] = \mathbb{Q}(Y = 1|X = x)$ the set

$$G^*_{\mathbb{Q}} := \left\{ x \;\middle|\; f_{\mathbb{Q}}(x) \geq \frac{1}{2} \right\}$$

minimizes the misclassification error. A set with this property is called a Bayes rule. Classification can equivalently be seen as estimation of $G^*_{\mathbb{Q}}$ by the set $\widehat{G}$, which is therefore equally referred to as a classifier. For a $\mathbb{Q}$-measurable set $G \subseteq \mathbb{R}^d$ let

$$R_n(G) = \frac{1}{n} \sum_{i=1}^{n} 1\big(Y_i \neq 1(X_i \in G)\big) = \frac{1}{n} \sum_{i=1}^{n} \big(Y_i - 1(X_i \in G)\big)^2$$

be the empirical version of the misclassification error $R(G)$. We consider empirical risk minimization classifiers defined by

$$\widehat{G}_n := \operatorname*{arg\,min}_{G \in \mathcal{N}_n} R_n(G)$$

where $\mathcal{N}_n$ is some finite collection of $\mathbb{Q}$-measurable sets for all $n \in \mathbb{N}$.

### *2.2. Consistency results*

Proposition 2.1 establishes convergences rates for estimating $G^*_{\mathbb{Q}}$ using $\widehat{G}_n$ under certain conditions on $\mathcal{N}_n$ and $\mathbb{Q}$. For the loss function we consider a slight generalization of the misclassification error

$$\mathbb{E}\big[\big(R(\widehat{G}_n) - R(G^*_{\mathbb{Q}})\big)^p\big] = \mathbb{E}\big[d^p_{f_{\mathbb{Q}}}(\widehat{G}_n, G^*_{\mathbb{Q}})\big]$$

for $p \geq 1$. The proposition is somewhat similar to Theorem 2 from [18]. In contrast to our approach, they consider the discrimination of two probability distributions with underlying distribution functions and do not allow for non-optimal convergence rates. The proposition is an important component for the proof of our main theorems given in Section 3. The proofs of this section can be found in Appendix B.

**Proposition 2.1.** *Let $\tau_n > 0$ be a monotonically increasing sequence. Let $\mathfrak{Q}$ be a class of potential joint distributions $\mathbb{Q}$ of $(X, Y)$ and $\mathcal{N}_n$ be a collection of subsets of $\mathbb{R}^d$ for all $n \in \mathbb{N}$ such that the following conditions hold.*

*(i) For all $\mathbb{Q} \in \mathfrak{Q}$ all sets in $\bigcup_{n\in\mathbb{N}} \mathcal{N}_n$ and $G^*_{\mathbb{Q}}$ are $\mathbb{Q}$-measurable.*
*(ii) There exists a constant $\kappa \geq 1$ such that*

$$d_{f_{\mathbb{Q}}}(G, G^*_{\mathbb{Q}}) \geq c_1 d^{\kappa}_{\Delta}(G, G^*_{\mathbb{Q}})$$

*for some constant $c_1 > 0$, all $G \in \bigcup_{n\in\mathbb{N}} \mathcal{N}_n$ and all $\mathbb{Q} \in \mathfrak{Q}$.*

*Additionally, we assume that there is a constant $N_0 \in \mathbb{N}$ such that for all $n \geq N_0$ the following holds.*

*(iii) There is a constant* $c_2 > 0$ *such that for all* $\mathbb{Q} \in \mathfrak{Q}$ *there is a* $G \in \mathcal{N}_n$ *with*

$$d_{f_{\mathbb{Q}}}(G, G^*_{\mathbb{Q}}) \leq c_2 \tau_n^{-\kappa}.$$

*(iv) There exist constants* $c_3, \rho > 0$ *such that*

$$\log\big(|\mathcal{N}_n|\big) \leq c_3 n^{\frac{\rho}{\rho+2\kappa-1}}.$$

*Then for all* $p \geq 1$ *we have*

$$\limsup_{n\to\infty} \sup_{\mathbb{Q}\in\mathfrak{Q}} \tilde{\tau}_n^{\kappa p}\, \mathbb{E}\big[d^p_{f_{\mathbb{Q}}}(\widehat{G}_n, G^*_{\mathbb{Q}})\big] < \infty,$$
$$\limsup_{n\to\infty} \sup_{\mathbb{Q}\in\mathfrak{Q}} \tilde{\tau}_n^{p}\, \mathbb{E}\big[d^p_{\Delta}(\widehat{G}_n, G^*_{\mathbb{Q}})\big] < \infty,$$

*where* $\tilde{\tau}_n := \min\{\tau_n, n^{\frac{1}{\rho+2\kappa-1}}\}$ *for all* $n \in \mathbb{N}$.

Condition (i) is needed for all terms to be well defined. Condition (iii) states that the set in question must be well approximated by elements of $\mathcal{N}_n$. A sufficient assumption is that $\mathcal{N}_n$ is an $\epsilon$-net of $\big\{G^*_{\mathbb{Q}} \;\big|\; \mathbb{Q} \in \mathfrak{Q}\big\}$, where $\epsilon := c_1 \tau_n^{-\kappa}$. Together with (iv) this indirectly bounds the complexity of $\mathfrak{Q}$. If the class of sets $\{G^*_{\mathbb{Q}} \mid \mathbb{Q} \in \mathfrak{Q}\}$ is too large one will not be able to find sets $\mathcal{N}_n$ that satisfy (iii) and (iv) at the same time. It is clear that the best rates are achieved with $\tau_n = n^{\frac{1}{\rho+2\kappa-1}}$. We do not use the same sequences in conditions (iii) and (iv) since one can prove non-optimal convergence rates using this version of the proposition.

The second condition is the noise condition described in the introduction. Note that following [24] for $\kappa > 1$ condition (ii) holds if

$$\mathbb{P}\Big(\big|2f_{\mathbb{Q}}(X) - 1\big| \leq t\Big) \leq c t^{\frac{1}{\kappa-1}}$$

for all $t > 0$ and some $c > 0$. Roughly speaking, this forces the mass to decay at a certain rate when approaching $f_{\mathbb{Q}} = \frac{1}{2}$. Note that we use (ii) instead of this assumption since it is slightly more general and includes the case $\kappa = 1$. Additionally, it appears more naturally in the proofs. Observing the alternative assumption $\kappa = 1$ corresponds to the case where $f_{\mathbb{Q}}$ does not take on values close to $\frac{1}{2}$. Using Proposition 2.1 one can achieve rates approaching $n^{-1}$ for small $\kappa, \rho$ i.e. if there is not much mass in the region around $f_{\mathbb{Q}}(x) = \frac{1}{2}$ and the complexity of $\mathcal{N}_n$, and consequently $\mathfrak{Q}$, is moderate. In Appendix B.1 we present convergence rates in the case when condition (ii) is not satisfied. In contrast to the results above, the rates are always slower than $n^{-\frac{1}{2}}$.

## 3. Convergence rates for neural networks

We begin by shortly introducing neural networks. The idea is to use Proposition 2.1 to obtain optimal convergence rates up to a log factor. Neural networks are used to define a suitable class of sets $\mathcal{N}_n$ for every $n \in \mathbb{N}$.

### *3.1. Definitions regarding neural networks*

**Definition 3.1.** *Let $L, m_0, \dots, m_{L+1} \in \mathbb{N}$. For $i = 1, \dots, L$, let $\sigma_i$ be a function*

$$\sigma_i : \mathbb{R} \to \mathbb{R}.$$

*For $b = (b_1, \dots, b_{m_i}) \in \mathbb{R}^{m_i}$, define a shifted $m_i$-dimensional version of $\sigma_i$ by*

$$\sigma_{i,b} : \mathbb{R}^{m_i} \to \mathbb{R}^{m_i},\ \sigma_{i,b}(y_1, \dots, y_{m_i}) = \big(\sigma_i(y_1 - b_1), \dots, \sigma_i(y_{m_i} - b_{m_i})\big).$$

*A **neural network** with network architecture $\big(L, (m_0, \dots, m_{L+1}), (\sigma_1, \dots, \sigma_L)\big)$ is a sequence*

$$\Phi := \big(W_1, b_1, \dots, W_L, b_L, W_{L+1}\big)$$

*where each $W_s \in \mathbb{R}^{m_s \times m_{s-1}}$ is a weight matrix and $b_s \in \mathbb{R}^{m_s}$ is a shift vector. The realization of a neural network $\Phi$ on a set $D \subseteq \mathbb{R}^{m_0}$ is the function*

$$R(\Phi) : D \to \mathbb{R}^{m_{L+1}},\ R(\Phi)(x) = W_{L+1}\sigma_{L,b_L}W_L \cdots W_2\sigma_{1,b_1}W_1x.$$

*We denote by*

$$\mathcal{N}_{L,m,\sigma} := \big\{R(\Phi) \mid \Phi = \big(W_1, b_1, \dots, b_L, W_{L+1}\big),\ W_s \in \mathbb{R}^{m_s \times m_{s-1}}, b_s \in \mathbb{R}^{m_s}\big\}$$

*the set of realizations of neural networks with network architecture $(L, m, \sigma)$, where $m := (m_0, \dots, m_{L+1}) \in \mathbb{N}^{L+2}$ and $\sigma := (\sigma_1, \dots, \sigma_L)$.*

Typically, for $i = 1, \dots, L$ the function $\sigma_i$ is called the activation function and $\sigma_{i,b}$ is named the shifted activation function. The constant $L$ denotes the number of hidden Layers. The values $d := m_0$ and $m_{L+1}$ are the input and output dimensions respectively. In this article, we are interested in the case where for $i = 1, \dots, L$ the activation function in the $i$-th layer is the rectifier linear unit (ReLU)

$$\sigma_i(x) := \max\{x, 0\}.$$

Additionally, if not further specified, we consider a compact domain $D := [0, 1]^d$ and a one-dimensional output $m_{L+1} = 1$. For the sake of completeness, we note that a network with $L = 0$ layers is of the form $\Phi := (W)$ for $W \in \mathbb{R}^{m_0 \times m_1}$ and has realization $R(\Phi)(x) = Wx$. Note that in general, the weights of a neural network $\Phi$, i.e. the entries of its shift vectors $(b_1, \dots, b_{L+1})$ and weight matrices $(W_1, \dots, W_{L+1})$, are not uniquely determined by its realization $R(\Phi)$. In the following, for brevity, we occasionally introduce a network by defining its realization. In such a case, it is clear from the presentation of the realization which precise neural network is considered.

As a first step, we wish to introduce a suitable finite class of sets parameterized by neural networks and count the number of elements. We define these sets as $R(\Phi)^{-1}(1)$ where $\Phi$ is a realization of a neural network. Equivalently, we could have considered neural networks with a binary step function in the output layer or find a neural network $\tilde{\Phi}$ and define the approximating set $R(\tilde{\Phi})^{-1}((0.5, 1])$, which is closer to the idea that the realization of the neural network represents

some sort of probability. Since this is not the idea of our approximation results, we stick to the version above. To obtain a finite class, we need to reduce the number of considered elements of $\mathcal{N}_{L,m,\sigma}$ while maintaining reasonable approximating capabilities. A typical approach in the theoretical literature is to use a sparsity constraint. For $s > 1$ we therefore only consider realizations of neural nets which have at most $s$ nonzero weights. If $s$ is the total number of nonzero weights, we say that the network has sparsity $s$. Additionally, we assume all weights to be elements of the set

$$\mathcal{W}_c := \left\{k2^{-c} \;\middle|\; c \in \mathbb{N},\ k \in \{-2^c, -2^c + 1, \dots, 2^c - 1, 2^c\}\right\}.$$

Thus we only consider weights $|w| \leq 1$. Concluding, we use the following notation to describe the collection of sets we are interested in.

**Definition 3.2.** *Let $L_0, c \in \mathbb{N}$ and $s_0 > 1$ be fixed. Denote by $\tilde{\mathcal{N}}_{L_0,s_0,c}$ the set of realizations of neural networks with $d$ dimensional input, one-dimensional output, at most $L_0$ layers, ReLU activation functions and sparsity at most $s_0$, where all weights are elements of $\mathcal{W}_c$. The class of corresponding sets given by neural networks is then*

$$\mathcal{N}_{L_0,s_0,c} := \left\{R(\Phi)^{-1}(1) \subseteq [0,1]^d,\ \middle|\ \Phi \in \tilde{\mathcal{N}}_{L_0,s_0,c}\right\}.$$

Note that the requirements from Definition 3.2 allow for realizations of neural networks with arbitrary hidden layer dimensions $(m_1, \dots, m_L) \in \mathbb{N}^L$. However, it is easy to see that every element of $\tilde{\mathcal{N}}_{L_0,s_0,c}$ is a realization of a neural net that satisfies the properties described in the Definition and $m_i \leq s_0$ for all $i \in \{1, \dots, L_0\}$. Using this we receive an upper bound on the number of elements of $\mathcal{N}_{L_0,s_0,c}$ by counting the number of corresponding neural networks. This bound is central in order to show (iv) in Proposition 2.1 when proving Theorem 3.5. Note that the following result is independent of the choice of activation functions $\sigma_i$. The proof can be found in Appendix C.1.

**Lemma 3.3.** *For $s_0 > 1$ and $L_0, c \in \mathbb{N}$ let $\mathcal{N}_{L_0,s_0,c}$ be the class of sets introduced in Definition 3.2. We have an upper bound on the number of elements given by*

$$\left|\mathcal{N}_{L_0,s_0,c}\right| \leq \left((ds_0 + \min\{s_0, L_0\}(s_0+1)^2)2^{c+2}\right)^{s_0}.$$

### *3.2. Conditions on the Bayes rule*

In order to define the set of probability distributions we consider for approximation, we restrict the possible Bayes rules. Intuitively the boundary should satisfy some kind of smoothness condition so that it can be approximated by neural networks. Additionally, the set must be discretizable in some sense. When considering $\mathcal{F} = \mathcal{F}_{\beta,B,d-1}$ the class of sets we use is similar to a class defined in [19]. Note, that the class used here is larger. This version depends on a set $\mathcal{F}$ which represents a class of boundary functions. The idea is that we can obtain different convergence rates for different classes using the same procedure. Furthermore, we add a condition corresponding to (ii) of Proposition 2.1. It bounds $f_{\mathbb{Q}}$ near the boundary of the respective Bayes rule.

**Definition 3.4.** *Let $\beta \geq 0$, $B > 0$ and $d \in \mathbb{N}$ with $d \geq 2$. Additionally, let $r \in \mathbb{N}$, $\epsilon_1, \epsilon_2 > 0$, $\mathbb{Q}$ be a probability measure on $[0,1]^d \times \{0,1\}$ and $\mathcal{F}$ be a class of functions $\gamma : [0,1]^{d-1} \to \mathbb{R}$. Define*

$$\mathcal{I} := \{1, \dots, d\} \times \{-1, 1\}, \quad \mathcal{D} := \left\{ \prod_{i=1}^{d} [a_i, b_i] \;\middle|\; 0 \leq a_i < b_i \leq 1 \right\}.$$

*and*

$$\mathcal{H}_{\beta,B} := \left\{ g \in \mathcal{F}_{\beta,B,1} \;\middle|\; \forall \alpha < \beta : \partial^\alpha g(0) = 0 \right\}.$$

*Let $\mathcal{K}^{\mathcal{F}}_{\mathbb{Q},\beta,B,\epsilon_1,\epsilon_2,r,d}$ be the class of all subsets $H = H_1 \cup \dots \cup H_u \subseteq [0,1]^d$ for $u \in \{0, \dots, r\}$ such that for $\nu = 1, \dots, u$ there exist $(j_\nu, \iota_\nu) \in \mathcal{I}$, $\gamma_\nu \in \mathcal{F}$, $D_\nu = \prod_{i=1}^{d} [a_i^\nu, b_i^\nu] \in \mathcal{D}$ with the following properties.*

1. *For all $\nu = 1, \dots, u$ we have*

$$H_\nu = D_\nu \cap \{x \in [0,1]^d \mid \iota_\nu x_{j_\nu} \leq \gamma_\nu(x_{-j_\nu})\}.$$

2. *$\lambda(D_{\nu_1} \cap D_{\nu_2}) = 0$ for $\nu_1 \neq \nu_2$.*
3. *$b^\nu_{j_\nu} - a^\nu_{j_\nu} \geq \epsilon_2$ for all $\nu = 1, \dots, u$.*
4. *If $\beta > 0$, the following holds. For $\nu = 1, \dots, u$ and all $x \in D_\nu \cap \partial H$ there exists $g_{\nu,x} \in \mathcal{H}_{\beta,B}$ such that for $y_{j_\nu} \in [\max\{0, x_{j_\nu} - \epsilon_1\}, \min\{1, x_{j_\nu} + \epsilon_1\}]$ we have*

$$\begin{aligned} |2f_{\mathbb{Q}}(y) - 1| &\leq g_{\nu,x}(x_{j_\nu} - y_{j_\nu}) \quad \text{for } x_{j_\nu} - \epsilon_1 \leq y_{j_\nu} \leq x_{j_\nu}, \\ |2f_{\mathbb{Q}}(y) - 1| &\leq g_{\nu,x}(y_{j_\nu} - x_{j_\nu}) \quad \text{for } x_{j_\nu} \leq y_{j_\nu} \leq x_{j_\nu} + \epsilon_1, \end{aligned}$$

   *where $y = (x_1, \dots, x_{j_\nu - 1}, y_{j_\nu}, x_{j_\nu + 1}, \dots, x_d)$.*

Note that $g \in \mathcal{H}_{\beta,B}$ implies $g(0) = \partial^0 g(0) = 0$. The idea is to use sets defined by realizations of neural networks to approximate $G^*_{\mathbb{Q}} \in \mathcal{K}^{\mathcal{F}}_{\mathbb{Q},\beta,B,\epsilon_1,\epsilon_2,r,d}$ from Definition 3.4 for a suitable class $\mathcal{F}$ in order to apply Proposition 2.1. The first three conditions in the definition are assumptions on the shape of $G^*_{\mathbb{Q}}$. We assume that we can divide the set into at most $u$ cubes where the boundary of $G^*_{\mathbb{Q}}$ looks like a function of $\mathcal{F}$. Condition (2.) states that the cubes at most overlap on their boundaries while condition (3.) forces the cubes to be large enough. Figure 1 shows an example for an element of $\mathcal{K}^{\mathcal{F}}_{\mathbb{Q},\beta,B,\epsilon_1,\epsilon_2,12,2}$, where $\mathcal{F}$ is the set of piecewise linear functions. It highlights the first three conditions. Additionally, with (4.) we require that $f_{\mathbb{Q}}$ is bounded towards $\frac{1}{2}$ close to the boundary of $G^*_{\mathbb{Q}}$. Similar conditions are used in [11, 23]. An intuition why an additional constraint is necessary is the following: Let $\mathfrak{Q}_\kappa$ contain all probability distributions satisfying (i) and (ii) from Proposition 2.1 for $\kappa \geq 1$ with constants $c_2, c_3 > 0$. We argue that there exists a $\kappa > 1$ such that it is not possible to find a set $\mathcal{N}^\kappa_n$ which satisfies (iii) and (iv) for $\mathfrak{Q} = \mathfrak{Q}_\kappa$ at the same time with $\tau_n = \tau_{n,\kappa} \approx n^{\frac{1}{\rho + 2\kappa - 1}}$. Note that if condition (ii) in Proposition 2.1 is satisfied for $\kappa_0 \geq 1$, it is also satisfied for all $\kappa > \kappa_0$ which implies $\mathfrak{Q}_\kappa \supseteq \mathfrak{Q}_1$ for all $\kappa \geq 1$.

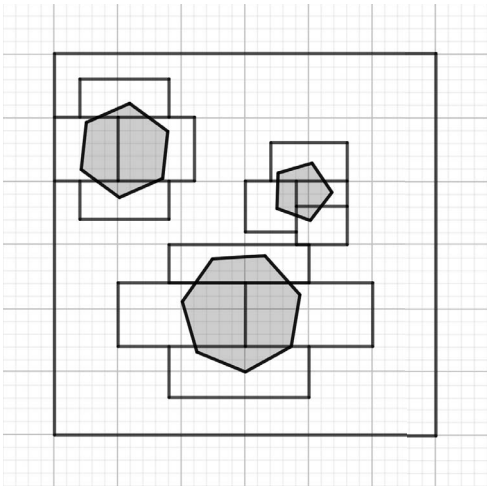

FIG 1. *Example for an element of* $\mathcal{K}^{\mathcal{F}}_{\mathbb{Q},\beta,B,\epsilon_1,\epsilon_2,12,2}$*, where* $\mathcal{F}$ *is the set of piecewise linear functions. The grey objects represent the set. The small boxes represent a possible choice for* $D_1,\dots,D_{12}$*.*

Thus, for all $n \in \mathbb{N}$ and $\kappa \geq 1$ we need a set $\mathcal{N}_n^\kappa$ such that $\forall \mathbb{Q} \in \mathfrak{Q}_1$ $\exists G \in \mathcal{N}_n^\kappa$ such that

$$d_{f_{\mathbb{Q}}}(G, G^*_{\mathbb{Q}}) \leq c_2 \tau_n^{-\kappa} \overset{\kappa\to\infty}{\to} c_2 n^{-\frac{1}{2}}, \quad \log\left(|\mathcal{N}_n^\kappa|\right) \leq c_3 \tau_n^{\rho} \overset{\kappa\to\infty}{\to} c_3.$$

This is not possible since $|\{G^*_{\mathbb{Q}} \mid \mathbb{Q} \in \mathfrak{Q}_1\}| = \infty$. Intuitively, condition (ii) from Proposition 2.1 requires that $f_{\mathbb{Q}}$ does not increase slower than $x^{\kappa-1} = x^\beta$ when $f_{\mathbb{Q}}$ is close to $\frac{1}{2}$. On the other hand (4.) from Definition 3.4 states that $f_{\mathbb{Q}}$ does not increase faster than $x^\beta$ close to the boundary of $G^*_{\mathbb{Q}}$. Thus if both hold, we require $\beta$ to be the "correct smoothness parameter" in some sense. We refer to Appendix A.1 for a more detailed explanation and comparison of (4.) and (ii).

### *3.3. Main theorems*

We begin by stating the central result of this article. We then use this result to show consistency results for more specific cases. The rates we obtain in Theorem 3.5 are optimal up to a log factor. In the following, all proofs of this section are given in Appendix C.

**Theorem 3.5.** *Let* $\beta \geq 0$*,* $B, \rho > 0$ *and* $d \in \mathbb{N}$ *with* $d \geq 2$*. Let* $\mathcal{F}$ *be a set of functions*

$$\gamma : [0,1]^{d-1} \to \mathbb{R}$$

*such that the following holds. There exist* $\epsilon_0, C_1, C_2 > 0$ *and* $C_3, C_4 \in \mathbb{N}$ *such that for any* $\gamma \in \mathcal{F}$ *and any* $\epsilon \in (0, \epsilon_0)$ *there is a neural network* $\Phi$ *with* $L \leq$

*$L_0(\epsilon) := C_1\lceil \log(\epsilon^{-1})\rceil$ layers, sparsity $s \leq s_0(\epsilon) := C_2\epsilon^{-\rho}\log(\epsilon^{-1})$ and weights in $\mathcal{W}_c$ with $c = c_0(\epsilon) := C_3 + C_4\lceil \log(\epsilon^{-1})\rceil$ such that*

$$\|R(\Phi) - \gamma\|_\infty \leq \epsilon.$$

*Define $\kappa := 1+\beta$ and let $\mathfrak{Q}$ be a class of potential joint distributions $\mathbb{Q}$ of $(X, Y)$ such that the following conditions hold.*

*(a) There is a constant $M > 1$ such that for all $\mathbb{Q} \in \mathfrak{Q}$ the marginal distribution of $\mathbb{Q}_X$ has a Lebesgue density bounded by $M$.*

*(b) There are constants $r \in \mathbb{N}$ and $\epsilon_1, \epsilon_2 > 0$ such that for all $\mathbb{Q} \in \mathfrak{Q}$ the Bayes rule satisfies $G^*_{\mathbb{Q}} \in \mathcal{K}^{\mathcal{F}}_{\mathbb{Q},\beta,B,\epsilon_1,\epsilon_2,r,d}$.*

*(c) We have*

$$d_{f_{\mathbb{Q}}}(G, G^*_{\mathbb{Q}}) \geq c_1 d^\kappa_\Delta(G, G^*_{\mathbb{Q}})$$

*for some constant $c_1 > 0$, all $G \in \mathcal{N}$ and all $\mathbb{Q} \in \mathfrak{Q}$, where*

$$\mathcal{N} := \bigcup_{L,s,c} \mathcal{N}_{L,s,c}$$

*is the class of sets corresponding to any neural network.*

*Let $\tau_n := \frac{n^{\frac{1}{\rho+2\kappa-1}}}{\log^{\frac{2}{\rho}}(n)}$. Then there exist constants $C_1', C_2' > 0$ and $C_3' \in \mathbb{N}$ such that for all $p \geq 1$ we have*

$$\limsup_{n\to\infty} \sup_{\mathbb{Q}\in\mathfrak{Q}} \tau_n^{p\kappa}\, \mathbb{E}\big[d^p_{f_{\mathbb{Q}}}(\widehat{G}_n, G^*_{\mathbb{Q}})\big] < \infty,$$

$$\limsup_{n\to\infty} \sup_{\mathbb{Q}\in\mathfrak{Q}} \tau_n^{p}\, \mathbb{E}\big[d^p_{\Delta}(\widehat{G}_n, G^*_{\mathbb{Q}})\big] < \infty,$$

*where $\widehat{G}_n := \operatorname{arg\,min}_{G\in\mathcal{N}_n} R_n(G)$ with $\mathcal{N}_n = \mathcal{N}_{C_1' L_0(\tau_n^{-1}), C_2' s_0(\tau_n^{-1}), C_3' c_0(\tau_n^{-1})}$.*

**Remark 3.6.** *In Theorem 3.5 we assumed that $\kappa$ coming from condition (ii) in Proposition 2.1 and $\beta$ coming from Definition 3.4 satisfy*

$$\kappa = 1 + \beta. \tag{3.1}$$

*With this assumption, we obtain the rates we desire. The question arises what happens if (3.1) does not hold? It is simple to show that $\kappa \leq 1+\beta$ can not hold. For $\kappa > 1 + \beta$ we note that the condition using $\beta$ is only used to prove (iii) in Proposition 2.1 (see Theorem C.3 in the Appendix). Regarding the proofs of Theorem C.3 and Theorem 3.5 we observe that we obtain the same result as in Theorem 3.5 by replacing $\tau_n$ with $\tau_n' = \tau_n^{\frac{1+\beta}{\kappa}}$. We do not claim optimality for this rate.*

### *3.4. Results for regular boundaries*

We can now prove results for specific classes of sets $\mathcal{F}$ to obtain convergence results. A first important example is the class $\mathcal{F}_{\beta,B,d}$. The following Lemma is a consequence of Theorem 5 in [22].

**Lemma 3.7.** *Let $\beta, B > 0$ and $d \in \mathbb{N}$. Then there exist $\epsilon_0, c_1, c_2 > 0, c_3, c_4 \in \mathbb{N}$ such that the following holds. For any function $\gamma \in \mathcal{F}_{\beta,B,d}$ and any $\epsilon \in (0, \epsilon_0)$, there exists a neural network $\Phi$ with $L \leq L_0(\epsilon) := c_1\lceil \log(\epsilon^{-1})\rceil$ layers, sparsity $s \leq s_0(\epsilon) := c_2\epsilon^{-\frac{d}{\beta}} \log(\epsilon^{-1})$ and weights in $\mathcal{W}_c$ with $c = c_0(\epsilon) := c_3 + c_4\lceil \log(\epsilon^{-1})\rceil$ such that*

$$\|R(\Phi) - \gamma\|_\infty \leq \epsilon.$$

**Corollary 3.8.** *Let $\beta_1 \geq 0$, $B_1, \beta_2, B_2 > 0$ and $d \in \mathbb{N}$ with $d \geq 2$. Let $\mathfrak{Q}$ be a class of potential joint distributions $\mathbb{Q}$ of $(X, Y)$. Assume (a), (b), (c) from Theorem 3.5 hold with $\rho := \frac{d-1}{\beta_2}$, $\beta := \beta_1$, $B := B_1$ as well as $\mathcal{F} := \mathcal{F}_{\beta_2,B_2,d-1}$. Let*

$$\tau_n := \frac{n^{\frac{1}{\rho+2\kappa-1}}}{\log^{\frac{2}{\rho}}(n)}.$$

*Then there exist constants $C_1', C_2' > 0$ and $C_3' \in \mathbb{N}$ such that for all $p \geq 1$ we have*

$$\lim_{n\to\infty} \sup_{\mathbb{Q}\in\mathfrak{Q}} \tau_n^{p\kappa}\, \mathbb{E}\big[d^p_{f_\mathbb{Q}}(\widehat{G}_n, G^*_\mathbb{Q})\big] < \infty,$$
$$\lim_{n\to\infty} \sup_{\mathbb{Q}\in\mathfrak{Q}} \tau_n^{p}\, \mathbb{E}\big[d^p_{\Delta}(\widehat{G}_n, G^*_\mathbb{Q})\big] < \infty,$$

*where $\widehat{G}_n := \operatorname*{arg\,min}_{G\in\mathcal{N}_n} R_n(G)$ with $\mathcal{N}_n = \mathcal{N}_{C_1' L_0(\tau_n^{-1}), C_2' s_0(\tau_n^{-1}), C_3' c_0(\tau_n^{-1})}$ and $L_0, s_0, c_0$ from Lemma 3.7.*

Corollary 3.8 together with Theorem 4.1 from the next chapter prove optimal convergence rates up to a log factor. Following up, note that the rates we receive from Corollary 3.8 are affected by the curse of dimensionality. Observe that the rates obtained by Theorem 3.5 are influenced by condition (c) on the one hand and the ability of neural networks to approximate sets in $\mathcal{F}$ on the other. The dependence on the dimension $d$ in Corollary 3.8 comes from the latter. Thus a natural approach to circumvent the curse of dimensionality is to approximate a smaller set $\mathcal{F}$. Intuitively it is clear that without strong restrictions on the distribution we can only overcome the curse if the complexity of the borders of the sets we approximate is small enough so that they can overcome the curse. In the literature, many different sets are considered which infer useful approximation capabilities of neural networks. Here, we use a class of sets introduced by [22].

**Definition 3.9.** *Let $r \in \mathbb{N}$, $t \in \mathbb{N}^r$, $d \in \mathbb{N}^{r+1}$, $\beta \in \mathbb{R}^r$ and $B > 0$ with $t_i \leq d_i$, $\beta_i > 0$ for $i = 1, \ldots, r$, $d_{r+1} = 1$. Define*

$$\mathcal{G}_{r,t,\beta,B,d} := \big\{\gamma = \gamma_r \circ \cdots \circ \gamma_1 \;\big|\; \gamma_i = (\gamma_{ij} \circ \iota_{ij})_{j=1}^{d_{i+1}},\ \gamma_{ij} \in \mathcal{F}_{\beta_i,B,t_i},\ \iota_{ij} \in \mathcal{ID}_i,$$
$$\gamma_{ij} : [0,1]^{t_i} \to [0,1] \text{ for } i = 1, \ldots, r-1,$$
$$\gamma_{r1} : [0,1]^{t_r} \to \mathbb{R}\big\},$$

*where*

$$\mathcal{ID}_i = \left\{\iota : [0,1]^{d_i} \to [0,1]^{t_i} \;\middle|\; \iota(x) = (x_{i_1}, \dots, x_{i_{t_i}}),\ i_j \in \{1, \dots, d_i\}\right\}.$$

Instead of requiring that $\gamma_{ij}$ is supported on $[0,1]^{t_i}$, we could have used the condition that $\gamma_{ij}$ is supported on $\prod_{k=1}^{t_i}[a_k, b_k]$ for some values $a_k, b_k \in \mathbb{R}$. However, this does not enlarge the class considerably. It can easily be seen that we can instead increase the bound $B$ to find an even larger class. The idea for using the set $\mathcal{G}_{r,t,\beta,B,d}$ is that its complexity does not depend on the input dimension $d_1$, but only on the most difficult component to approximate. The complexity of the components depends on their effective dimension $t_i$ and their implied smoothness. As described by [22], the correct smoothness parameter to consider is

$$\beta_i^* := \beta_i \prod_{k=i+1}^{r} \min\{\beta_k, 1\}.$$

Examples for sets that can profit from Definition 3.9 are additive models ($r = 1$, $t_1 = 1$), interaction models of order $k$ ($r = 1$, $t_1 = k$), and multiplicative models (they are a subset of $\mathcal{G}_{r,t,\beta,B,d}$ when $r = \log_2(d) + 1$, $t_i = 2$ for all $i$). Next, our goal is to establish a convergence result when the set of boundary functions is $\mathcal{G}_{r,t,\beta,B,d}$. Similarly to the approach above, we first provide a lemma that provides approximation results using neural networks.

**Lemma 3.10.** *Let $r \in \mathbb{N}$, $t \in \mathbb{N}^r$, $d \in \mathbb{N}^{r+1}$, $\beta \in \mathbb{R}^r$, and $B > 0$ with $t_i \le d_i$, $\beta_i > 0$ for $i = 1, \dots, r$, $d_{r+1} = 1$. Let $\mathcal{G}_{r,t,\beta,B,d}$ be defined as in Definition 3.9 and define*

$$\rho := \max_{i=1,\dots,r} \left(\frac{t_i}{\beta_i^*}\right).$$

*Then there exist $\epsilon_0, c_1, c_2 > 0, c_3, c_4 \in \mathbb{N}$ such that the following holds. For any function $\gamma \in \mathcal{G}_{r,t,\beta,B,d}$ and any $\epsilon \in (0, \epsilon_0)$, there exists a neural network $\Phi$ with $L \le L_0(\epsilon) := c_1\lceil \log(\epsilon^{-1})\rceil$ layers, sparsity $s \le s_0(\epsilon) := c_2\epsilon^{-\rho}\log(\epsilon^{-1})$ and weights in $\mathcal{W}_c$ with $c = c_0(\epsilon) := c_3 + c_4\lceil \log(\epsilon^{-1})\rceil$ such that*

$$\|R(\Phi) - \gamma\|_\infty \le \epsilon.$$

The following corollary establishes the corresponding convergence result. Theorem 4.2 provides the lower bound in the case where $t_i \le \min\{d_1, \dots, d_i\}$.

**Corollary 3.11.** *Let $r_2 \in \mathbb{N}$, $t \in \mathbb{N}^{r_2}$, $d \in \mathbb{N}^{r_2+1}$, $\beta_1 \ge 0$, $\beta_2 \in \mathbb{R}^{r_2}$, and $B_1, B_2 > 0$ with $\beta_{2,i} > 0$ for $i = 1, \dots, r_2$, $d_{r_2+1} = 1$. Additionally, $t_1 < d_1$ and $t_i \le d_i$ for $i \ne 1$. Define $d' \in \mathbb{N}^{r_2+1}$ with $d_1' = d_1 - 1$, $d_i' = d_i$ for $i \ne 1$ and*

$$\rho := \max_{i=1,\dots,r_2} \frac{t_i}{\beta_{2,i}^*}$$

*Let $\kappa := 1 + \beta_1$ and $\mathfrak{Q}$ be a class of potential joint distributions $\mathbb{Q}$ of $(X, Y)$ such that the following conditions hold.*

(a) *There is a constant $M > 1$ such that for all $\mathbb{Q} \in \mathfrak{Q}$ the marginal distribution of $\mathbb{Q}_X$ has a Lebesgue density bounded by $M$.*

(b) *There are constants $r_1 \in \mathbb{N}$ and $\epsilon_1, \epsilon_2 > 0$ such that for all $\mathbb{Q} \in \mathfrak{Q}$ the Bayes rule satisfies $G^*_{\mathbb{Q}} \in \mathcal{K}^{\mathcal{F}}_{\mathbb{Q},\beta_1,B_1,\epsilon_1,\epsilon_2,r_1,d_1}$ with*

$$\mathcal{F} := \mathcal{G}_{r_2,t,\beta_2,B_2,d'}.$$

(c) *We have*

$$d_{f_{\mathbb{Q}}}(G, G^*_{\mathbb{Q}}) \geq c_1 d^{\kappa}_{\Delta}(G, G^*_{\mathbb{Q}})$$

*for some constant $c_1 > 0$, all $G \in \mathcal{N}$ and all $\mathbb{Q} \in \mathfrak{Q}$, where*

$$\mathcal{N} := \bigcup_{L,s,c} \mathcal{N}_{L,s,c}$$

*is the class of sets corresponding to any neural network.*

*Let $\tau_n := \frac{n^{\frac{1}{\rho+2\kappa-1}}}{\log^{\frac{2}{\rho}}(n)}$. Then there exist constants $C_1', C_2' > 0$ and $C_3' \in \mathbb{N}$ such that for all $p \geq 1$ we have*

$$\limsup_{n\to\infty} \sup_{\mathbb{Q}\in\mathfrak{Q}} \tau_n^{p\kappa}\, \mathbb{E}\big[d^p_{f_{\mathbb{Q}}}(\widehat{G}_n, G^*_{\mathbb{Q}})\big] < \infty,$$

$$\limsup_{n\to\infty} \sup_{\mathbb{Q}\in\mathfrak{Q}} \tau_n^{p}\, \mathbb{E}\big[d^p_{\Delta}(\widehat{G}_n, G^*_{\mathbb{Q}})\big] < \infty,$$

*where $\widehat{G}_n := \operatorname{arg\,min}_{G\in\mathcal{N}_n} R_n(G)$ with $\mathcal{N}_n = \mathcal{N}_{C_1' L_0(\tau_n^{-1}), C_2' s_0(\tau_n^{-1}), C_3' c_0(\tau_n^{-1})}$ and $L_0, s_0, c_0$ from Lemma 3.10.*

Note that Corollary 3.11 is a generalization of Corollary 3.8. The rate now depends on $\rho$ which in turn depends on $t_1, \ldots, t_{r_2}$ instead of the input dimension $d_1$. Clearly, the effective dimensions $t_i$ can be much smaller than the input dimension $d_1$, for example, when the boundary functions are additive.

## 4. Lower bounds

We now establish lower bounds on the convergence rates from corollaries 3.8 and 3.11. Note that the lower bounds also prove that the rates obtained in Theorem 3.5 can not be improved up to a log factor. Since Corollary 3.11 is a generalization of 3.8, we only have to prove a lower bound for the setting given in the former. For clarity, we provide both statements. The proofs of this section can be found in Appendix D.

**Theorem 4.1.** *Let $\beta_1 \geq 0$, $B_1, \beta_2, B_2, \rho > 0$, and $d \in \mathbb{N}$ with $d \geq 2$. Let $\mathfrak{Q}$ be the class of all potential joint distributions $\mathbb{Q}$ of $(X, Y)$ such that (a), (b) from Theorem 3.5 hold with $\rho := \frac{d-1}{\beta_2}$, $\kappa := 1+\beta_1$, $\beta := \beta_1$, $B := B_1$, $\mathcal{F} := \mathcal{F}_{\beta_2,B_2,d-1}$, and some $M > 1$, $r \in \mathbb{N}$, $\epsilon_1, \epsilon_2 > 0$. Let (c) hold with $c_1 > 0$ large enough and set $\tau_n := n^{\frac{1}{\rho+2\kappa-1}}$. Then*

$$\liminf_{n\to\infty} \inf_{G_n\in\mathfrak{G}} \sup_{\mathbb{Q}\in\mathfrak{Q}} \tau_n^p \mathbb{E}\left[d^p_{\Delta}(G_n, G^*_{\mathbb{Q}})\right] > 0,$$

$$\liminf_{n\to\infty} \inf_{G_n\in\mathfrak{G}} \sup_{\mathbb{Q}\in\mathfrak{Q}} \tau_n^{p\kappa}\mathbb{E}\left[d_{f_{\mathbb{Q}}}^p(G_n, G_{\mathbb{Q}}^*)\right] > 0$$

*for every $p \geq 0$, where $\mathfrak{G}$ contains all estimators depending on the data.*

Intuitively $B_1$ bounds the factor of the $x^{\beta_2}$-term of $f_{\mathbb{Q}}$ close to the boundary from above. On the other hand, $c_1$ bounds this term from below. Thus not every combination of $B_1, c_1 > 0$ is possible. We prove Theorem 4.1 for large $c_1 > 0$. We do not provide the exact ratio of $B$ and $c_1$ required since it is not important for the statement. Lastly, the lower bound corresponding to Corollary 3.11 is given.

**Theorem 4.2.** *Let $r_2 \in \mathbb{N}$, $t \in \mathbb{N}^{r_2}$, $d \in \mathbb{N}^{r_2+1}$, $\beta_1 \geq 0$, $\beta_2 \in \mathbb{R}^{r_2}$, and $B_1, B_2 > 0$ with $\beta_{2,i} > 0$ for $i = 1, \ldots, r_2$, $d_{r_2+1} = 1$. Additionally, $t_1 < d_1$ and $t_i \leq \min\{d_1, \ldots, d_i\}$ for $i \neq 1$. Let*

$$\rho := \max_{i=1,\ldots,r_2} \frac{t_i}{\beta_{2,i}^*}$$

*Define $\kappa := 1 + \beta_1$ and let $\mathfrak{Q}$ be the class of all potential joint distributions $\mathbb{Q}$ of $(X, Y)$ such that (a), (b) from Corollary 3.11 hold for some $M > 1$, $r \in \mathbb{N}$, $\epsilon_1, \epsilon_2 > 0$. Let (c) hold with $c_1 > 0$ large enough and set $\tau_n := n^{\frac{1}{\rho+2\kappa-1}}$. Then*

$$\liminf_{n\to\infty} \inf_{G_n\in\mathfrak{G}} \sup_{\mathbb{Q}\in\mathfrak{Q}} \tau_n^{p}\mathbb{E}\left[d_{\Delta}^p(G_n, G_{\mathbb{Q}}^*)\right] > 0,$$

$$\liminf_{n\to\infty} \inf_{G_n\in\mathfrak{G}} \sup_{\mathbb{Q}\in\mathfrak{Q}} \tau_n^{p\kappa}\mathbb{E}\left[d_{f_{\mathbb{Q}}}^p(G_n, G_{\mathbb{Q}}^*)\right] > 0$$

*for every $p \geq 0$, where $\mathfrak{G}$ contains all estimators depending on the data.*

## 5. Concluding remarks

We establish optimal convergence rates up to a log factor in a classification setting under (1.1) using neural networks. Theorem 3.5 can be applied for many different boundary functions. The complexity of the class of boundary functions $\mathcal{F}$ is one of the main driving factors of the convergence rate. In particular many approaches which circumvent the curse of dimensionality in a regression setting can be used to circumvent the curse in this classification setting.

Note that this paper is of a theoretical nature. While sparsity constraints are considered thoroughly in the theoretical literature, they are not widely used in practice. Additionally, we did not discuss the minimization required for the calculation of $\widehat{G}_n$. This is a very interesting but complicated topic which is not in the scope of this article. Observe that the class of neural networks used in Theorem 3.5 depends on $\kappa$ as well as $\rho$. We believe that one can extend the results of this paper by either having adaptive classes of neural networks or a class independent of $\kappa$ and $\rho$ in a similar manner to [24]. One obstacle to overcome is the fact that the conditions on the probability distribution $\mathfrak{Q}$ required are not strictly weaker for larger $\kappa$ and $\rho$.

## Appendix A: Explanation of Definition 3.4

### *A.1. Comparison of (ii) and (4.)*

As discussed above (ii) states that $f_{\mathbb{Q}}$ does not increase slower than $x^{\kappa-1} = x^{\beta}$ when $f_{\mathbb{Q}}$ is close to $\frac{1}{2}$ while (4.) states that $f_{\mathbb{Q}}$ does not increase faster than $x^{\beta}$ close to the boundary of $G^*_{\mathbb{Q}}$. Note that the regions in which the requirements are important are slightly different. $f_{\mathbb{Q}}$ may be close (or even equal) to $\frac{1}{2}$ far away from the boundary such that (ii) is relevant. On the other hand (4.) implies that $f_{\mathbb{Q}}$ is similar to $\frac{1}{2}$ near the boundary of $G^*_{\mathbb{Q}}$, such that both are relevant near the boundary. Note that even functions $f_{\mathbb{Q}}$ that are highly discontinuous near the boundary of $G^*_{\mathbb{Q}}$ may satisfy both (ii) and (4.). This is essential in order to obtain the minimax rates we desire, which can be seen in the construction of functions $f_{\mathbb{Q}}$ in the proofs of Theorems 4.1 and Theorem 4.2.

**Remark A.1.** *We provide two examples of naive requirements which do not satisfy the properties we desire. First of all, we do not assume $f_{\mathbb{Q}}$ to be of order $x^{\beta}$ close to the boundary of $G^*_{\mathbb{Q}}$. This leads to a different problem as can be seen in Theorem 3.3 in [11]. Secondly, looking at the proofs we would wish to use the following condition to counteract (ii): There exists a constant $\kappa \geq 1$ such that*

$$d_{f_{\mathbb{Q}}}(G, G^*_{\mathbb{Q}}) \leq c_1 d^{\kappa}_{\Delta}(G, G^*_{\mathbb{Q}}) \tag{A.1}$$

*for some constant $c_1 > 0$ and all relevant $G$. However, considering for example all $G \in \bigcup_{L,s,c} \mathcal{N}_{L,s,c}$ one can show that this is not satisfied for almost any probability distribution $\mathbb{Q}$. Thus, the assumption is too restrictive and one can obtain better rates in this case. In essence (4.) implies (A.1) for all relevant sets $G$.*

### *A.2. Comparison to [11]*

In this section, we shortly compare Definition 3.4 in the case $\mathcal{F} = \mathcal{F}_{\beta,B,d-1}$ to the definitions in [11]. They differ in three aspects. (1) In [11] the possible Bayes rules are of the form

$$G = \sum_{t=1}^{T} A_t, \quad A_t = \bigcap_{k=1}^{K} \{x_{j_{k,t}} \geq g_{k,t}(x_{-j_{k,t}})\} \tag{A.2}$$

for functions $g_{k,t} \in \mathcal{F}_{\beta,B,d-1}$. This is a generalization of boundary fragments and is a somewhat global approach to defining sets compared to our local approach. For example, any set containing a hole is not of the form (A.2) for any $K, T \in \mathbb{N}$. Figure 2 provides a more broad example for a set that is not of the form (A.2) and is considered in $\mathcal{K}_{\mathbb{Q},\beta,B,\epsilon_1,\epsilon_2,r,d}$ for come $r \in \mathbb{N}$. (2) Condition (3.5) in [11] puts a requirement on all $x \in [0,1]^d$ where (4.) in Definition 3.4 only considers values close to the boundary of $G$. (3) Condition (R) as well as the condition in (3.5) in [11] depend on the precise choice of $g_{k,t}$ in (A.2) including points which

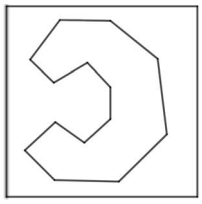

FIG 2. *Example of a set that is not of the form (A.2) for any $K, T \in \mathbb{N}$ but considered in $\mathcal{K}_{\mathbb{Q},\beta,B,\epsilon_1,\epsilon_2,r,d}$ when $F = \mathcal{F}_{\beta,B,d-1}$ and $r$ is large enough.*

are not on the boundary of $G$. This can lead to unexpected requirements far away from the boundary of $G$.

We thus argue that Definition 3.4 is more refined and allows for a significantly larger variety of sets. This comes with the cost of additional complications in the proofs since handling the boundary functions locally is more complex than handling them globally.

## Appendix B: General convergence results

The proof of Proposition 2.1 is similar to the proof of Theorem 2 in [18]. For the sake of completion, we provide the entire argumentation here anyway.

*Proof of Proposition 2.1.* Let $n \geq N_0$. Without loss of generality, we may assume that $\tau_n \leq n^{\frac{1}{\rho+2\kappa-1}}$, since otherwise the conditions are also satisfied when using $\bar{\tau}_n = n^{\frac{1}{\rho+2\kappa-1}}$. We begin by proving the assertion for the first term. The idea is to bound

$$\mathbb{P}\big(d_{f_{\mathbb{Q}}}(\widehat{G}_n, G^*_{\mathbb{Q}}) > t\tau_n^{-\kappa}\big)$$

for some $t > 0$. First, observe that for any $G \in \mathcal{N}_n$

$$\begin{aligned}
&R_n(G) - R_n(G^*_{\mathbb{Q}}) - d_{f_{\mathbb{Q}}}(G, G^*_{\mathbb{Q}}) \\
&= \frac{1}{n}\sum_{i=1}^n \big(Y_i - 1(X_i \in G)\big)^2 - \frac{1}{n}\sum_{i=1}^n \big(Y_i - 1(X_i \in G^*_{\mathbb{Q}})\big)^2 \\
&\quad - \Big(\mathbb{E}\Big[\big(Y - 1(X \in G)\big)^2\Big] - \mathbb{E}\Big[\big(Y - 1(X \in G^*_{\mathbb{Q}})\big)^2\Big]\Big) \\
&= \frac{1}{n}\sum_{i=1}^n h_G(X_i, Y_i) - \mathbb{E}\big[h_G(X_i, Y_i)\big] =: \frac{1}{n}\sum_{i=1}^n U_i(G)
\end{aligned}$$

holds, where

$$h_G : \mathbb{R}^d \times \{0,1\} \to \mathbb{R},\ h_G(x,y) = \big(y - 1(x \in G)\big)^2 - \big(y - 1(x \in G^*_{\mathbb{Q}})\big)^2.$$

Regarding (iii), for every $n \in \mathbb{N}$ there exists a $G_n \in \mathcal{N}_n$ such that

$$d_{f_{\mathbb{Q}}}(G_n, G^*_{\mathbb{Q}}) \leq c_2\tau_n^{-\kappa}.$$

For $t > 0$, define

$$\Xi_t := \Big\{ G \in \mathcal{N}_n \;\Big|\; d_{f_{\mathbb{Q}}}(G, G^*_{\mathbb{Q}}) \geq t\tau_n^{-\kappa} \Big\}.$$

Then, for $t \geq 4c_2$ and $G \in \Xi_t$ we have

$$\frac{1}{2} d_{f_{\mathbb{Q}}}(G, G^*_{\mathbb{Q}}) - d_{f_{\mathbb{Q}}}(G_n, G^*_{\mathbb{Q}}) \geq c_2 \tau_n^{-\kappa}. \tag{B.1}$$

Recall that by definition $\widehat{G}_n$ minimizes $R_n(\cdot)$. Therefore, in view of the calculations above, for $t \geq 4c_2$

$$\begin{aligned}
&\mathbb{P}\big(d_{f_{\mathbb{Q}}}(\widehat{G}_n, G^*_{\mathbb{Q}}) > t\tau_n^{-\kappa}\big) \\
&\leq \mathbb{P}\big(\exists G \in \Xi_t :\; R_n(G) - R_n(G_n) \leq 0\big) \\
&= \mathbb{P}\Big(\exists G \in \Xi_t :\; R_n(G) - R_n(G^*_{\mathbb{Q}}) - \big(R_n(G_n) - R_n(G^*_{\mathbb{Q}}\big) \leq 0\Big) \\
&= \mathbb{P}\left(\exists G \in \Xi_t :\; d_{f_{\mathbb{Q}}}(G, G^*_{\mathbb{Q}}) + \frac{1}{n}\sum_{i=1}^n U_i(G) - d_{f_{\mathbb{Q}}}(G_n, G^*_{\mathbb{Q}}) - \frac{1}{n}\sum_{i=1}^n U_i(G_n) \leq 0\right)
\end{aligned}$$

holds. Using inequality (B.1) in the third row yields

$$\begin{aligned}
&\mathbb{P}\left(\exists G \in \Xi_t :\; d_{f_{\mathbb{Q}}}(G, G^*_{\mathbb{Q}}) + \frac{1}{n}\sum_{i=1}^n U_i(G) - d_{f_{\mathbb{Q}}}(G_n, G^*_{\mathbb{Q}}) - \frac{1}{n}\sum_{i=1}^n U_i(G_n) \leq 0\right) \\
&= \mathbb{P}\left(\exists G \in \Xi_t :\; \left(\frac{1}{2} d_{f_{\mathbb{Q}}}(G, G^*_{\mathbb{Q}}) + \frac{1}{n}\sum_{i=1}^n U_i(G)\right)\right. \\
&\qquad\qquad \left. + \left(\frac{1}{2} d_{f_{\mathbb{Q}}}(G, G^*_{\mathbb{Q}}) - d_{f_{\mathbb{Q}}}(G_n, G^*_{\mathbb{Q}}) - \frac{1}{n}\sum_{i=1}^n U_i(G_n)\right) \leq 0\right) \\
&\leq \mathbb{P}\left(\exists G \in \Xi_t :\; \frac{1}{2} d_{f_{\mathbb{Q}}}(G, G^*_{\mathbb{Q}}) + \frac{1}{n}\sum_{i=1}^n U_i(G) \leq 0\right) \\
&\quad + \mathbb{P}\left(c_2\tau_n^{-\kappa} - \frac{1}{n}\sum_{i=1}^n U_i(G_n) \leq 0\right) \\
&\leq \mathbb{P}\left(\exists G \in \Xi_t :\; \frac{1}{n}\sum_{i=1}^n U_i(G) \leq -\frac{1}{2} d_{f_{\mathbb{Q}}}(G, G^*_{\mathbb{Q}})\right) + \mathbb{P}\left(c_2\tau_n^{-\kappa} \leq \frac{1}{n}\sum_{i=1}^n U_i(G_n)\right)
\end{aligned}$$

and thus

$$\begin{aligned}
&\mathbb{P}\big(d_{f_{\mathbb{Q}}}(\widehat{G}_n, G^*_{\mathbb{Q}}) > t\tau_n^{-\kappa}\big) \\
&\leq \mathbb{P}\left(\exists G \in \Xi_t :\; \frac{1}{n}\sum_{i=1}^n U_i(G) \leq -\frac{1}{2} d_{f_{\mathbb{Q}}}(G, G^*_{\mathbb{Q}})\right) + \mathbb{P}\left(c_2\tau_n^{-\kappa} \leq \frac{1}{n}\sum_{i=1}^n U_i(G_n)\right).
\end{aligned}$$

It remains to find upper bounds for the two terms above. In order to bound the first term, note that for $(x,y) \in \mathbb{R}^d \times \{0,1\}$ and any $G \in \mathcal{N}_n$ we have

$$\begin{aligned} |h_G(x,y)| &= \begin{cases} \big|1 - 1(x \in G) - \big(1 - 1(x \in G^*_{\mathbb{Q}})\big)\big|, & \text{for } y = 1, \\ \big|1(x \in G) - 1(x \in G^*_{\mathbb{Q}})\big|, & \text{for } y = 0 \end{cases} \\ &= 1\big(x \in G \Delta G^*_{\mathbb{Q}}\big). \end{aligned}$$

For all $i = 1, \dots, n$ this implies $|U_i(G)| \le 2$ and

$$\begin{aligned} \mathbb{E}\big[U_i(G)^2\big] &\le \mathbb{E}\big[h_G(X_i, Y_i)^2\big] = \mathbb{E}\big[1\big(x \in G \Delta G^*_{\mathbb{Q}}\big)\big] \\ &= d_\Delta(G, G^*_{\mathbb{Q}}) \le c_1^{-\frac{1}{\kappa}} d_{f_{\mathbb{Q}}}(G, G^*_{\mathbb{Q}})^{\frac{1}{\kappa}} \end{aligned}$$

where the last inequality follows from (ii). By Bernstein's inequality, for all $a > 0$

$$\mathbb{P}\left( \left| \frac{1}{n} \sum_{i=1}^n U_i(G) \right| \ge a \right) \le 2 \exp\left( - \frac{k_1 n a^2}{a + c_1^{-\frac{1}{\kappa}} d_{f_{\mathbb{Q}}}(G, G^*_{\mathbb{Q}})^{\frac{1}{\kappa}}} \right)$$

holds, where $k_1 > 0$ is a constant. By setting $a = \frac{1}{2} d_{f_{\mathbb{Q}}}(G, G^*_{\mathbb{Q}})$ and observing that $d_{f_{\mathbb{Q}}}(G, G^*_{\mathbb{Q}}) \le 1$, we have

$$\mathbb{P}\left( \left| \frac{1}{n} \sum_{i=1}^n U_i(G) \right| \ge \frac{1}{2} d_{f_{\mathbb{Q}}}(G, G^*_{\mathbb{Q}}) \right) \le 2 \exp\Big( - k_2 n d_{f,g}(G, G^*_{\mathbb{Q}})^{\frac{2\kappa - 1}{\kappa}} \Big)$$

for some constant $k_2 > 0$. Noting that by definition $\tau_n \le n^{\frac{1}{\rho + 2\kappa - 1}}$ and $\kappa \ge 1$, by (iv) we have

$$\begin{aligned} &\mathbb{P}\left( \exists G \in \Xi_t : \ \left| \frac{1}{n} \sum_{i=1}^n U_i(G) \right| \ge \frac{1}{2} d_{f_{\mathbb{Q}}}(G, G^*_{\mathbb{Q}}) \right) \\ &\quad \le 2 \exp\big(c_3 n^{\frac{\rho}{\rho + 2\kappa - 1}}\big) \exp\big( - k_2 n t^{\frac{2\kappa - 1}{\kappa}} \tau_n^{1 - 2\kappa}\big) \\ &\quad \le 2 \exp\big(c_3 n^{\frac{\rho}{\rho + 2\kappa - 1}}\big) \exp\big( - k_2 t^{\frac{2\kappa - 1}{\kappa}} n^{\frac{1 - 2\kappa}{\rho + 2\kappa - 1} + 1}\big) \\ &\quad \le 2 \exp\Big( \big(c_3 - k_2 t^{\frac{2\kappa - 1}{\kappa}}\big) n^{\frac{\rho}{\rho + 2\kappa - 1}} \Big) \\ &\quad \le 2 \exp\Big( - c_3 \tau_n^\rho \Big) \end{aligned}$$

for all $t \ge \left(\frac{2c_3}{k_2}\right)^{\frac{\kappa}{2\kappa - 1}}$. To bound the second term we use Bernstein's inequality with $a = c_2 \tau_n^{-\kappa}$ and receive

$$\begin{aligned} \mathbb{P}\left( c_2 \tau_n^{-\kappa} \le \frac{1}{n} \sum_{i=1}^n U_i(G_n) \right) &\le \exp\left( - \frac{k_1 n c_2^2 \tau_n^{-2\kappa}}{c_2 \tau_n^{-\kappa} + c_1^{-\frac{1}{\kappa}} d_{f_{\mathbb{Q}}}(G^*_{\mathbb{Q}}, G_n)^{\frac{1}{\kappa}}} \right) \\ &\le \exp\big( - k_3 n \tau_n^{-2\kappa + 1}\big) \\ &\le \exp\big( - k_3 \tau_n^\rho\big) \end{aligned}$$

for some constant $k_3 > 0$. Therefore, for $t \geq \max\left\{4c_2, \left(\frac{2c_3}{k_2}\right)^{\frac{\kappa}{2\kappa-1}}\right\}$ we find an upper bound

$$\begin{aligned}
&\mathbb{P}\big(d_{f_{\mathbb{Q}}}(\widehat{G}_n, G^*_{\mathbb{Q}}) > t\tau_n^{-\kappa}\big) \\
&\leq \mathbb{P}\left(\exists G \in \Xi_t : \ \frac{1}{n}\sum_{i=1}^n U_i(G) \leq -\frac{1}{2} d_{f_{\mathbb{Q}}}(G, G^*_{\mathbb{Q}})\right) + \mathbb{P}\left(c_2\tau_n^{-\kappa} \leq \frac{1}{n}\sum_{i=1}^n U_i(G_n)\right) \\
&\leq 2\exp\Big(-c_3\tau_n^{\rho}\Big) + \exp\big(-k_3\tau_n^{\rho}\big).
\end{aligned}$$

Observing that $d_{f_{\mathbb{Q}}}(\widehat{G}_n, G^*_{\mathbb{Q}}) \leq 1$ we conclude

$$\begin{aligned}
&\mathbb{E}\big[d^p_{f_{\mathbb{Q}}}(\widehat{G}_n, G^*_{\mathbb{Q}})\big] \\
&\leq \mathbb{E}\big[1\big(d_{f_{\mathbb{Q}}}(\widehat{G}_n, G^*_{\mathbb{Q}}) > t\tau_n^{-\kappa}\big)\big] + t\tau_n^{-p\kappa}\mathbb{E}\big[1\big(d_{f_{\mathbb{Q}}}(\widehat{G}_n, G^*_{\mathbb{Q}}) \leq t\tau_n^{-\kappa}\big)\big] \\
&\leq 2\exp\big(-c_3\tau_n^{\rho}\big) + \exp\big(-k_3\tau_n^{\rho}\big) + t\tau_n^{-\kappa p}
\end{aligned}$$

and thus

$$\begin{aligned}
&\limsup_{n\to\infty} \sup_{\mathbb{Q}\in\mathfrak{Q}} \tau_n^{\kappa p}\mathbb{E}\big[d^p_{f_{\mathbb{Q}}}(\widehat{G}_n, G^*_{\mathbb{Q}})\big] \\
&\leq \limsup_{n\to\infty} \tau_n^{\kappa p}\Big(2\exp\big(-c_3\tau_n^{\rho}\big) + \exp\big(-k_3\tau_n^{\rho}\big) + t\tau_n^{-\kappa p}\Big) \\
&< \infty.
\end{aligned}$$

Proving that the second term in the assertion is finite follows directly, since regarding (ii) for all $\mathbb{Q} \in \mathfrak{Q}$ and sets $G \in \mathcal{N}_n$ it holds hat

$$d_{f_{\mathbb{Q}}}(G, G^*_{\mathbb{Q}}) \geq c_1 d^{\kappa}_{\Delta}(G, G^*_{\mathbb{Q}}). \qquad \square$$

### *B.1. Result with absence of Tsybakovs noise condition*

**Proposition B.1.** *Let $\tau_n > 0$ be a monotonically increasing sequence. Let $\mathfrak{Q}$ be a class of potential joint distributions $\mathbb{Q}$ of $(X, Y)$ and $\mathcal{N}_n$ be a collection of subsets of $\mathbb{R}^d$ for all $n \in \mathbb{N}$ such that the following conditions hold.*

*(i) For all $\mathbb{Q} \in \mathfrak{Q}$ all sets in $\bigcup_{n\in\mathbb{N}} \mathcal{N}_n$ and $G^*_{\mathbb{Q}}$ are $\mathbb{Q}$-measurable.*

*Additionally, we assume that there is a constant $N_0 \in \mathbb{N}$ such that for all $n \geq N_0$ the following holds.*

*(ii) There is a constant $c_2 > 0$ such that for all $n \in \mathbb{N}$ and $\mathbb{Q} \in \mathfrak{Q}$ there is a $G \in \mathcal{N}_n$ with*

$$d_{f_{\mathbb{Q}}}(G, G^*_{\mathbb{Q}}) \leq c_2\tau_n^{-1}.$$

*(iii) There exist $c_3, \rho > 0$ such that*

$$\log\big(|\mathcal{N}_n|\big) \leq c_3 n^{\frac{\rho}{\rho+2}}.$$

*Then for all $p \geq 1$ we have*

$$\limsup_{n\to\infty} \sup_{\mathbb{Q}\in\mathfrak{Q}} \tilde{\tau}_n^p \, \mathbb{E}\big[d_{f_\mathbb{Q}}^p(\widehat{G}_n, G_\mathbb{Q}^*)\big] < \infty,$$

*where $\tilde{\tau}_n := \min\{\tau_n, n^{\frac{1}{\rho+2}}\}$ for all $n \in \mathbb{N}$.*

Proposition B.1 provides rates in the absence of condition (ii) of Proposition 2.1. For $p = 1$ the best rate achievable is of order $n^{-\frac{1}{\rho+2}}$, which is always slower than $n^{-\frac{1}{2}}$. To compare the rates of Propositions 2.1 and B.1 assume we are given $\mathfrak{Q}$ and $\mathcal{N}_n$ satisfying the conditions of Proposition 2.1 with $\tau_n = n^{\frac{1}{\rho+2\kappa-1}}$. The conditions then read

$$d_{f_\mathbb{Q}}(G, G_\mathbb{Q}^*) \leq c_2 \tau_n^{-\kappa}, \quad \log\big(|\mathcal{N}_n|\big) \leq c_3 \tau_n^\rho$$

for some $G \in \mathcal{N}_n$. Thus, the conditions of Proposition B.1 are satisfied for $\mathfrak{Q}$, $\rho' = \frac{\rho}{\kappa}$ with $\tau_n' = n^{\frac{1}{\rho'+2}}$ and some subsequence $\mathcal{N}_{k_n}$. The rate we obtain is

$$n^{-\frac{\kappa}{\rho+2\kappa}} > n^{-\frac{\kappa}{\rho+2\kappa-1}}. \tag{B.2}$$

Thus, we lose something if we don't know the correct value for $\kappa$. Note that we do not claim optimality for the rates obtained by Proposition B.1. However, we note that optimal rates can not be faster than $n^{-\frac{1}{2}}$ for any $\rho > 0$ since the right-hand side of (B.2) goes to $n^{-\frac{1}{2}}$ for $\kappa \to \infty$ which corresponds to $\rho' \to 0$.

*Proof of Proposition B.1.* Let $n \geq N_0$. Without loss of generality, we may assume that $\tau_n \leq n^{\frac{1}{\rho+2}}$, since otherwise the conditions are also satisfied when using $\tau_n = n^{\frac{1}{\rho+2}}$. The idea is to bound

$$\mathbb{P}\big(d_{f_\mathbb{Q}}(\widehat{G}_n, G_\mathbb{Q}^*) > t\tau_n^{-1}\big)$$

for some $t > 0$. First, observe that for any $G \in \mathcal{N}_n$

$$\begin{aligned}
&R_n(G) - R_n(G_\mathbb{Q}^*) - d_{f_\mathbb{Q}}(G, G_\mathbb{Q}^*) \\
&= \frac{1}{n}\sum_{i=1}^n \big(Y_i - 1(X_i \in G)\big)^2 - \frac{1}{n}\sum_{i=1}^n \big(Y_i - 1(X_i \in G_\mathbb{Q}^*)\big)^2 \\
&\quad - \Big(\mathbb{E}\Big[\big(Y - 1(X \in G)\big)^2\Big] - \mathbb{E}\Big[\big(Y - 1(X \in G_\mathbb{Q}^*)\big)^2\Big]\Big) \\
&= \frac{1}{n}\sum_{i=1}^n h_G(X_i, Y_i) - \mathbb{E}\big[h_G(X_i, Y_i)\big] =: \frac{1}{n}\sum_{i=1}^n U_i(G)
\end{aligned}$$

holds, where

$$h_G : \mathbb{R}^d \times \{0,1\} \to \mathbb{R}, \; h_G(x,y) = \big(y - 1(x \in G)\big)^2 - \big(y - 1(x \in G_\mathbb{Q}^*)\big)^2.$$

Regarding (ii), for every $n \in \mathbb{N}$ there exists a $G_n \in \mathcal{N}_n$ such that

$$d_{f_\mathbb{Q}}(G_n, G_\mathbb{Q}^*) \leq c_2 \tau_n^{-1}.$$

For $t > 0$, define

$$\Xi_t := \Big\{ G \in \mathcal{N}_n \;\Big|\; d_{f_{\mathbb{Q}}}(G, G^*_{\mathbb{Q}}) \geq t\tau_n^{-1} \Big\}.$$

Then, for $t \geq 4c_2$ and $G \in \Xi_t$ we have

$$\frac{1}{2} d_{f_{\mathbb{Q}}}(G, G^*_{\mathbb{Q}}) - d_{f_{\mathbb{Q}}}(G_n, G^*_{\mathbb{Q}}) \geq c_2 \tau_n^{-1}. \tag{B.3}$$

Recall that by definition $\widehat{G}_n$ minimizes $R_n(\cdot)$. Therefore, in view of the calculations above, for $t \geq 4c_2$

$$\begin{aligned}
&\mathbb{P}\big(d_{f_{\mathbb{Q}}}(\widehat{G}_n, G^*_{\mathbb{Q}}) > t\tau_n^{-1}\big) \\
&\leq \mathbb{P}\big(\exists G \in \Xi_t :\ R_n(G) - R_n(G_n) \leq 0\big) \\
&= \mathbb{P}\Big(\exists G \in \Xi_t :\ R_n(G) - R_n(G^*_{\mathbb{Q}}) - \big(R_n(G_n) - R_n(G^*_{\mathbb{Q}}\big) \leq 0\Big) \\
&= \mathbb{P}\left(\exists G \in \Xi_t :\ d_{f_{\mathbb{Q}}}(G, G^*_{\mathbb{Q}}) + \frac{1}{n}\sum_{i=1}^n U_i(G) - d_{f_{\mathbb{Q}}}(G_n, G^*_{\mathbb{Q}}) - \frac{1}{n}\sum_{i=1}^n U_i(G_n) \leq 0\right)
\end{aligned}$$

holds. Using inequality (B.3) in the third row yields

$$\begin{aligned}
&\mathbb{P}\left(\exists G \in \Xi_t :\ d_{f_{\mathbb{Q}}}(G, G^*_{\mathbb{Q}}) + \frac{1}{n}\sum_{i=1}^n U_i(G) - d_{f_{\mathbb{Q}}}(G_n, G^*_{\mathbb{Q}}) - \frac{1}{n}\sum_{i=1}^n U_i(G_n) \leq 0\right) \\
&= \mathbb{P}\left(\exists G \in \Xi_t :\ \left(\frac{1}{2} d_{f_{\mathbb{Q}}}(G, G^*_{\mathbb{Q}}) + \frac{1}{n}\sum_{i=1}^n U_i(G)\right)\right. \\
&\qquad\qquad \left. + \left(\frac{1}{2} d_{f_{\mathbb{Q}}}(G, G^*_{\mathbb{Q}}) - d_{f_{\mathbb{Q}}}(G_n, G^*_{\mathbb{Q}}) - \frac{1}{n}\sum_{i=1}^n U_i(G_n)\right) \leq 0\right) \\
&\leq \mathbb{P}\left(\exists G \in \Xi_t :\ \frac{1}{2} d_{f_{\mathbb{Q}}}(G, G^*_{\mathbb{Q}}) + \frac{1}{n}\sum_{i=1}^n U_i(G) \leq 0\right) \\
&\quad + \mathbb{P}\left(c_2\tau_n^{-1} - \frac{1}{n}\sum_{i=1}^n U_i(G_n) \leq 0\right) \\
&\leq \mathbb{P}\left(\exists G \in \Xi_t :\ \frac{1}{n}\sum_{i=1}^n U_i(G) \leq -\frac{1}{2} d_{f_{\mathbb{Q}}}(G, G^*_{\mathbb{Q}})\right) + \mathbb{P}\left(c_2\tau_n^{-1} \leq \frac{1}{n}\sum_{i=1}^n U_i(G_n)\right)
\end{aligned}$$

and thus

$$\begin{aligned}
&\mathbb{P}\big(d_{f_{\mathbb{Q}}}(\widehat{G}_n, G^*_{\mathbb{Q}}) > t\tau_n^{-1}\big) \\
&\leq \mathbb{P}\left(\exists G \in \Xi_t :\ \frac{1}{n}\sum_{i=1}^n U_i(G) \leq -\frac{1}{2} d_{f_{\mathbb{Q}}}(G, G^*_{\mathbb{Q}})\right) + \mathbb{P}\left(c_2\tau_n^{-1} \leq \frac{1}{n}\sum_{i=1}^n U_i(G_n)\right)
\end{aligned}$$

It remains to find upper bounds for the two terms above. In order to bound the first term, note that for $(x, y) \in \mathbb{R}^d \times \{0, 1\}$ and any $G \in \mathcal{N}_n$ we have

$$\begin{aligned} |h_G(x,y)| &= \begin{cases} \big|1 - 1(x \in G) - \big(1 - 1(x \in G^*_{\mathbb{Q}})\big)\big|, & \text{for } y = 1, \\ \big|1(x \in G) - 1(x \in G^*_{\mathbb{Q}})\big|, & \text{for } y = 0 \end{cases} \\ &= \begin{cases} \big|1(x \in G^*_{\mathbb{Q}}) - 1(x \in G)\big|, & \text{for } y = 1, \\ \big|1(x \in G) - 1(x \in G^*_{\mathbb{Q}})\big|, & \text{for } y = 0 \end{cases} \\ &= 1\big(x \in G \Delta G^*_{\mathbb{Q}}\big). \end{aligned}$$

For all $i = 1, \dots, n$ this implies $|U_i(G)| \le 2$ and

$$\mathbb{E}\big[U_i(G)^2\big] \le \mathbb{E}\big[h_G(X_i, Y_i)^2\big] = \mathbb{E}\big[1\big(x \in G \Delta G^*_{\mathbb{Q}}\big)\big] = d_\Delta(G, G^*_{\mathbb{Q}}) \le 1.$$

By Bernstein's inequality, for all $a > 0$

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^n U_i(G)\right| \ge a\right) \le 2\exp\left(-\frac{k_1 n a^2}{a+1}\right)$$

holds, where $k_1 > 0$ is a constant. By setting $a = \frac{1}{2} d_{f_{\mathbb{Q}}}(G, G^*_{\mathbb{Q}})$, we have

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^n U_i(G)\right| \ge \frac{1}{2} d_{f_{\mathbb{Q}}}(G, G^*_{\mathbb{Q}})\right) \le 2\exp\Big(-k_2 n d_{f,g}(G, G^*_{\mathbb{Q}})^2\Big)$$

for some constant $k_2 > 0$. Noting that by definition $\tau_n \le n^{\frac{1}{\rho+2}}$, by (iii) we have

$$\begin{aligned} &\mathbb{P}\left(\exists G \in \Xi_t : \left|\frac{1}{n}\sum_{i=1}^n U_i(G)\right| \ge \frac{1}{2} d_{f_{\mathbb{Q}}}(G, G^*_{\mathbb{Q}})\right) \\ &\quad \le 2\exp\big(c_3 n^{\frac{\rho}{\rho+2}}\big)\exp\big(-k_2 n t^2 \tau_n^{-2}\big) \\ &\quad \le 2\exp\big(c_3 n^{\frac{\rho}{\rho+2}}\big)\exp\big(-k_2 t^2 n^{-\frac{2}{\rho+2}+1}\big) \\ &\quad \le 2\exp\Big(\big(c_3 - k_2 t^2\big) n^{\frac{\rho}{\rho+2}}\Big) \\ &\quad \le 2\exp\Big(-c_3 \tau_n^\rho\Big) \end{aligned}$$

for all $t \ge \sqrt{\frac{2c_3}{k_2}}$. To bound the second term we use Bernstein's inequality with $a = c_2 \tau_n^{-1}$ and receive

$$\begin{aligned} \mathbb{P}\left(c_2\tau_n^{-1} \le \frac{1}{n}\sum_{i=1}^n U_i(G_n)\right) &\le \exp\left(-\frac{k_1 n c_2^2 \tau_n^{-2}}{c_2\tau_n^{-1} + 1}\right) \\ &\le \exp\big(-k_3 n \tau_n^{-2}\big) \\ &\le \exp\big(-k_3 \tau_n^\rho\big) \end{aligned}$$

for a constant $k_3 > 0$. Therefore, for $t \geq \max\{4c_2, \sqrt{\frac{2c_3}{k_2}}\}$ we find an upper bound

$$\begin{aligned}
&\mathbb{P}\big(d_{f_{\mathbb{Q}}}(\widehat{G}_n, G^*_{\mathbb{Q}}) > t\tau_n^{-1}\big) \\
&\leq \mathbb{P}\left(\exists G \in \Xi_t : \ \frac{1}{n}\sum_{i=1}^n U_i(G) \leq -\frac{1}{2} d_{f_{\mathbb{Q}}}(G, G^*_{\mathbb{Q}})\right) + \mathbb{P}\left(c_2\tau_n^{-1} \leq \frac{1}{n}\sum_{i=1}^n U_i(G_n)\right) \\
&\leq 2\exp\Big(-c_3\tau_n^{\rho}\Big) + \exp\big(-k_3\tau_n^{\rho}\big).
\end{aligned}$$

Observing that $d_{f_{\mathbb{Q}}}(\widehat{G}_n, G^*_{\mathbb{Q}}) \leq 1$, we conclude

$$\begin{aligned}
&\mathbb{E}\big[d^p_{f_{\mathbb{Q}}}(\widehat{G}_n, G^*_{\mathbb{Q}})\big] \\
&\leq \mathbb{E}\big[1\big(d_{f_{\mathbb{Q}}}(\widehat{G}_n, G^*_{\mathbb{Q}}) > t\tau_n^{-1}\big)\big] + t\tau_n^{-p}\mathbb{E}\big[1\big(d_{f_{\mathbb{Q}}}(\widehat{G}_n, G^*_{\mathbb{Q}}) \leq t\tau_n^{-1}\big)\big] \\
&\leq 2\exp\Big(-c_3\tau_n^{\rho}\Big) + \exp\big(-k_3\tau_n^{\rho}\big) + t\tau_n^{-p}
\end{aligned}$$

and thus

$$\begin{aligned}
&\limsup_{n\to\infty} \sup_{f\in\mathcal{F}} \tau_n^p \mathbb{E}\big[d^p_{f_{\mathbb{Q}}}(\widehat{G}_n, G^*_{\mathbb{Q}})\big] \\
&\leq \limsup_{n\to\infty} \tau_n^p\Big(2\exp\Big(-c_3\tau_n^{\rho}\Big) + \exp\big(-k_3\tau_n^{\rho}\big) + t\tau_n^{-p}\Big) \\
&< \infty.
\end{aligned}$$

□

## Appendix C: Convergence rates for neural networks

The main goal of this section is to prove Theorem 3.5. We begin by proving Lemma 3.3. After introducing a few additional lemmas we prove the theorem. We follow this up by proving Lemma 3.7, Corollary 3.8, Lemma 3.10, and Corollary 3.11.

### *C.1. Number of elements of $\mathcal{N}_{L_0,s_0,c}$*

*Proof of Lemma 3.3.* First of all, if $s_0 \leq L_0$ clearly only the last $s_0$ layers influence the realization of a neural network. Thus an upper bound is given by counting the number of neural nets with at most $\min\{s_0, L_0\}$ layers, at most sparsity $s_0$, weights in $\mathcal{W}_c$ and $m_i \leq s_0$ for all $i \in \{1, \dots, L\}$. Each weight can take on $|\mathcal{W}_c| = 2^{c+1} + 1$ different values. The total number of weights can be bounded by

$$\begin{aligned}
m_0 m_1 + \sum_{i=1}^{\min\{s_0, L\}} (m_i + 1) m_{i+1} &\leq d s_0 + \sum_{i=1}^{\min\{s_0, L_0\}} (s_0 + 1) s_0 \\
&\leq d s_0 + \min\{s_0, L_0\}(s_0 + 1)^2 \\
&=: V.
\end{aligned}$$

Note that if $s_0 \leq L_0$, the input dimension does not influence the outcome. Therefore, there are at most

$$\binom{V}{s_0} \leq V^{s_0}$$

possible combinations to pick $s_0$ (possibly) nonzero weights. Thus

$$\left|\mathcal{N}_{L_0,s_0,c}\right| \leq V^{s_0}\left(2^{c+1}+1\right)^{s_0} \leq \left(V2^{c+2}\right)^{s_0}. \qquad \square$$

### *C.2. Proof of the main result*

In order to simplify the approximation results below, we introduce a lemma considering the parallelization and concatenation of two networks $\Phi_1$ and $\Phi_2$. Since these results have been shown in many other articles e.g. [19, 22], we omit the proof.

**Lemma C.1.** *Let $R(\Phi_1) : \mathbb{R}^{d_1} \to \mathbb{R}^{d_3}$ and $R(\Phi_2) : \mathbb{R}^{d_2} \to \mathbb{R}^{d_4}$ be realizations of neural networks with $L_1, L_2$ layers, sparsity $s_1, s_2$ and weights in $\mathcal{W}_{c_1}, \mathcal{W}_{c_2}$ respectively.*

- *If $d_4 = d_1$, the concatenation of the functions $R(\Phi_1) \circ R(\Phi_2)$ can be realized by a neural network with $L = L_1 + L_2 + 1$ layers, sparsity $s \leq 2s_1 + 2s_2$ and weights in $\mathcal{W}_{\max\{c_1,c_2\}}$.*
- *If $d_1 = d_2$, the parallelization of the functions $P(R(\Phi_1), R(\Phi_2)) : \mathbb{R}^{d_1} \to \mathbb{R}^{d_3+d_4}$ given by*

  $$P(R(\Phi_1), R(\Phi_2))(x) := (R(\Phi_1)(x), R(\Phi_2)(x))$$

  *can be realized by a neural network with $L = \max\{L_1, L_2\}$ layers, sparsity $s \leq s_1 + s_2 + 2dL$ and weights in $\mathcal{W}_{\max\{c_1,c_2\}}$.*

Note that we are only using weights $|w| \leq 1$. In order to approximate high numbers, we use the following lemma.

**Lemma C.2.** *Let $c, M \in \mathbb{N}$. Then there exist neural networks $\Phi_1, \Phi_2$ with input dimensions $m_0 = 1$, at most $L = M + 1$ layers, sparsity $s \leq 4M + 1$ and weights in $\mathcal{W}_c$ such that*

$$R(\Phi_1)(x) = 2^M x,$$
$$R(\Phi_2)(x) = 2^M$$

*for all $x \in [0, 1]$.*

*Proof.* The network $\Phi_1$ is given by

$$\Phi_1 := \left(W_1, b_1, \ldots, W_{M+1}, b_{M+1}\right)$$

where $W_{M+1} = W_1^T = (1\ 1)$,

$$W_i = \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}$$

for $i = 2, \dots, M$, $b_i = (0,0)$ for $i \leq M$ and $b_{M+1} = 0$. The other network is

$$\Phi_2 := (W_0, b_0, \dots, W_{M+1}, b_{M+1}) = (W_0, b_0) \times \Phi_1$$

where $W_0 = 0$ and $b_0 = 1$.

The layers of both networks are bounded by $M + 1$. The sparsity of both networks can be bounded by

$$s \leq 2 * 2 + 4(M-1) + 1 = 4M + 1. \qquad \square$$

Next, we construct a neural network for each $G \in \mathcal{K}^{\mathcal{F}}_{\mathbb{Q},\beta,B,\epsilon_1,\epsilon_2,r,d}$ which approximates $G$ well with respect to the metric $d_{f_{\mathbb{Q}}}$. The rough idea for the construction of the network is similar to ideas used in [19]. However, the precise construction in order to adapt to the metric in question differs substantially. The proof of the following theorem is one of the main contributions of this paper.

**Theorem C.3.** *Let $\beta \geq 0$, $B, \rho > 0$ and $d \in \mathbb{N}$ with $d \geq 2$. Let $\mathcal{F}$ be a set of functions*

$$\gamma : [0,1]^{d-1} \to \mathbb{R}$$

*such that the following holds. There exist $\epsilon_0, C_1, C_2 > 0$ and $C_3, C_4 \in \mathbb{N}$ such that for any $\gamma \in \mathcal{F}$ and any $\epsilon \in (0, \epsilon_0)$ there is a neural network $\Phi$ with $L \leq L_0(\epsilon) := C_1 \lceil \log(\epsilon^{-1}) \rceil$ layers, sparsity $s \leq s_0(\epsilon) := C_2 \epsilon^{-\rho} \log(\epsilon^{-1})$ and weights in $\mathcal{W}_c$ with $c = c_0(\epsilon) := C_3 + C_4 \lceil \log(\epsilon^{-1}) \rceil$ such that*

$$\|R(\Phi) - \gamma\|_\infty \leq \epsilon.$$

*Define $\kappa = 1 + \beta$ and let $\mathfrak{Q}$ be a class of potential joint distributions $\mathbb{Q}$ of $(X, Y)$ such that the following conditions hold.*

- *(a) There is a constant $M > 1$ such that for all $\mathbb{Q} \in \mathfrak{Q}$ the marginal distribution of $\mathbb{Q}_X$ has a Lebesgue density bounded by $M$.*
- *(b) There are constants $r \in \mathbb{N}$ and $\epsilon_1, \epsilon_2 > 0$ such that for all $\mathbb{Q} \in \mathfrak{Q}$ the Bayes rule satisfies $G^*_{\mathbb{Q}} \in \mathcal{K}^{\mathcal{F}}_{\mathbb{Q},\beta,B,\epsilon_1,\epsilon_2,r,d}$.*

*Let*

$$\tau_n := \frac{n^{\frac{1}{\rho + 2\kappa - 1}}}{\log^{\frac{2}{\rho}}(n)}.$$

*Then there exist constants $C_1', C_2' > 0$ and $C_3' \in \mathbb{N}$ such that the set*

$$\mathcal{N}_n = \mathcal{N}_{C_1' L_0(\tau_n^{-1}), C_2' s_0(\tau_n^{-1}), C_3' c_0(\tau_n^{-1})}$$

*satisfies the following property. There is a constants $c_2 > 0$ and $N_0 \in \mathbb{N}$ such that for all $n \geq N_0$ and $\mathbb{Q} \in \mathfrak{Q}$ there is a $G \in \mathcal{N}_n$ with*

$$d_{f_{\mathbb{Q}}}(G, G^*_{\mathbb{Q}}) \leq c_2 \tau_n^{-\kappa}.$$

*Proof.* Set $\epsilon_0 := \min\{\epsilon_1, \frac{\epsilon_2}{4}\}$. Choose $N_0$ large enough such that $\tau_{N_0} \geq \epsilon_0^{-1}$. The proof is outlined as follows. We first construct a candidate set $G$ using neural networks. Then, we show that it satisfies the desired properties.

Let $n \geq N_0$, $\mathbb{Q} \in \mathfrak{Q}$ and

$$G^*_{\mathbb{Q}} = H_1 \cup \cdots \cup H_u$$

as in Definition 3.4 with $u \leq r$. We begin with the construction of the candidate set $G$. The idea is to define a network which approximates $G^*_{\mathbb{Q}}$ well on each set $H_\nu$ separately. Define $\iota_\nu, j_\nu, a^\nu_i, b^\nu_i, D_\nu, \gamma_\nu$ and $g_{\nu,x}$ as in Definition 3.4. First, for each $\nu = 1, \ldots, u$ we consider a set $\tilde{D}_\nu$ with borders lying on a grid. The advantage of using $\tilde{H}_\nu = \tilde{D}_\nu \cap H_\nu$ instead of $H_\nu$ is twofold. On the one hand, the grid and parameters of $\mathcal{N}_n$ are defined such that the borders of $\tilde{D}_\nu$ can be constructed precisely. On the other, using the grid, two sets $\tilde{H}_{\nu_1}$, $\tilde{H}_{\nu_2}$ have a minimum distance for $\nu_1 \neq \nu_2$, which is important for our method to work. For $\delta > 0$ let

$$h_\delta = \max\left\{h = 2^{-c} \,\middle|\, h \leq \delta,\ c \in \mathbb{N}\right\}.$$

Set $\epsilon := \tau_n^{-1}$. Define $I := \left\{0, h_{\epsilon^\kappa}, 2h_{\epsilon^\kappa} \ldots, 1 - h_{\epsilon^\kappa}\right\}$ and let

$$\tilde{a}^\nu_j := \min\{a \in I \mid a > a^\nu_j\},$$
$$\tilde{b}^\nu_j := \max\{b \in I \mid b < b^\nu_j\},$$

for $\nu = 1, \ldots, u$, $j = 1, \ldots, d$. Now, set

$$\tilde{D}_\nu := \begin{cases} \prod_{j=1}^d \left[\tilde{a}^\nu_j, \tilde{b}^\nu_j\right], & \text{if } \forall j = 1, \ldots, d\ \tilde{a}^\nu_j < \tilde{b}_j, \\ \emptyset, & \text{otherwise.} \end{cases}$$

Note that $\tilde{b}^\nu_{j_\nu} - \tilde{a}^\nu_{j_\nu} \geq 2\epsilon$ for all $\nu = 1, \ldots, u$ by the choice of $\epsilon_0$. Figure 3 shows the collection of sets $\tilde{D}_\nu$ in the example considered in Figure 1. Obviously we have $\tilde{D}_\nu \subseteq D_\nu$. Let

$$\tilde{H}_\nu = \tilde{D}_\nu \cap \{x \in [0,1]^d \mid \iota_\nu x_{j_\nu} \leq \gamma_\nu(x_{-j_\nu})\}.$$

The idea is to construct a neural network for every $\nu = 1, \ldots, u$ with $\tilde{D}_\nu \neq \emptyset$ which approximates $\tilde{H}_\nu$. We obtain the final neural network by parallelizing and adding up these networks. More specifically, we construct a network that approximates the product of $\mathbb{1}(x \in \tilde{D}_\nu)$ and $\mathbb{1}(\iota_\nu x_{j_\nu} \leq \gamma_\nu(x_{-j_\nu}))$. The latter is approximated by a network $\Phi_{\gamma_\nu}$ which is the concatenation of a network approximating the Heaviside function $\mathbb{1}(x_{j_\nu} \geq 0)$ and a network approximating

$$\tilde{\gamma}_\nu(x) := \big(x_1, \ldots, x_{j_\nu - 1}, \gamma_\nu(x_{-j_\nu}) - \iota_\nu x_{j_\nu}, x_{j_\nu + 1}, \ldots, x_d\big).$$

For $\nu = 1, \ldots, u$, from the prerequisites given in the Theorem, we obtain a network $\Phi^1_{\gamma_\nu}$ with $L^1_{\gamma_\nu} \leq C^1_1 \lceil \log \epsilon^{-1} \rceil$, $s^1_{\gamma_\nu} \leq C^1_2 \epsilon^{-\rho} \lceil \log \epsilon^{-1} \rceil$ and weights in $\mathcal{W}_{c^1}$ with $c^1 := C^1_3 + C^1_4 \lceil \log \epsilon^{-1} \rceil$, such that

$$\|R(\Phi^1_{\gamma_\nu}) - \gamma_\nu\|_\infty \leq \frac{\epsilon}{4}.$$

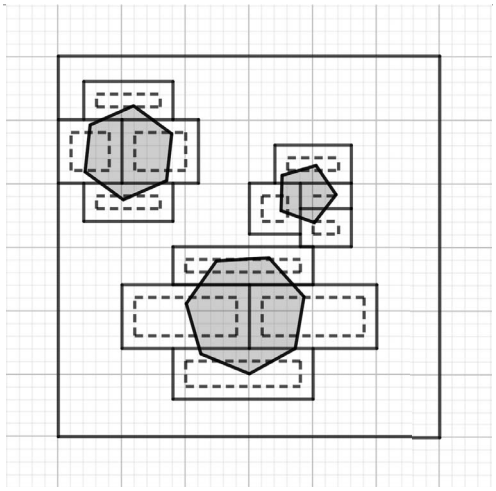

FIG 3. *The collection of sets $\tilde{D}_\nu$ when considering the example from Figure 1. The dotted lines are the borders of the sets $D_1, \dots, D_{12}$. Note that $\delta$ is quite large in this example and observe, that the distance between two sets $\tilde{D}_{\nu_1}, \tilde{D}_{\nu_2}$ is at least $2^{\delta+1}$.*

For technical reasons, we need to slightly change the realizations in order to handle the behavior of the approximations at the border of $\tilde{D}_\nu$. Define $\Phi^2_{\gamma_\nu}$ by its realization

$$\begin{aligned} R(\Phi^2_{\gamma_\nu})(x) &:= \frac{\tilde{b}^\nu_{j_\nu} + a^{\tilde{\nu}}_{j_\nu}}{2} + \sigma\left(R(\Phi^1_{\gamma_\nu})(x) + h_{\frac{\epsilon}{2}} - \frac{\tilde{b}^\nu_{j_\nu} + a^{\tilde{\nu}}_{j_\nu}}{2}\right) \\ &\quad - \sigma\left(\frac{\tilde{b}^\nu_{j_\nu} + a^{\tilde{\nu}}_{j_\nu}}{2} - R(\Phi^1_{\gamma_\nu})(x) + h_{\frac{\epsilon}{2}}\right) \\ &= \begin{cases} R(\Phi^1_{\gamma_\nu})(x) + h_{\frac{\epsilon}{2}}, & \text{if } R(\Phi^1_{\gamma_\nu})(x) \geq \frac{\tilde{b}^\nu_{j_\nu} + a^{\tilde{\nu}}_{j_\nu}}{2} + h_{\frac{\epsilon}{2}}, \\ R(\Phi^1_{\gamma_\nu})(x) - h_{\frac{\epsilon}{2}}, & \text{if } R(\Phi^1_{\gamma_\nu})(x) \leq \frac{\tilde{b}^\nu_{j_\nu} + a^{\tilde{\nu}}_{j_\nu}}{2} - h_{\frac{\epsilon}{2}}, \\ 2R(\Phi^1_{\gamma_\nu})(x) - \frac{\tilde{b}^\nu_{j_\nu} + a^{\tilde{\nu}}_{j_\nu}}{2}, & \text{otherwise.} \end{cases} \end{aligned}$$

Let $\widehat{\gamma}_\nu := R(\Phi^2_{\gamma_\nu})$. Note that

$$\|\widehat{\gamma}_\nu - \gamma_\nu\|_\infty \leq \|R(\Phi^1_{\gamma_\nu}) - \gamma_\nu\|_\infty + h_{\frac{\epsilon}{2}} \leq \frac{\epsilon}{4} + \frac{\epsilon}{2} \leq \epsilon$$

as well as

$$\begin{aligned} \widehat{\gamma}_\nu(x) &\geq \gamma_\nu(x) \text{ for } x : \gamma_\nu(x) \geq \tilde{b}^\nu_{j_\nu}, \\ \widehat{\gamma}_\nu(x) &\leq \gamma_\nu(x) \text{ for } x : \gamma_\nu(x) \leq \tilde{a}^\nu_{j_\nu}. \end{aligned}$$

Using parallelization and concatenation of Lemma C.1, the function

$$R(\Phi^3_{\gamma_\nu})(x) := \big(x_1, \dots, x_{j_\nu - 1}, \widehat{\gamma_\nu}(x_{-j_\nu}) - \iota_\nu x_{j_\nu}, x_{j_\nu+1}, \dots, x_d\big)$$

is a realization of a neural network with at most $L^3_{\gamma_\nu} \leq L^1_{\gamma_\nu} + 3$ layers, sparsity

$$s^3_{\gamma_\nu} \leq 2s^1_{\gamma_\nu} + 14 + 6d + 2dL^1_{\gamma_\nu},$$

and weights in $\mathcal{W}_{c^3}$ with $c^3 = c^1 + 2$. Now let

$$R(\Phi_H)(x) := \sigma(x_{j_\nu} + 1) - \sigma(x_{j_\nu}) = \begin{cases} 0, & \text{for } x_{j_\nu} \leq -1, \\ x_{j_\nu} + 1, & \text{for } -1 < x_{j_\nu} < 0, \\ 1, & \text{for } x_{j_\nu} \geq 0. \end{cases}$$

Note that $R(\Phi_H)(x) \in (0,1)$ if $x_{j_\nu} \in (-1,0)$. Define $\Phi_{\gamma_\nu} := \Phi_H \circ \Phi^3_{\gamma_\nu}$ as in Lemma C.1 concatenation. The network $\Phi_{\gamma_\nu}$ has $L_{\gamma_\nu} \leq C_1^{\gamma_\nu} \lceil \log \epsilon^{-1} \rceil$ layers, sparsity $s_{\gamma_\nu} \leq C_2^{\gamma_\nu} \epsilon^{-\rho} \lceil \log(\epsilon^{-1}) \rceil$ and weights in $\mathcal{W}_{c^{\gamma_\nu}}$ with $c^{\gamma_\nu} := C_3^{\gamma_\nu} + C_4^{\gamma_\nu} \lceil \log(\epsilon^{-1}) \rceil$ for some constants $C_1^{\gamma_\nu}, C_2^{\gamma_\nu} > 0$, $C_3^{\gamma_\nu}, C_4^{\gamma_\nu} \in \mathbb{N}$.

Then $R(\Phi_{\gamma_\nu})(x) = 1$ if $\iota_\nu x_{j_\nu} \leq \widehat{\gamma}_\nu(x_{-j_\nu})$ and $0 \leq R(\Phi_{\gamma_\nu})(x) < 1$ otherwise. Next, for $\nu = 1, \ldots, u$ with $\tilde{H}_\nu \neq \emptyset$ and $i \in \{1, \ldots, d\}$, define the network $\Phi_{\nu,i}$ with realization

$$\begin{aligned} R(\Phi_{\nu,i})(x) &:= 2h_{\epsilon^\kappa}^{-1} \Bigg( \sigma\left( x_i - \tilde{a}_i^\nu + \frac{h_{\epsilon^\kappa}}{2} \right) - \sigma\left( x_i - \tilde{a}_i^\nu \right) - \sigma\left( x_i - \tilde{b}_i^\nu \right) \\ &\qquad + \sigma\left( x_i - \tilde{b}_i^\nu - \frac{h_{\epsilon^\kappa}}{2} \right) \Bigg) \\ &= \begin{cases} 0, & \text{for } x_i \leq \tilde{a}_i^\nu - \frac{h_{\epsilon^\kappa}}{2}, \\ 2h_{\epsilon^\kappa}^{-1} \left( x_i - \tilde{a}_i^\nu + \frac{h_{\epsilon^\kappa}}{2} \right), & \text{for } \tilde{a}_i^\nu - \frac{h_{\epsilon^\kappa}}{2} < x_i < \tilde{a}_i^\nu, \\ 1, & \text{for } \tilde{a}_i^\nu \leq x_i \leq \tilde{b}_i \nu, \\ 1 - 2h_{\epsilon^\kappa}^{-1}(x_i - \tilde{b}_i^\nu), & \text{for, } \tilde{b}_i^\nu < x_i < \tilde{b}_i^\nu + \frac{h_{\epsilon^\kappa}}{2}, \\ 0, & \text{for } x_i \geq \tilde{b}_i^\nu + \frac{h_{\epsilon^\kappa}}{2}. \end{cases} \end{aligned}$$

Note that $\Phi_{\nu,i}$ is a concatenation of two neural networks since $2h_{\epsilon^\kappa}^{-1} \geq 1$. By Lemma C.2, we can realize the function $x \mapsto 2h_{\epsilon^\kappa}^{-1} x$ using a neural network $\Phi_\epsilon$ with $L_\epsilon \leq 1 + \lceil \kappa \rceil c^{\gamma_\nu}$ layers, sparsity $s_\epsilon \leq 4\lceil \kappa \rceil c^{\gamma_\nu} + 5$ and weights in $\mathcal{W}_{c^{\gamma_\nu}}$. Thus $\Phi_{\nu,i}$ has $L_{\nu_i} \leq 4 + \lceil \kappa \rceil c^{\gamma_\nu}$ layers, sparsity $s_{\nu,i} \leq 32\lceil \kappa \rceil c^{\gamma_\lambda} + 32$ and weights in $\mathcal{W}_{\lceil \kappa \rceil c^{\gamma_\nu}}$. We then define

$$R(\Phi_\nu)(x) := \sigma\Bigg( \sum_{i=1}^d R(\Phi_{\nu,i})(x) + R(\Phi_{\gamma_\nu})(x) - d \Bigg).$$

For $x \in [0,1]^d$ we have $R(\Phi_\nu)(x) = 1$ if $x \in \tilde{D}_\nu$ and $\iota_\nu x_{j_\nu} \leq \widehat{\gamma}_\nu(x_{-j_\nu})$. Otherwise $0 \leq R(\Phi_\nu(x)) < 1$ holds. Note that by regarding the construction of $\tilde{D}$, we have $R(\Phi_{\nu_1}) R(\Phi_{\nu_2}) = 0$ for $\nu_1 \neq \nu_2$. In order to construct the sum, we used a parallelization of the networks $\Phi_{\nu,1}, \ldots, \Phi_{\nu,d}$ and $\Phi_{\gamma_\nu}$. Thus the network $\Phi_\nu$ has $L_\nu \leq C_1^\nu \lceil \log \epsilon^{-1} \rceil$ layers, sparsity $s_\nu \leq C_2^\nu \epsilon^{-\rho} \lceil \log \epsilon^{-1} \rceil$ and weights in $\mathcal{W}_{c^\nu}$ with $c^\nu = C_3^\nu + C_4^\nu \lceil \log \epsilon^{-1} \rceil$ for some constants $C_1^\nu, C_2^\nu > 0$, $C_3^\nu, C_4^\nu \in \mathbb{N}$[1]. Note that

[1]Note that the notation suggests that the constants differ depending on $\nu = 1, \ldots, u$. However, due to the construction, this is not the case. The superscripts in the notations above are given in order to describe where the constant comes from. For our analysis below, it does not make a difference if the constants change with $\nu$ or not.

Lemma C.2 was used to construct $d \geq 1$. The realization of the final network is given by

$$R(\Phi)(x) = \sum_{\nu:\tilde{D}_\nu \neq \emptyset} R(\Phi_\nu)(x).$$

Define $G := R(\Phi)^{-1}(1)$. We now verify the desired properties.

We begin by finding constants $C_1', C_2', C_3' > 0$ such that

$$G \in \mathcal{N}_{C_1' L_0(\tau_n^{-1}), C_2' s_0(\tau_n^{-1}), C_3' c_0(\tau_n^{-1})}.$$

Clearly this realization $R(\Phi)$ can be achieved with

$$L \leq \max_{\nu:\tilde{D}_\nu \neq \emptyset} (C_1^\nu) \lceil \log \epsilon^{-1} \rceil + 1 = \max_{\nu:\tilde{D}_\nu \neq \emptyset} (C_1^\nu + 1) \lceil \log \tau_n \rceil =: C_1' L_0(\tau_n^{-1})$$

layers, sparsity

$$s \leq 2u(C_2^\nu \epsilon^{-\rho} \lceil \log \epsilon^{-1} \rceil + Ld) \leq 2r(C_2^\nu \epsilon^{-\rho} \lceil \log \epsilon^{-1} \rceil + Ld) = C_2' s_0(\tau_n^{-1})$$

and weights in $\mathcal{W}_{C_3' c_0(\tau_n)}$ with

$$C_3' c_0(\tau_n^{-1}) \geq C_3^\nu + C_4^\nu \lceil \log \epsilon^{-1} \rceil,$$

with suitably chosen $C_1', C_2' > 0$, $C_3' \in \mathbb{N}$. Note that the constants do not depend on $u$.

Next, we show that the set $G := R(\Phi)^{-1}(1)$ satisfies the desired approximation property $d_{f_\mathbb{Q}}(G, G_\mathbb{Q}^*) \leq \tau_n^{-\kappa}$. First, for $\nu = 1, \ldots, u$ define $E_\nu$ as follows. Let

$$E_\nu := \bigcup_{j=1}^{d} \left( \prod_{i=1}^{j-1} [0,1] \right) \times \left( \left[ a_j^\nu, \tilde{a_j}^\nu \right] \cup \left[ \tilde{b_j}^\nu, b_j^\nu \right] \right) \times \left( \prod_{i=j+1}^{d} [0,1] \right).$$

It is easy to see that $\tilde{D}_\nu = \emptyset$ implies $D_\nu \subseteq E_\nu$. Set $E := \bigcup_{\nu=1}^{u} E_\nu \cup \tilde{D}_\nu$. Figure 4 shows $E$ in the example considered in Figures 1 and 4. Clearly $G_\mathbb{Q}^*, G \subseteq E$. Thus

$$\begin{aligned} d_{f_\mathbb{Q}}(G_\mathbb{Q}, G) &= \int_{G_\mathbb{Q}^* \Delta G} |2f_\mathbb{Q}(x) - 1| \mathbb{Q}_X(\mathrm{d}x) \\ &\leq M \left( \sum_{\nu=1}^{u} \int_{E_\nu} 1 \mathrm{d}x + \sum_{\nu=1}^{u} \int_{(G_\mathbb{Q}^* \Delta G) \cap \tilde{D}_\nu} |2f_\mathbb{Q}(x) - 1| \mathrm{d}x \right) \\ &=: M\big((I) + (II)\big). \end{aligned}$$

We need to bound both terms. For $(I)$ we observe that by construction $h_{\epsilon^\kappa} \leq 2\tau_n^{-\kappa}$ we have

$$(I) \leq \sum_{\nu=1}^{u} \sum_{j=1}^{d} 4h_{\epsilon^\kappa} \leq 4rd\tau_n^{-\kappa}.$$

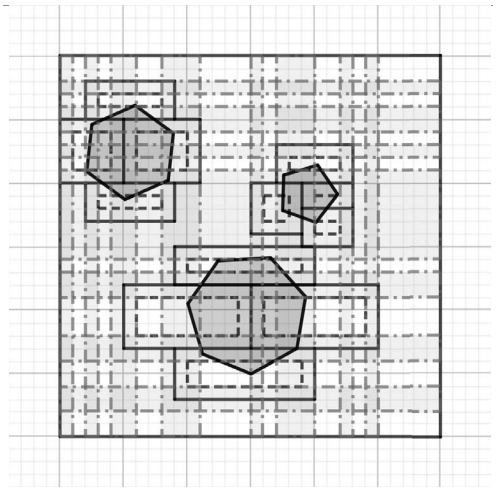

FIG 4. *The set $E$ when considering the example from figures 1 and 4. The light grey set represents $E$. Note that $E$ covers a majority of the space since $\epsilon = \tau_n^{-1}$ is quite large in this example. Observe that $G_{\mathbb{Q}}^*, G \subseteq E$.*

The calculations for the second term are a bit more involved. First, observe that by construction of $G$, for all $\nu = 1, \dots, u$ we have

$$\begin{aligned}&\int_{(G_{\mathbb{Q}}^* \Delta G) \cap \tilde{D}_\nu} |2f_{\mathbb{Q}}(x) - 1| \mathrm{d}x \\ &= \int_{\prod_{i \neq j_\nu} [\tilde{a}_i^\nu, \tilde{b}_i^\nu]} \int_{a_\nu(x_{-j_\nu})}^{b_\nu(x_{-j_\nu})} |2f_{\mathbb{Q}}(x) - 1| \mathrm{d}x_{j_\nu} \mathrm{d}x_{-j_\nu}.\end{aligned}$$

where

$$b_\nu(x_{-j_\nu}) := \begin{cases} \tilde{a}_{j_\nu}^\nu, & \text{if } \widehat{\gamma}_\nu(x_{-j_\nu}), \gamma_\nu(x_{-j_\nu}) < \tilde{a}_{j_\lambda}^\nu, \\ \tilde{b}_{j_\nu}^\nu, & \text{if } \tilde{b}_{j_\nu}^\nu < \widehat{\gamma}_\nu(x_{-j_\nu}), \gamma_\nu(x_{-j_\nu}), \\ \max\{\widehat{\gamma}_\nu(x_{-j_\nu}), \gamma_\nu(x_{-j_\nu})\}, & \text{otherwise,} \end{cases}$$

$$a_\nu(x_{-j_\nu}) := \begin{cases} \tilde{a}_{j_\nu}^\nu, & \text{if } \widehat{\gamma}_\nu(x_{-j_\nu}), \gamma_\nu(x_{-j_\nu}) < \tilde{a}_{j_\lambda}^\nu, \\ \tilde{b}_{j_\nu}^\nu, & \text{if } \tilde{b}_{j_\nu}^\nu < \widehat{\gamma}_\nu(x_{-j_\nu}), \gamma_\nu(x_{-j_\nu}), \\ \min\{\widehat{\gamma}_\nu(x_{-j_\nu}), \gamma_\nu(x_{-j_\nu})\}, & \text{otherwise.} \end{cases}$$

Let $x_{-j_\nu} \in \prod_{i \neq j_\nu} [\tilde{a}_i^\nu, \tilde{b}_i^\nu]$ be fixed. We have the following cases.

- Assume $\gamma_\nu(x_{-j_\nu}) \geq \tilde{b}_{j_\nu}^\nu$. Then by construction, we have

  $$\widehat{\gamma}_\nu(x_{-j_\nu}) \geq \gamma_\nu(x_{-j_\nu}) \geq \tilde{b}_{j_\nu}^\nu$$

  and thus

  $$\int_{a_\nu(x_{-j_\nu})}^{b_\nu(x_{-j_\nu})} |2f_{\mathbb{Q}}(x) - 1| \mathrm{d}x_{j_\nu} = \int_{\tilde{b}_{j_\nu}^\nu}^{\tilde{b}_{j_\nu}^\nu} |2f_{\mathbb{Q}}(x) - 1| \mathrm{d}x_{j_\nu} = 0.$$

- Assume $\gamma_\nu(x_{-j_\nu}) \le \tilde{a}^\nu_{-j_\nu}$. Then by construction, we have

$$\widehat{\gamma}_\nu(x_{-j_\nu}) \ge \gamma_\nu(x_{-j_\nu}) \ge \tilde{a}^\nu_{j_\nu}$$

and thus

$$\int_{a_\nu(x_{-j_\nu})}^{b_\nu(x_{-j_\nu})} |2f_{\mathbb{Q}}(x) - 1| \mathrm{d}x_{j_\nu} = \int_{\tilde{a}^\nu_{j_\nu}}^{\tilde{a}^\nu_{j_\nu}} |2f_{\mathbb{Q}}(x) - 1| \mathrm{d}x_{j_\nu} = 0.$$

- Assume $\gamma_\nu(x_{-j_\nu}) \in (\tilde{a}^\nu_{j_\nu}, \tilde{b}^\nu_{j_\nu})$, by construction

$$x^*_\vdots = (x_1, \dots, x_{j_\nu - 1}, \gamma_\nu(x_{-j_\nu}), x_{j_\nu+1}, \dots, x_d) \in \partial G^*_{\mathbb{Q}}.$$

Consider $\gamma_\nu(x_{-j_\nu}) \le x_{j_\nu} \le b_\nu(x_{-j_\nu})$. Let

$$x = (x_1, \dots, x_{j_\nu - 1}, x_{j_\nu}, x_{j_\nu+1}, \dots, x_d).$$

Now, for $\beta = 0$ we have

$$\begin{aligned}
&\int_{a_\nu(x_{-j_\nu})}^{b_\nu(x_{-j_\nu})} |2f_{\mathbb{Q}}(x) - 1| \mathrm{d}x_{j_\nu} \\
&\quad \le \int_{\gamma_\nu(x_{-j_\nu})}^{\widehat{\gamma}_\nu(x_{-j_\nu})} 1 \mathrm{d}x_{j_\nu} = \widehat{\gamma}_\nu(x_{-j_\nu}) - \gamma_\nu(x_{-j_\nu}) \le \tau_n^{-\kappa}.
\end{aligned}$$

For $\beta > 0$, by definition of $\mathcal{K}^{\mathcal{F}}_{\mathbb{Q},\epsilon_1,\epsilon_2,,r,d}$ we have

$$|2f_{\mathbb{Q}}(x) - 1| \le \left| g_{\nu,x^*}\big(x_{j_\nu} - \gamma_\nu(x_{-j_\nu})\big) \right|$$

Let $m := \max\{k \in \mathbb{N} \mid k < \beta\}$ and $\omega := \beta - m$. Using a Taylor expansion, there exists $y_{j_\nu} \in \big(0, x_{j_\nu} - \gamma_\nu(x_{-j_\nu})\big)$ such that

$$\begin{aligned}
g_{\nu,x^*}\big(x_{j_\nu} - \gamma_\nu(x_{-j_\nu})\big) &= g_{\nu,x^*}\big(x_{j_\nu} - \gamma_\nu(x_{-j_\nu})\big) - g_{\nu,x^*}(0) \\
&= \sum_{i=1}^{m-1} \frac{1}{i!} \partial^i_{j_\nu} g_{\nu,x^*}(0) \big(x_{j_\nu} - \gamma_\nu(x_{-j_\nu})\big)^i \\
&\qquad + \frac{1}{m!} \partial^m_{j_\nu} g_{\nu,x^*}(y_{j_\nu}) \big(x_{j_\nu} - \gamma_\nu(x_{-j_\nu})\big)^m \\
&= \frac{1}{m!} \partial^m_{j_\nu} g_{\nu,x^*}(y_{j_\nu}) \big(x_{j_\nu} - \gamma_\nu(x_{-j_\nu})\big)^m.
\end{aligned}$$

Note that we used the definition of $\mathcal{H}_{\beta,B}$ in the last equality for all $i \le m < \beta$. Thus

$$\begin{aligned}
&|2f_{\mathbb{Q}}(x) - 1| \\
&\le \left| g_{\nu,x^*}\big(x_{j_\nu} - \gamma_\nu(x_{-j_\nu})\big) \right| \\
&\le \frac{1}{m!} \frac{|\partial^m_{j_\nu} g_{\nu,x^*}(y_{j_\nu}) - \partial^m_{j_\nu} g_{\nu,x^*}(0)|}{(y_{j_\nu} - 0)^\omega} \big(x_{j_\nu} - \gamma_\nu(x_{-j_\nu})\big)^\beta
\end{aligned}$$

$$\leq \frac{B}{m!}\big(x_{j_\nu} - \gamma_\lambda(x_{-j_\nu})\big)^\beta.$$

Similarly, for $a_\nu(x_{-j\nu}) \leq x'_{j_\nu} \leq \gamma_\nu(x_{-j_\nu})$ we obtain

$$|2f_{\mathbb{Q}}(x) - 1| \leq \frac{B}{m!}\big(x_{j_\nu} - \gamma_\lambda(x_{-j_\nu})\big)^\beta.$$

This implies

$$\begin{aligned}
&\int_{a_\nu(x_{-j_\nu})}^{b_\nu(x_{-j_\nu})} |2f_{\mathbb{Q}}(x) - 1| \mathrm{d}x_{j_\nu} \mathrm{d}x_{-j_\nu} \\
&\leq \int_{\gamma_\nu(x_{-j_\nu})}^{\widehat{\gamma}_\nu(x_{-j_\nu})} \frac{B}{m!}\big(x_{j_\nu} - \gamma_\nu(x_{-j_\nu})\big)^\beta \mathrm{d}x_{j_\nu} \mathrm{d}x_{-j_\nu} \\
&= \frac{B}{m!(\beta+1)}\big(\widehat{\gamma}_\nu(x_{-j_\nu}) - \gamma_\nu(x_{-j_\nu})\big)^{\beta+1} \\
&\leq \frac{B}{m!(\beta+1)}\tau_n^{-\kappa}.
\end{aligned}$$

Therefore, we have

$$\int_{a_\nu(x_{-j_\nu})}^{b_\nu(x_{-j_\nu})} |2f_{\mathbb{Q}}(x) - 1| \mathrm{d}x_{j_\nu} \mathrm{d}x_{-j_\nu} \leq \max\left\{\frac{B}{m!(\beta+1)}, 1\right\}\tau_n^{-\kappa},$$

for all $\nu = 1, \ldots, u$ and $x_{-j_\nu} \in \prod_{i \neq j_\nu}[\tilde{a}_i^\nu, \tilde{b}_i^\nu]$ which yields

$$(II) \leq \max\left\{\frac{B}{m!(\beta+1)}, 1\right\}\tau_n^{-\kappa}.$$

Thus

$$d_f(G, G^*_{\mathbb{Q}}) \leq \left(4rd + \max\left\{\frac{B}{m!(\beta+1)}, 1\right\}\right)\tau_n^{-\kappa}$$

which concludes the proof. □

The remainder of the proof of our main result is now simple.

*Proof of Theorem 3.5.* We check the requirements in Proposition 2.1.

Conditions (i) and (ii) are clear.

Condition (iii) follows from Theorem C.3.

Lastly, we need to prove (iv). Let $n \geq N_0$ where $N_0$ is defined in Theorem C.3. By Lemma 3.3 we have

$$\begin{aligned}
|\mathcal{N}_n| &= \left|\mathcal{N}_{C'_1 L_0(\tau_n), C'_2 s_0(\tau_n), C'_3 c_0(\tau_n^{-1})}\right| \\
&\leq \big((dC'_2 s_n(\tau_n) + \min\{C'_1 L_n(\tau_n), C'_2 s_0(\tau_n^{-1})\} \\
&\quad \times (C'_2 s_0(\tau_n) + 1)^2) 2^{C'_3 c_0(\tau_n^{-1})+2}\big)^{C'_2 s_0(\tau_n)}.
\end{aligned}$$

Inserting all variables yields

$$|\mathcal{N}_n| \leq \left(k_1 \tau_n^{k_2} \log^2(\tau_n)\right)^{k_3 \tau_n^\rho \log(\tau_n)}$$

for some constants $k_1, k_2, k_3 > 0$. Thus

$$\log\left(|\mathcal{N}_n|\right) \leq k_4 \tau_n^\rho \log^2(\tau_n)$$

for some constant $k_4 > 0$. By setting

$$\tau_n := \frac{n^{\frac{1}{\rho+2\kappa-1}}}{\log^{\frac{2}{\rho}}(n)}$$

we obtain assumption (iv)

$$\log|\mathcal{N}_n| \leq c_3 n^{\frac{\rho}{\rho+2\kappa-1}}.$$ □

### *C.3. Proofs for regular boundaries*

Next, we prove Lemma 3.7. We first provide the corresponding statement from [22]. Lemma 3.7 is a reformulated version.

**Theorem C.4** (Theorem 5 in [22])**.** *For any function $f \in \mathcal{F}_{\beta,B,d}$ and any integers $m \geq 1$ and $N \geq \max\left\{(\beta+1)^d, B+1\right\}$ there exists a neural network $\Phi$ with*

$$L = 8 + (m+5)(1 + \lceil d \log_2 d \rceil)$$

*layers, sparsity*

$$s \leq 94 d^2 (\beta+1)^{2d} N (m+6)(1 + \lceil \log_2 d \rceil)$$

*and weights $|w| \leq 1$ such that*

$$\|R(\Phi) - f\|_\infty \leq (2B+1)3^{d+1} N 2^{-m} + B 2^\beta N^{-\frac{\beta}{d}}.$$

*Proof.* Theorem 5 in [22]. □

*Proof of Lemma 3.7.* Let $N$ be the smallest integer satisfying

$$N \geq k_1 \epsilon^{-\frac{d}{\beta}} \geq \max\left\{\left(B 2^{\beta+2} \epsilon^{-1}\right)^{\frac{d}{\beta}}, (\beta+1)^d, B+1\right\},$$

where $k_1 := \max\left\{\left(B 2^{\beta+2}\right)^{\frac{d}{\beta}}, (\beta+1)^d, B+1\right\}$. Let $k_2$ be the smallest integer satisfying

$$k_2 \geq \log(k_1+1) + \left(\frac{d}{\beta} + 1\right) + \log((2B+1)3^{d+1}) + 1$$

and define $m := k_2 \lceil \log(\epsilon^{-1} \rceil \in \mathbb{N}$. Since $\epsilon < 3^{-1}$ and thus $\log(\epsilon^{-1}) \geq 1$ we have

$$m \geq \left(\log(k_1+1) + \left(\frac{d}{\beta} + 2\right) + \log((2B+1)3^{d+1}) + 1\right) \log(\epsilon^{-1})$$

$$\geq \log(k_1+1) + \left(\frac{d}{\beta}+1\right)\log(\epsilon^{-1}) + \log((2B+1)3^{d+1}) + 2$$
$$\geq N + \log((2B+1)3^{d+1}) + \log(\epsilon^{-1}) + 2.$$

Theorem C.4 implies the following. For any $f \in \mathcal{F}_{\beta,B,d}$ there exists a network $\tilde{\Phi}$ with

$$L = 8 + (k_2\lceil \log \epsilon^{-1}\rceil + 5)(1 + \lceil \log_2 d\rceil)$$

layers, sparsity

$$s \leq 94d^2(\beta+1)^{2d}k_1\epsilon^{-\frac{d}{\beta}}(k_2\lceil \log \epsilon^{-1}\rceil + 6)(1+\lceil \log_2 d\rceil)$$

and weights $|w_i| \leq 1$ such that

$$\|R(\tilde{\Phi}) - f\|_\infty \leq \frac{\epsilon}{2}.$$

Note that

$$L \leq (8 + (2k_2+5)(1+\lceil \log_2 d\rceil))\log \epsilon^{-1} =: c_1 \log \epsilon^{-1},$$
$$s \leq \left(94d^2(\beta+1)^{2d}k_1(2k_2+6)(1+\lceil \log_2 d\rceil)\right)\epsilon^{-\frac{d}{\beta}}\log\epsilon^{-1} =: c_2\epsilon^{-\frac{d}{\beta}}\log\epsilon^{-1}.$$

Let $V$ be defined as in the proof of Lemma 3.3. Following the proof of Lemma 12 of [22], we see that for any $g \leq \frac{\epsilon}{4(L+1)V}$ there is a neural network $\Phi$ with $L$ layers and sparsity $s$ such that

$$\|R(\Phi) - R(\tilde{\Phi})\|_\infty \leq \frac{\epsilon}{2},$$

where the nonzero weights of $\Phi$ are discretized with grid size $g$. Now, define

$$\frac{\epsilon}{4(L+1)V}$$
$$= \frac{\epsilon}{4(L+1)\left(ds + L(s+1)^2\right)}$$
$$\geq \frac{\epsilon}{4(c_1\lceil \log\epsilon^{-1}\rceil + 1)\left(dc_2\epsilon^{-\frac{d}{\beta}}\lceil \log\epsilon^{-1}\rceil + c_1\lceil \log\epsilon^{-1}\rceil(c_2\epsilon^{-\frac{d}{\beta}}\lceil \log\epsilon^{-1}\rceil + 1)^2\right)}$$
$$\geq \frac{1}{4(c_1+1)(dc_2 + c_1(c_2+1)^2}\epsilon^{2+2\frac{d}{\beta}}$$
$$\geq 2^{-\left(c_3 + c_4\lceil \log(\epsilon^{-1})\rceil\right)} =: g,$$

with

$$c_3 := \lceil \log\left(4(c_1+1)(dc_2 + c_1(c_2+1)^2\right)\rceil,$$
$$c_4 := \left\lceil 2 + 2\frac{d}{\beta}\right\rceil.$$

Therefore, all weights are elements of $\mathcal{W}_c$ with $c = c_3 + c_4\lceil \log\epsilon^{-1}\rceil$ and

$$\|R(\Phi) - f\|_\infty \leq \|R(\Phi) - R(\tilde{\Phi})\|_\infty + \|R(\tilde{\Phi}) - f\|_\infty \leq \frac{\epsilon}{2} + \frac{\epsilon}{2} = \epsilon. \qquad \square$$

Lastly, we prove Lemma 3.10. The extension to this case is similar to the extension in [22].

*Proof of Lemma 3.10.* Let

$$\gamma = \gamma_r \circ \cdots \circ \gamma_1 \in \mathcal{G}_{r,t,\beta,B,d}$$

with $\gamma_i = (\gamma_{ij} \circ \iota_{ij})_{j=1}^{d_{i+1}}$. We first construct a candidate network $\Phi$ for $\gamma$. Then, we show that it approximates gamma well and satisfies the required properties.

In order to construct a network that approximates $\gamma$ well, we first approximate $\gamma_{ij}$ and $\tau_{ij}$ using neural networks. The final network is constructed using concatenation and parallelization.

Let $i = 1, \ldots, r$, $j = 1, \ldots, d_i + 1$ and $\epsilon_i > 0$. Using Lemma 3.7 there exist constants $\epsilon_0^i, c_1^i, c_2^i > 0, c_3^i, c_4^i \in \mathbb{N}$ such that there exists a neural network $\Phi_{ij}$ with $L^{ij} \leq c_1^i \lceil \log(\epsilon_i^{-1}) \rceil$ layers, sparsity $s^{ij} \leq c_2^i \epsilon_i^{-\frac{t_i}{\beta_i}} \log(\epsilon_i^{-1})$ and weights in $\mathcal{W}_{c^i}$ with $c^i := c_3^i + c_4^i \lceil \log(\epsilon_i^{-1}) \rceil$ such that

$$\|R(\Phi_{ij})(x) - \gamma_{ij}\|_\infty \leq \epsilon_i$$

if $\epsilon_i < \epsilon_0^i$. Let $\widehat{\gamma}_{ij} := R(\Phi_{ij})(x)$. Additionally, the function $\iota_{ij}$ is the realization of a network with 0 Layers and sparsity $t_i$.

Since concatenating and parallelizing networks using Lemma C.1 leads to linear transformations on the upper bounds on the Layers, sparsity, and the constant $c'$, there exist constants $c_1', c_2' > 0$, $c_3', c_4' \in \mathbb{N}$ such that the function

$$\widehat{\gamma} = \widehat{\gamma}_r \circ \cdots \circ \widehat{\gamma}_1$$

with $\widehat{\gamma}_i = (\widehat{\gamma}_{ij} \circ \iota_{ij})_{j=1}^{d-1}$ is the realization of a network with

$$\begin{aligned} L &\leq c_1' \max_{i=1,\ldots,r} \lceil \log(\epsilon_i^{-1}) \rceil \text{ layers,} \\ s &\leq c_2' \max_{i=1,\ldots,r} \epsilon_i^{-\frac{t_i}{\beta_i}} \log(\epsilon_i^{-1}) \text{ sparsity,} \\ c' &:= c_3' + c_4' \max_{i=1,\ldots,r} \lceil \log(\epsilon_i^{-1}) \rceil \end{aligned}$$

and weights in $\mathcal{W}_{c'}$.

Now, let $\epsilon > 0$ be small enough. We show that $\widehat{\gamma}$ approximates $\gamma$ well for suitabily chosen $\epsilon_i$. Following Lemma 9 in [22] we have

$$\begin{aligned} \|\gamma - \widehat{\gamma}\|_\infty &\leq C \sum_{i=1}^{r} \| \max_{j=1,\ldots,d_{i+1}} |\gamma_{ij} - \widehat{\gamma}_{ij}| \|_\infty^{\prod_{k=i+1}^{r} \min\{\beta_k, 1\}} \\ &\leq C \sum_{i=1}^{r} \epsilon_i^{\prod_{k=i+1}^{r} \min\{\beta_k, 1\}} \\ &\leq Cr \max_{i=1,\ldots,r} \epsilon_i^{\prod_{k=1}^{i+1} \min\{\beta_k, 1\}} \end{aligned}$$

for some constant $C > 0$. Set

$$\epsilon_i := \left(\frac{\epsilon}{Cr}\right)^{\frac{1}{\prod_{k=1}^{i+1} \min\{\beta_k, 1\}}}.$$

First, this implies

$$\|\gamma - \widehat{\gamma}\|_\infty \leq \epsilon.$$

Additionally, the network $\Phi$ has

$$\begin{aligned} L &\leq c_1' \max_{i=1,\dots,r} \lceil \log(\epsilon_i^{-1}) \rceil \leq c_1 \lceil \log(\epsilon^{-1}) \rceil \text{ layers}, \\ s &\leq c_2' \max_{i=1,\dots,r} \epsilon_i^{-\frac{t_i}{\beta_i}} \log(\epsilon_i^{-1}) = c_2 \max_{i=1,\dots,r} \epsilon^{-\frac{t_i}{\beta_i \prod_{k=1}^{i+1} \min\{\beta_k, 1\}}} \log(\epsilon^{-1}) \\ &= c_2 \epsilon^{-\rho} \log(\epsilon^{-1}) \text{ sparsity}, \\ c' &= c_3' + c_4' \max_{i=1,\dots,r} \lceil \log(\epsilon_i^{-1}) \rceil \leq c_3 + c_4 \lceil \log(\epsilon^{-1}) \rceil := c \end{aligned}$$

and weights in $\mathcal{W}_c$ for some constants $c_1, c_2 > 0$, $c_3, c_4 \in \mathbb{N}$. □

## Appendix D: Lower bounds

We first prove Theorem 4.1. The outline of the proof is similar to the proof of Theorem 3 in [18]. However, the setting of Theorem 3.5 differs substantially from theirs. This leads to a new situation and new technical challenges to overcome in the proof of Theorem 4.1.

*Proof of Theorem 4.1.* By Hölders inequality and condition (c) it is enough to consider the case $p = 1$ and the first inequality. Let $\mathfrak{Q}_1 \subseteq \mathfrak{Q}$ be a finite set of potential probability measures of $(X_1, Y_1), \dots, (X_n, Y_n)$. Then

$$\sup_{\mathbb{Q} \in \mathfrak{Q}} \mathbb{E}[d_\Delta(G_n, G_{\mathbb{Q}}^*)] \geq \frac{1}{\#\mathfrak{Q}_1} \sum_{\mathbb{Q} \in \mathfrak{Q}_1} \mathbb{E}[d_\Delta(G_n, G_{\mathbb{Q}}^*)]$$

Hence, it suffices to show that for any estimator $G_n$ we have

$$\frac{1}{\#\mathfrak{Q}_1} \sum_{\mathbb{Q} \in \mathfrak{Q}_1} \mathbb{E}[d_\Delta(G_n, G_{\mathbb{Q}}^*)] \geq c n^{-\frac{1}{2\kappa - 1 + \rho}} \quad \text{a.s.}, \tag{D.1}$$

for some constant $c > 0$. We now define the set $\mathfrak{Q}_1$. Then, we prove $\mathfrak{Q}_1 \subseteq \mathfrak{Q}$. Lastly, we show that $\mathfrak{Q}_1$ satisfies (D.1).

Let $\phi : \mathbb{R} \to [0, 1]$ be an infinitely many times differentiable function with the following properties:

- $\phi(t) = 0$ for $|t| \geq 1$,
- $\phi(0) = 1$.

Let $K \geq 2$ be an integer. For $i \in \{1, \dots, K\}^{d-1}$ define

$$\phi_i : [0,1]^{d-1} \to [0,1],\ \phi_i(y) = k_1 K^{-\beta_2} \prod_{j=1}^{d-1} \phi\left(K\left(y_j - \frac{2i_j - 1}{K}\right)\right)$$

for some $0 < k_1$ small enough. Define

$$W := \prod_{i \in \{1,\dots,K\}^{d-1}} \{0,1\}.$$

For $w \in W$ let

$$\gamma_w : [0,1]^{d-1} \to [0,1],\ \gamma_w(y) = \sum_{i \in \{1,\dots,K\}^{d-1}} w_i \phi_i(y).$$

Now, for $w \in W$ define $\mathbb{Q}_w$ as follows. The marginal distribution $\mathbb{Q}_{w,X}$ is the uniform distribution on $[0,1]^d$ and

$$\begin{aligned} f_{\mathbb{Q}_w}(x) := \frac{1}{2}\left(1 + k_2\left(\gamma_w(x_{-1}) - x_1\right)^{\beta_1}\right) \mathbb{1}\left(x_1 \leq \gamma_w(x_{-1})\right) \\ + \frac{1}{2}\left(1 - k_2 x_1^{\beta_1}\right) \mathbb{1}\left(0 < x_1 \leq \gamma_{1-w}(x_{-1})\right) \\ + \frac{1}{2}\left(1 - k_3\left(x_1 - \gamma_1(x_{-1})\right)^{\beta_1}\right) \mathbb{1}(\gamma_1(x_{-1}) < x_1) \end{aligned}$$

for some $k_2, k_3 > 0$. Finally, let

$$\mathfrak{Q}_1 := \left\{\mathbb{Q}_w \;\middle|\; w \in W\right\}.$$

We now show that $\mathfrak{Q}_1 \subseteq \mathfrak{Q}$ by properly selecting the constants $c_1, k_1, k_2, k_3$ such that $f_{\mathbb{Q}_w}$ is well defined for all $w \in W$ and showing that $\mathfrak{Q}_1$ satisfies the conditions (a), (b), (c).

First of all, we choose $k_1, k_3$ small enough and (given $k_2 > 0$) $K_0$ large enough such that for all $K \geq K_0$ we have

$$\frac{1}{4} \leq f_{\mathbb{Q}_w}(x) \leq 1$$

for all $x \in [0,1]^d$ and $w \in W$.

(a) Clearly for all $w \in W$ the marginal distribution of $\mathbb{Q}_w$ with respect to $X$ has a Lebesgue density which bounded by $1 \leq M$.

(b) We need to show

$$G^*_{\mathbb{Q}_w} = \left\{x \in [0,1]^d \;\middle|\; f_{\mathbb{Q}_w}(x) \geq \frac{1}{2}\right\} \in \mathcal{K}^{\mathcal{F}_{\beta_2, B_2, d-1}}_{\mathbb{Q}, \beta, B, \epsilon_1, \epsilon_2, r, d}$$

for all $w \in W$.

1. Clearly by selecting $\nu = u = 1$, $j = j_\nu = 1$, $\iota_2 = 1$, $D_\nu = [0,1]^d$ and $\gamma = \gamma_w$ we have

$$G^*_{\mathbb{Q}_w} = H_1 = D_\nu \cap \big\{x \in [0,1]^d \mid \iota_2 x_1 \leq \gamma(x_{-1})\big\}.$$

   For $k_1$ small enough we also have $\gamma \in \mathcal{F}_{\beta_2, B_2, d-1}$ for all $w \in W$.
2. Clear.
3. If $\beta_1 > 0$, for $w \in W$ and $x \in \partial G^*_{\mathbb{Q}_w}$ we have $x_1 = \gamma_w(x_{-1})$. Let

$$g_{\nu,x} : [0,1] \to \mathbb{R},\ g_{\nu,x}(y) = \max\{k_2, k_3\} y^{\beta_1}.$$

   Note that for $k_2, k_3$ small enough we have $g_{\nu,x} \in \mathcal{H}_{\beta_1, B_1}$. Additionally, we have

$$|2f_{\mathbb{Q}_w}(y) - 1| \leq g_{\nu,x}(y - x_1),\ \text{for } y \geq x_1,$$
$$|2f_{\mathbb{Q}_w}(y) - 1| \leq g_{\nu,x}(x_1 - y),\ \text{for } y \leq x_1.$$

4. Clear.

This implies the assertion.

(c) Let $w \in W$. For $\beta_1 = 0$ we have

$$d^\kappa_\Delta(G, G^*_{\mathbb{Q}_w}) = d_\Delta(G, G^*_{\mathbb{Q}_w}) = \frac{1}{\min\{k_2, k_3\}} d_{f_{\mathbb{Q}_w}}(G, G^*_{\mathbb{Q}_w})$$

For $\beta_1 > 0$, there is an $\eta_0 > 0$ such that for all $0 < \eta \leq \eta_0$ we have

$$\begin{aligned}
&\lambda\Big(\big\{x \in [0,1]^d \;\big|\; |2f_{\mathbb{Q}_w}(x) - 1| \leq \eta\big\}\Big) \\
&\leq \lambda\Bigg(\Big\{x \in [0,1]^d \;\Big|\; x_1 \leq \gamma_w(x_{-1}),\ k_2(\gamma_w(x_{-1}) - x_1)^{\beta_1} \leq \eta\Big\} \\
&\quad \cup \Big\{x \in [0,1]^d \;\Big|\; x_1 \leq \gamma_{1-w}(x_{-1}),\ k_2 x_1^{\beta_1} \leq \eta\Big\} \\
&\quad \cup \Big\{x \in [0,1]^d \;\Big|\; \gamma_1(x_{-1}) \leq x_1,\ k_3\Big(x_1 - \gamma_1(x_{-1})\Big)^{\beta_1} \leq \eta\Big\}\Bigg) \\
&\leq \lambda\Bigg(\left\{x \in [0,1]^d \;\middle|\; \gamma_2(x_{-1}) - \frac{1}{k_2^{\frac{1}{\beta_1}}} \eta^{\frac{1}{\beta_1}} \leq x_1 \leq \gamma_w(x_{-1})\right\} \\
&\quad \cup \left\{x \in [0,1]^d \;\middle|\; x_1 \leq \frac{1}{k_2^{\frac{1}{\beta_1}}} \eta^{\frac{1}{\beta_1}}\right\} \\
&\quad \cup \left\{x \in [0,1]^d \;\middle|\; \gamma_1(x_{-1}) \leq x_1,\ \gamma_1(x_{-1}) + \frac{1}{k_3^{\frac{1}{\beta_1}}} \eta^{\frac{1}{\beta_1}}\right\}\Bigg) \\
&\leq \left(\frac{2}{k_2^{\frac{1}{\beta_1}}} + \frac{1}{k_3^{\frac{1}{\beta_1}}}\right) \eta^{\frac{1}{\beta_1}}.
\end{aligned}$$

Following Proposition 1 in of [24] there exists $\tilde{c}_1, \tilde{\eta}_0 > 0$ such that

$$d_\Delta^\kappa(G, G^*_{\mathbb{Q}_w}) \leq \tilde{c}_1 d_{f_{\mathbb{Q}_w}}(G, G^*_{\mathbb{Q}_w})$$

for all $G$ such that $d_\Delta(G, G^*_{\mathbb{Q}_w}) \leq \tilde{\eta}_0$. If $\tilde{\eta}_0 \geq 1$ this implies the assertion with $c_1 := \tilde{c}_1$. If not, the assertion is implied by setting $c_1 := \frac{\tilde{c}_1}{\tilde{\eta}_0^\kappa}$.

Next, we prove Inequality (D.1). For $w \in W$ write $\mathbb{Q}_w^n$ for the probability measure of the distribution of $(X_1, Y_1), \ldots, (X_n, Y_n)$ when the underlying distribution is $\mathbb{Q}_w$. Define the product measure $\psi = \zeta \times \lambda$, where $\zeta$ is the counting measure on $\{0, 1\}$. Note that $\mathbb{Q}_w$ has a density with respect to $\psi$ which is given by

$$\mathrm{d}\mathbb{Q}_w = \frac{\mathrm{d}\mathbb{Q}_w}{\mathrm{d}\psi}(y, x) := \mathbb{1}(y = 1) f_w(x) + \mathbb{1}(y = 0) \cdot (1 - f_w(x)).$$

Assume $w_1, w_2 \in W$ differ by only 1 entry. We obtain

$$\int \min\{\mathrm{d}\mathbb{Q}^n_{w_1}, \mathrm{d}\mathbb{Q}^n_{w_2}\}\mathrm{d}\psi = \int \min\{\mathrm{d}\mathbb{Q}^n_0, \mathrm{d}\mathbb{Q}^n_1\}\mathrm{d}\psi,$$

where for $s = 0, 1$ we write $\mathbb{Q}^n_s = \mathbb{Q}^n_{w^s}$ with

$$w_i^s := \begin{cases} s, & \text{for } i_1 = \cdots = i_{d-1} = 1, \\ 0, & \text{otherwise} \end{cases}$$

for $i \in \{1, \ldots, K\}^{d-1}$. Then, using Assouad's Lemma (version from [15] for set estimation) we get

$$\begin{aligned} &\frac{1}{\#\mathfrak{Q}_1} \sum_{\mathbb{Q} \in \mathfrak{Q}_1} \mathbb{E}[d_\Delta(G_n, G^*_{\mathbb{Q}})] \\ &\geq \frac{1}{2} K^{d-1} \lambda\left( \left\{ x \in [0, 1]^d \;\middle|\; x_1 \leq \phi_1(x_{-1}) \right\} \right) \int \min\{\mathrm{d}\mathbb{Q}^n_0, \mathrm{d}\mathbb{Q}^n_1\}\mathrm{d}\psi \\ &= \frac{1}{2} k_1 K^{d-1-\beta_2} \int_{\mathbb{R}^{d-1}} \prod_{j=1}^{d-1} \phi(K x_{j+1}) \mathrm{d}x_{-1} \int \min\{\mathrm{d}\mathbb{Q}^n_0, \mathrm{d}\mathbb{Q}^n_1\}\mathrm{d}\psi. \end{aligned}$$

We first bound the term $\int \min\{\mathrm{d}\mathbb{Q}^n_0, \mathrm{d}\mathbb{Q}^n_1\}\mathrm{d}\psi$. By using the fact that

$$\int \mathrm{d}\mathbb{Q}_0 \mathrm{d}\psi = 1$$

and Hölders inequality in the fourth row we calculate

$$\begin{aligned} &\int \min\{\mathrm{d}\mathbb{Q}^n_0, \mathrm{d}\mathbb{Q}^n_1\}\mathrm{d}\psi \\ &= 1 - \frac{1}{2} \int |\mathrm{d}\mathbb{Q}^n_0 - \mathrm{d}\mathbb{Q}^n_1| \mathrm{d}\psi \end{aligned}$$

$$
\begin{aligned}
&= 1 - \frac{1}{2}\int \left|\sqrt{\mathrm{d}\mathbb{Q}_0^n} - \sqrt{\mathrm{d}\mathbb{Q}_1^n}\right| \cdot \left|\sqrt{\mathrm{d}\mathbb{Q}_0^n} + \sqrt{\mathrm{d}\mathbb{Q}_1^n}\right| \mathrm{d}\psi \\
&\geq 1 - \frac{1}{2}\left(\int \left(\sqrt{\mathrm{d}\mathbb{Q}_0^n} - \sqrt{\mathrm{d}\mathbb{Q}_1^n}\right)^2 \mathrm{d}\psi\right)^{\frac{1}{2}} \left(\int \left(\sqrt{\mathrm{d}\mathbb{Q}_0^n} + \sqrt{\mathrm{d}\mathbb{Q}_1^n}\right)^2 \mathrm{d}\psi\right)^{\frac{1}{2}}.
\end{aligned}
$$

By repeatedly using the fact that $\int \mathrm{d}\mathbb{Q}_0 \mathrm{d}\psi = 1$, this implies

$$
\begin{aligned}
&\int \min\{\mathrm{d}\mathbb{Q}_0^n, \mathrm{d}\mathbb{Q}_1^n\}\mathrm{d}\psi \\
&= 1 - \frac{1}{2}\left(2\left(1 - \int \sqrt{\mathrm{d}\mathbb{Q}_0^n \mathrm{d}\mathbb{Q}_1^n}\mathrm{d}\psi\right)\right)^{\frac{1}{2}} \left(2\left(1 + \int \sqrt{\mathrm{d}\mathbb{Q}_0^n \mathrm{d}\mathbb{Q}_1^n}\mathrm{d}\psi\right)\right)^{\frac{1}{2}} \\
&= 1 - \left(1 - \left(\int \sqrt{\mathrm{d}\mathbb{Q}_0^n \mathrm{d}\mathbb{Q}_1^n}\mathrm{d}\psi\right)^2\right)^{\frac{1}{2}} \\
&\geq 1 - \left(1 - \left(\int \sqrt{\mathrm{d}\mathbb{Q}_0^n \mathrm{d}\mathbb{Q}_1^n}\mathrm{d}\psi\right)^2 + \frac{1}{4}\left(\int \sqrt{\mathrm{d}\mathbb{Q}_0^n \mathrm{d}\mathbb{Q}_1^n}\mathrm{d}\psi\right)^4\right)^{\frac{1}{2}} \\
&= 1 - \left(1 - \frac{1}{2}\left(\int \sqrt{\mathrm{d}\mathbb{Q}_0^n \mathrm{d}\mathbb{Q}_1^n}\mathrm{d}\psi\right)^2\right) \\
&= \frac{1}{2}\left(\int \sqrt{\mathrm{d}\mathbb{Q}_0^n \mathrm{d}\mathbb{Q}_1^n}\mathrm{d}\psi\right)^2.
\end{aligned}
$$

With independence we have

$$
\int \sqrt{\mathrm{d}\mathbb{Q}_0^n \mathrm{d}\mathbb{Q}_1^n}\mathrm{d}\psi = \left(\int \sqrt{\mathrm{d}\mathbb{Q}_0 \mathrm{d}\mathbb{Q}_1}\mathrm{d}\psi\right)^n.
$$

Additionally, observe that

$$
\begin{aligned}
\int \sqrt{\mathrm{d}\mathbb{Q}_0 \mathrm{d}\mathbb{Q}_1}\mathrm{d}\psi &= \frac{1}{2}\int \mathrm{d}\mathbb{Q}_0 \mathrm{d}\psi + \frac{1}{2}\int \mathrm{d}\mathbb{Q}_1 \mathrm{d}\psi - \frac{1}{2}\int \left(\sqrt{\mathrm{d}\mathbb{Q}_0} - \sqrt{\mathrm{d}\mathbb{Q}_1}\right)^2 \mathrm{d}\psi \\
&= 1 - \frac{1}{2}\int \left(\sqrt{\mathrm{d}\mathbb{Q}_0} - \sqrt{\mathrm{d}\mathbb{Q}_1}\right)^2 \mathrm{d}\psi
\end{aligned}
$$

and

$$
\begin{aligned}
&\int \left(\sqrt{\mathrm{d}\mathbb{Q}_0} - \sqrt{\mathrm{d}\mathbb{Q}_1}\right)^2 \mathrm{d}\psi \\
&\leq \int \left(\sqrt{f_{w^0}(x)} - \sqrt{f_{w^1}(x)}\right)^2 \mathrm{d}x + \int \left(\sqrt{1 - f_{w^0}(x)} - \sqrt{1 - f_{w^1}(x)}\right)^2 \mathrm{d}x \\
&\leq 2\int \left(f_{w^0}(x) - f_{w^1}(x)\right)^2 \mathrm{d}x + 2\int \left(1 - f_{w^0}(x) - \left(1 - f_{w^1}(x)\right)\right)^2 \mathrm{d}x \\
&= 4\int \left(f_{w^0}(x) - f_{w^1}(x)\right)^2 \mathrm{d}x
\end{aligned}
$$

where we used that $f_w(x) \geq \frac{1}{4}$ for all $x \in [0,1]^d$ and $w \in W$. Next, we calculate

$$\begin{aligned}
&\int \left(f_{w^0}(x) - f_{w^1}(x)\right)^2 \mathrm{d}x \\
&\geq \frac{1}{4} \int_{[0,1]^{d-1}} \int_0^{\phi_j(x_{-1})} \left(k_2 \left(\phi_j(x_{-1}) - x_1\right)^{\beta_1} + k_2 x_1^{\beta_1}\right)^2 \mathrm{d}x_1 \mathrm{d}x_{-1} \\
&\geq \frac{k_2^2}{4} \int_{[0,1]^{d-1}} \int_0^{\phi_j(x_{-1})} \left(\phi_j(x_{-1}) - x_1\right)^{2\beta_1} \mathrm{d}x_1 \mathrm{d}x_{-1} \\
&\quad + \frac{k_2^2}{4} \int_{[0,1]^{d-1}} \int_0^{\phi_j(x_{-1})} x_1^{2\beta_1} \mathrm{d}x_1 \mathrm{d}x_{-1} \\
&= \frac{k_2^2}{4}(I_1 + I_2).
\end{aligned}$$

We need to control the terms $I_1$ and $I_2$. For the first we obtain

$$\begin{aligned}
I_1 &:= \int_{[0,1]^{d-1}} \int_0^{\phi_j(x_{-1})} \left(\phi_j(x_{-1}) - x_1\right)^{2\beta_1} \mathrm{d}x_1 \mathrm{d}x_{-1} \\
&= \int_{[0,1]^{d-1}} \int_0^{\phi_j(x_{-1})} x_1^{2\beta_1} \mathrm{d}x_1 \mathrm{d}x_{-1} \\
&= \frac{1}{1+2\beta_1} \int_{[0,1]^{d-1}} \phi_j(x_{-1})^{1+2\beta_1} \mathrm{d}x_{-1} \\
&\leq \frac{k_1^{1+2\beta_1}}{1+2\beta_1} K^{-\beta_2(1+2\beta_1)} \int_{\mathbb{R}^{d-1}} \prod_{j=1}^{d-1} \phi(K x_{j+1})^{1+2\beta_1} \mathrm{d}x_{-1} \\
&\leq 2\frac{k_1^{1+2\beta_1}}{1+2\beta_1} K^{-\beta_2(1+2\beta_1)-(d-1)} \\
&= 2\frac{k_1^{1+2\beta_1}}{1+2\beta_1} K^{-\beta_2(2\kappa-1+\rho)}
\end{aligned}$$

and similarly

$$\begin{aligned}
I_2 &:= \int_{[0,1]^{d-1}} \int_0^{\phi_j(x_{-1})} x_1^{2\beta_1} \mathrm{d}x_1 \mathrm{d}x_{-1} \\
&\leq 2\frac{k_1^{1+2\beta_1}}{1+2\beta_1} K^{-\beta_2(2\kappa-1+\rho)}.
\end{aligned}$$

This implies

$$\int \min\{\mathrm{d}\mathbb{Q}_0^n, \mathrm{d}\mathbb{Q}_1^n\} \mathrm{d}\psi \geq \frac{1}{2}\left(1 - c^* K^{-\beta_2(2\kappa-1+\rho)}\right)^{2n}$$

for some constant $c^* > 0$. By setting $K := n^{\frac{1}{\beta_2}\frac{1}{2\kappa-1+\rho}}$ we obtain

$$\int \min\{\mathrm{d}\mathbb{Q}_0^n, \mathrm{d}\mathbb{Q}_1^n\} \mathrm{d}\psi \geq \frac{1}{2}\left(1 - c^* \frac{1}{n}\right)^{2n} > c'$$

for some constant $c' > 0$ for $n$ large enough. Thus

$$\begin{aligned}
&\frac{1}{\#\mathfrak{Q}_1} \sum_{\mathbb{Q}\in\mathfrak{Q}_1} \mathbb{E}[d_\Delta(G_n, G^*_{\mathbb{Q}})] \\
&= \frac{1}{2} k_1 K^{d-1-\beta_2} \int_{\mathbb{R}^{d-1}} \prod_{j=1}^{d-1} \phi(Kx_{j+1}) \mathrm{d}x_{-1} \int \min\{\mathrm{d}\mathbb{Q}_0^n, \mathrm{d}\mathbb{Q}_1^n\} \mathrm{d}\psi \\
&\geq \frac{1}{2} k_1 K^{-\beta_2} \int_{\mathbb{R}^{d-1}} \prod_{j=1}^{d-1} \phi(x_{j+1}) \mathrm{d}x_{-1} \cdot c' \\
&\geq c n^{-\frac{1}{2\kappa-1+\rho}}
\end{aligned}$$

for some constant $c > 0$. This concludes the proof. □

Lastly, the proof of Theorem 4.2 is provided. The ideas used in the proof are very similar to those used in the proof of Theorem 4.1 above. We therefore only focus on the differences.

*Proof of Theorem 4.2.* As in the proof of Theorem 4.1 the strategy is to show that for any estimator $G_n$ we have

$$\frac{1}{\#\mathfrak{Q}_1} \sum_{\mathbb{Q}\in\mathfrak{Q}_1} \mathbb{E}[d_\Delta(G_n, G^*_{\mathbb{Q}})] \geq c n^{\frac{1}{2\kappa-1+\rho}} \quad \text{a.s.},$$

for some constant $c > 0$ and some finite set $\mathfrak{Q}_1 \subseteq \mathfrak{Q}$. Let $K \geq 2$ be an integer and let

$$i_{\text{opt}} := \arg\max_{i=1,\dots,r_2} \frac{t_i}{\beta^*_{2,i}}.$$

As in the proof of Theorem 4.1 define $\phi : \mathbb{R} \to [0,1]$ to be an infinitely many times differentiable function with the following two properties:

- $\phi(t) = 0$ for $|t| \geq 1$,
- $\phi(0) = 1$.

Note that $\phi^\alpha$ also fulfills both properties for any $\alpha > 0$, though it may not be infinitely many times differentiable. For $i \in \{1,\dots,K\}^{t_{i_{\text{opt}}}}$ define

$$\phi_i : [0,1]^{d_1-1} \to [0,1], \ \phi_i(y) = k_1 K^{-\beta^*_{2,i_{\text{opt}}}} \prod_{j=1}^{t_{i_{\text{opt}}}} \phi^\alpha \left( K \left( y_j - \frac{2i_j - 1}{K} \right) \right)$$

for $\alpha := \prod_{k=i_{\text{opt}}}^{r_2} \min\{\beta_k, 1\}$ and some $0 < k_1$ small enough. Define

$$W := \prod_{i\in\{1,\dots,K\}^{d-1}} \{0,1\}.$$

For $w \in W$ let

$$\gamma_w : [0,1]^{d_1-1} \to [0,1], \ \gamma_w(y) = \sum_{i\in\{1,\dots,K\}^{t_{i_{\text{opt}}}}} w_i \phi_i(y).$$

Now, for $w \in W$ we define $\mathbb{Q}_w$ as before. The marginal distribution $\mathbb{Q}_X$ is the uniform distribution on $[0,1]^d$ and

$$\begin{aligned} f_{\mathbb{Q}_w}(x) := \frac{1}{2}\left(1 + k_2\left(\gamma_w(x_{-1}) - x_1\right)^{\beta_1}\right) \mathbb{1}\left(x_1 \le \gamma_w(x_{-1})\right) \\ + \frac{1}{2}\left(1 - k_2 x_1^{\beta_1}\right) \mathbb{1}\left(0 < x_1 \le \gamma_{1-w}(x_{-1})\right) \\ + \frac{1}{2}\left(1 - k_3\left(x_1 - \gamma_1(x_{-1})\right)^{\beta_1}\right) \mathbb{1}(\gamma_1(x_{-1}) < x_1) \end{aligned}$$

for some $k_2, k_3 > 0$. Note that for $k_1 > 0$ small enough $\gamma := \gamma_w \in \mathcal{G}_{r_2,t,\beta_2,B_2,d'}$ by defining

$$\begin{aligned} \gamma_i(y) &= (y_1, \dots, y_{t_i}, 0, \dots, 0), \text{ for } i < i_{\text{opt}}, \\ \gamma_i(y) &= (\psi(y), 0, \dots, 0), \text{ for } i = i_{\text{opt}}, \\ \gamma_i(y) &= \left(k_1^{\alpha_i} y_1^{\min\{\beta_i,1\}}, 0, \dots, 0\right), \text{ for } i > i_{\text{opt}}, \end{aligned}$$

where

$$\begin{aligned} \psi(y) &:= \sum_{i \in \{1,\dots,K\}^{t_{i_{\text{opt}}}}} w_i k_1^{\alpha_{i_{\text{opt}}}} K^{-\beta_{2,i_{\text{opt}}}} \prod_{j=1}^{t_{i_{\text{opt}}}} \phi\left(K\left(y_j - \frac{2i_j - 1}{K}\right)\right), \\ \alpha_i &:= \frac{1}{(r_2 - i_{\text{opt}} + 1)\prod_{k=i+1}^{r_2} \min\{\beta_k, 1\}} \end{aligned}$$

for $i_{\text{opt}} \le i \le r_2$. The rest of the proof is analogous to the proof of Theorem 4.1. □

## Acknowledgments

First and foremost, I would like to thank Enno Mammen for supporting me with some helpful comments and inspiring insights during the creation of this paper. Additionally, many thanks goes to Munir Hiabu for assisting with comments during the final stages of the working process.

## References

[1] Jean-Yves Audibert and Alexandre B. Tsybakov. Fast learning rates for plug-in classifiers. *The Annals of Statistics*, 35(2):608–633, 2007. MR2336861

[2] Andrew R. Barron. Approximation and estimation bounds for artificial neural networks. *Machine Learning*, 14(1):115–133, 1994.

[3] Thijs Bos and Johannes Schmidt-Hieber. Convergence rates of deep relu networks for multiclass classification. *Electronic Journal of Statistics*, 16(1):2724–2773, 2022. MR4406243

[4] Ronan Collobert and Jason Weston. A unified architecture for natural language processing: Deep neural networks with multitask learning. In *Proceedings of the 25th International Conference on Machine Learning*, pages 160–167, 2008.

[5] George Cybenko. Approximation by superpositions of a sigmoidal function. *Mathematics of Control, Signals and Systems*, 2(4):303–314, 1989. MR1015670

[6] Richard M. Dudley. Metric entropy of some classes of sets with differentiable boundaries. *Journal of Approximation Theory*, 10(3):227–236, 1974. MR0358168

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

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

[9] Masaaki Imaizumi and Kenji Fukumizu. Deep neural networks learn non-smooth functions effectively. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 869–878. PMLR, 2019.

[10] Javed Khan, Jun S. Wei, Markus Ringner, Lao H. Saal, Marc Ladanyi, Frank Westermann, Frank Berthold, Manfred Schwab, Cristina R. Antonescu, Carsten Peterson, et al. Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. *Nature Medicine*, 7(6):673–679, 2001.

[11] Yongdai Kim, Ilsang Ohn, and Dongha Kim. Fast convergence rates of deep neural networks for classification. *Neural Networks*, 138:179–197, 2021. MR3796894

[12] Michael Kohler, Adam Krzyżak, and Sophie Langer. Estimation of a function of low local dimensionality by deep neural networks. *IEEE Transactions on Information Theory*, 2022. MR4433267

[13] 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.

[14] Michael Kohler and Sophie Langer. On the rate of convergence of fully connected deep neural network regression estimates. *The Annals of Statistics*, 49(4):2231–2249, 2021. MR4319248

[15] Aleksandr Petrovich Korostelev and Alexandre B. Tsybakov. *Minimax Theory of Image Reconstruction*, volume 82. Springer Science & Business Media, 2012. MR1226450

[16] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521(7553):436–444, 2015.

[17] Christian Leibig, Vaneeda Allken, Murat Seçkin Ayhan, Philipp Berens, and Siegfried Wahl. Leveraging uncertainty information from deep neural networks for disease detection. *Scientific Reports*, 7(1):1–14, 2017.

[18] Enno Mammen and Alexander B. Tsybakov. Smooth discrimination analysis. *The Annals of Statistics*, 27(6):1808–1829, 1999. MR1765618

[19] Philipp Petersen and Felix Voigtlaender. Optimal approximation of piecewise smooth functions using deep relu neural networks. *Neural Networks*, 108:296–330, 2018.

[20] Philipp Petersen and Felix Voigtlaender. Optimal learning of high-dimensional classification problems using deep neural networks. *arXiv preprint arXiv:2112.12555*, 2021.

[21] Jürgen Schmidhuber. Deep learning in neural networks: An overview. *Neural Networks*, 61:85–117, 2015.

[22] Johannes Schmidt-Hieber et al. Nonparametric regression using deep neural networks with relu activation function. *The Annals of Statistics*, 48(4):1875–1897, 2020. MR4134774

[23] Bernadetta Tarigan and Sara A. Van De Geer. Classifiers of support vector machine type with $\ell_1$ complexity regularization. *Bernoulli*, 12(6):1045–1076, 2006. MR2274857

[24] Alexander B. Tsybakov. Optimal aggregation of classifiers in statistical learning. *The Annals of Statistics*, 32(1):135–166, 2004. MR2051002

[25] Qiang Wu and Ding-Xuan Zhou. SVM soft margin classifiers: Linear programming versus quadratic programming. *Neural Computation*, 17(5):1160–1187, 2005. MR2176072

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