Risk bounds for deep learning

Bos, J.M.

Citation

Bos, J. M. (2024, June 19). *Risk bounds for deep learning*. Retrieved from https://hdl.handle.net/1887/3763887

Version:	Publisher's Version
License:	Licence agreement concerning inclusion of doctoral thesis in the Institutional Repository of the University of Leiden
Downloaded from:	https://hdl.handle.net/1887/3763887

Note: To cite this publication please use the final published version (if applicable).

Chapter 2

Convergence rates of deep ReLU networks for multiclass classification

Abstract

For classification problems, trained deep neural networks return probabilities of class memberships. In this chapter we study convergence of the learned probabilities to the true conditional class probabilities. More specifically we consider sparse deep ReLU network reconstructions minimizing cross-entropy loss in the multiclass classification setup. Interesting phenomena occur when the class membership probabilities are close to zero. Convergence rates are derived that depend on the near-zero behaviour via a margin-type condition.

2.1 Introduction

The classification performance of a procedure is often evaluated by considering the percentage of test samples that is assigned to the correct class. The corresponding loss for this performance criterion is called the 0-1 loss. Theoretical results for this loss are often related to the the margin condition [94, 145, 5], which allows for fast convergence rates. Empirical risk minimization with respect to the non-convex 0-1 loss is computationally hard and convex surrogate losses are used instead, see for example [9, 140]. More recently, similar results have been obtained for deep neural networks in the binary classification setting. This includes results for standard deep neural networks in combination with the hinge and logistic loss as surrogate losses [67], as

well as results for deep convolutional neural networks with the least squares loss [74] and logistic loss [76] as surrogate losses. More details can be found in the discussion following Theorem 2.3.3.

Trained neural networks provide more information than just a guess of the class membership. For each class and each input, they return an estimate for the probability that the true label is in this class. For an illustration, see for example Figure 4 in the seminal work [80]. In applications it is often important how certain a network is about class memberships, especially in safety-critical systems where a wrong decision can have serious consequences such as automated driving [22] and AI based disease detection [87, 52]. In fact, the conditional class probabilities provide us with a notion of confidence. If the probability of the largest class is nearly one, it is likely that this class is indeed the true one. On the other hand, if there is no clear largest class and the conditional class probabilities of several classes are close to each other, it might be advisable to let a human examine the case instead of basing the decision only on the outcome of the algorithm.

To evaluate how fast the estimated conditional class probabilities of deep ReLU networks approach the true conditional class probabilities, we consider in this chapter convergence with respect to the cross-entropy (CE) loss. If the conditional class probabilities are bounded away from zero or one, the problem is related to regression and density estimation. Therefore, it seems that one could simply modify the existing proofs on convergence rates for deep ReLU networks in the regression context under the least squares loss [127, 11]. This does, however, not work since the behaviour of the CE loss differs fundamentally from that of the least squares loss for small conditional class probabilities. The risk associated with the CE loss is the expectation with respect to the input distribution of the Kullback-Leibler divergence of the conditional class probabilities. If an estimator becomes zero for one of the conditional class probabilities while the underlying conditional class probability is positive, the risk can even become infinite, see Section 2.2. In many applications where deep learning is state-of-the-art, the covariates contain nearly all information about the label and hence the conditional class probabilities are close to zero or one. For example in image classification it is often clear which object is shown on a picture. To deal with the behaviour near zero, we introduce a truncation of the CE loss function. This allows us to obtain convergence rates without bounding either the true underlying conditional class probabilities or the estimators away from zero. Instead our rates depend on an index quantifying the behaviour of the conditional class probabilities near zero. Convergence rates and the condition on the conditional class probabilities can be found in Section 2.3.

Notation: We denote vectors and vector valued functions by bold letters. For two vector valued functions $\mathbf{f} = (f_1, \ldots, f_d)$ and $\mathbf{g} = (g_1, \ldots, g_d)$ mapping $\mathcal{D}$ to $\mathbb{R}^d$, we set $\|\mathbf{f} - \mathbf{g}\|_{\mathcal{D},\infty} := \left\| \max_{j=1,\ldots,d} |f_j(\mathbf{x}) - g_j(\mathbf{x})| \right\|_{L^\infty(\mathcal{D})}$. If it is clear to which

domain $\mathcal{D}$ we refer to, we also simply write $\|\mathbf{f} - \mathbf{g}\|_\infty$. For a vector $\mathbf{v} = (v_1, \ldots, v_m)$ and a matrix $W = (W_{i,j})_{i=1,\ldots,n;j=1,\ldots,m}$ we define the maximum entry norms as $\|\mathbf{v}\|_\infty := \max_{i=1,\ldots,m} |v_i|$ and $\|W\|_\infty := \max_{i=1,\ldots,n} \max_{j=1,\ldots,m} |W_{i,j}|$. The counting 'norm' $\|\mathbf{v}\|_0, \|W\|_0$ is the number of nonzero entries in the vector $\mathbf{v}$ and matrix W, respectively. For a vector $\mathbf{v} = (v_1, \ldots, v_r)^\top$ and g a univariate function, we write $g(\mathbf{v}) := (g(v_1), \ldots, g(v_r))^\top$. We often apply this to the activation function or the logarithm $g(u) = \log(u)$. Similarly, we define for two vectors of the same length $\mathbf{v}, \mathbf{v}'$, $\log(\mathbf{v}/\mathbf{v}') = \log(\mathbf{v}) - \log(\mathbf{v}')$. For any natural number γ, we set $0\log^\gamma(0) := 0$. For a real number $x \in \mathbb{R}$, $\lfloor x \rfloor$ is the largest integer strictly smaller than x and $\lceil x \rceil$ is the smallest integer $\geq x$. A K-dimensional standard basis vector is a vector of length K that can be written as $(0, \ldots, 0, 1, 0, \ldots, 0)^\top$. We use $\mathcal{S}^K$ to denote the $(K-1)$-simplex in $\mathbb{R}^K$, that is, $\mathcal{S}^K = \{\mathbf{v} \in \mathbb{R}^K : \sum_{k=1}^K v_k = 1, v_k \geq 0, k = 1, \ldots, K\}$. For two probability measures P and Q, the Kullback-Leibler divergence $\mathrm{KL}(P, Q)$ is defined as $\mathrm{KL}(P, Q) := \int \log(dP/dQ)\,dP$ if P is dominated by Q and as $\mathrm{KL}(P, Q) := \infty$ otherwise.

2.2   The multiclass classification model

In multiclass classification with $K \geq 2$ classes and design on $[0,1]^d$, we observe a dataset $\mathcal{D}_n = \{(\mathbf{X}_i, \mathbf{Y}_i) : i = 1, \ldots, n\}$ of n i.i.d. copies of pairs $(\mathbf{X}, \mathbf{Y})$ with design/input vector $\mathbf{X}$ taking values in $[0,1]^d$ and the corresponding response vector $\mathbf{Y}$ being one of the K-dimensional standard basis vectors. The response decodes the label of the class: the output $\mathbf{Y}$ is the k-th standard basis vector if the label of the k-th class is observed. As a special case, for binary classification the output is decoded as $(1, 0)^T$ if the first class is observed and as $(0, 1)^T$ if the second class is observed. We write $\mathbb{P}$ for the joint distribution of the random vector $(\mathbf{X}, \mathbf{Y})$ and $\mathbb{P}_\mathbf{X}$ for the marginal distribution of $\mathbf{X}$. The conditional probability $\mathbb{P}_{\mathbf{Y}|\mathbf{X}}$ exists since $\mathbf{Y}$ is supported on finitely many points.

An alternative model is to assume that each of the K classes is observed roughly n/K times. To derive statistical risk bounds, there is hardly any difference and the fact that the i.i.d. model generates with small probability highly unbalanced designs will not change the analysis.

The task is now to estimate/learn from the dataset $\mathcal{D}_n$ the probability that a new input vector $\mathbf{X}$ is in class k. If $\mathbf{Y} = (Y_1, \ldots, Y_K)^\top$, the true conditional class probabilities are

$$p_k^0(\mathbf{x}) := \mathbb{P}(Y_k = 1|\mathbf{X} = \mathbf{x}), \quad k = 1, \ldots, K.$$

For any $\mathbf{x}$ this gives a probability vector, that is, $\sum_{k=1}^K p_k^0(\mathbf{x}) = 1$. For notational convenience, we also define the vector of conditional class probabilities $\mathbf{p}_0(\mathbf{x}) := (p_1^0(\mathbf{x}), \cdots, p_K^0(\mathbf{x}))^\top$.

To learn the conditional class probabilities from data, the commonly employed strategy in deep learning is to minimize the log-likelihood over the free parameters of a deep neural network using (stochastic) gradient descent. The likelihood for the conditional class probability vector $\mathbf{p}(\mathbf{x}) := (p_1(\mathbf{x}), \cdots, p_K(\mathbf{x}))^\top$ is given by

$$\mathcal{L}(\mathbf{p}|\mathcal{D}_n) = \prod_{i=1}^{n} \prod_{k=1}^{K} (p_k(\mathbf{X}_i))^{Y_{ik}},$$

with Y_{ik} the k-th entry of $\mathbf{Y}_i$. The negative log-likelihood or cross-entropy loss is then

$$\mathbf{p} \mapsto \ell(\mathbf{p}, \mathcal{D}_n) := -\frac{1}{n} \sum_{i=1}^{n} \sum_{k=1}^{K} Y_{ik} \log(p_k(\mathbf{X}_i)) = -\frac{1}{n} \sum_{i=1}^{n} \mathbf{Y}_i^\top \log(\mathbf{p}(\mathbf{X}_i)), \quad (2.2.1)$$

where the logarithm in the last expression is taken component-wise as explained in the notation section above and $\mathbf{Y}^T \log(\mathbf{p}(\mathbf{X}_i))$ is understood as the scalar product of the vectors $\mathbf{Y}$ and $\log(\mathbf{p}(\mathbf{X}_i))$. The response vectors $\mathbf{Y}_i$ are standard basis vectors and in particular have nonnegative entries. The cross-entropy loss is thus always nonnegative and consequently defines indeed a proper statistical loss function. The cross-entropy loss is also convex, but not strictly convex and thus also not strongly convex, see [149], Chapter III-B for a proof. For binary classification ($K = 2$), the cross-entropy loss coincides with the logistic loss. Throughout this chapter, we consider estimators/learners $\widehat{\mathbf{p}}(\mathbf{X})$ with the property that $\widehat{\mathbf{p}}(\mathbf{x})$ is a probability vector for all $\mathbf{x}$, or equivalently, $\widehat{\mathbf{p}}(\mathbf{x})$ lies in the simplex $\mathcal{S}^K$ for all $\mathbf{x}$. This is in particular true for neural networks with softmax activation function in the output layer. Recall that $\mathbf{p}_0(\mathbf{x})$ is the vector of true class probabilities. If $(\mathbf{X}, \mathbf{Y})$ has the same distribution as each of the observations and is independent of the dataset $\mathcal{D}_n$, the statistical estimation risk associated with the CE loss is

$$\mathbb{E}_{\mathcal{D}_n, (\mathbf{X}, \mathbf{Y})} \left[\mathbf{Y}^\top \log \left(\frac{\mathbf{p}_0(\mathbf{X})}{\widehat{\mathbf{p}}(\mathbf{X})} \right) \right] = \mathbb{E}_{\mathcal{D}_n, \mathbf{X}} \left[\mathbf{p}_0(\mathbf{X})^\top \log \left(\frac{\mathbf{p}_0(\mathbf{X})}{\widehat{\mathbf{p}}(\mathbf{X})} \right) \right]$$

$$= \mathbb{E}_{\mathcal{D}_n, \mathbf{X}} \left[\mathrm{KL} \left(\mathbf{p}_0(\mathbf{X}), \widehat{\mathbf{p}}(\mathbf{X}) \right) \right],$$

where the first equality follows from conditioning on the design vector $\mathbf{X}$ and $\mathrm{KL}(\mathbf{p}_0(\mathbf{X}), \widehat{\mathbf{p}}(\mathbf{X}))$ is understood as the Kullback-Leibler divergence of the discrete distributions with probability mass functions $\mathbf{p}_0(\mathbf{X})|\mathbf{X}$ and $\widehat{\mathbf{p}}(\mathbf{X})|(\mathbf{X}, \mathcal{D}_n)$.

(Stochastic) gradient descent methods aim to minimize the CE loss (2.2.1) over a function class $\mathcal{F}$ induced by the method. In the context of neural networks, this class is generated by all network functions with a pre-specified network architecture. In particular, the class is parametrized through the network parameters. The maximum likelihood estimator (MLE) is by definition any global minimizer of (2.2.1). For

some function classes the MLE can be given explicitly. In the extreme case that $\mathbf{x} \mapsto \mathbf{p}(\mathbf{x})$ is constraint to constant functions, the problem is equivalent to estimation of the probability vector of a multinomial distribution and the MLE is the average $\widehat{\mathbf{p}}^{\mathrm{MLE}} = \frac{1}{n} \sum_{i=1}^{n} \mathbf{Y}_i$. The other extreme is the case of training error zero. If the observed design vectors are all different, training error zero is achieved whenever there exists $\mathbf{p} \in \mathcal{F}$ such that $\mathbf{Y}_i = \mathbf{p}(\mathbf{X}_i)$ for all $i = 1, \ldots, n$. This follows from $0 \log(0) = 1 \log(1) = 0$. To achieve training error zero, we therefore need to interpolate all data points. Notice that misclassification error zero does not necessarily require interpolation of the data points.

Already for small function classes, the MLE has infinite risk if the statistical risk is as defined above. The next lemma makes this precise.

Lemma 2.2.1. *Consider binary classification ($K = 2$) with uniform design $\mathbf{X} \sim \mathrm{Unif}([0, 1]^d)$ and $\mathbf{p}_0(\mathbf{x}) := (1/2, 1/2)^{\top}$ for all $\mathbf{x} \in [0, 1]^d$. Suppose that the function class $\mathcal{F}$ contains an element $\mathbf{p}(\mathbf{x}) = (p_1(\mathbf{x}), p_2(\mathbf{x}))^{\top}$ such that $p_1(\mathbf{x}) = 0$ for all $\mathbf{x} \in [0, 1/3]^d$ and $p_1(\mathbf{x}) = 1$ for all $\mathbf{x} \in [2/3, 1]^d$. Then, there exists a MLE $\widehat{\mathbf{p}}$ with*

$$\mathbb{E}_{\mathcal{D}_n, \mathbf{X}} \left[\mathbf{p}_0(\mathbf{X})^{\top} \log \left(\frac{\mathbf{p}_0(\mathbf{X})}{\widehat{\mathbf{p}}(\mathbf{X})} \right) \right] = \infty.$$

The assumption on the function class $\mathcal{F}$ in the previous statement is quite weak and is satisfied if $\mathcal{F}$ contains all piecewise constant conditional class probabilities with at most two pieces or all piecewise linear conditional class probabilities with at most three pieces. A large statistical risk occurs also in the case of zero training error or if the estimator $\widehat{\mathbf{p}}$ severely underestimates the true probabilities.

To overcome the shortcomings of the Kullback-Leibler risk, one possibility is to regularize the Kullback-Leibler divergence and to consider for some $B > 0$ the truncated Kullback-Leibler risk

$$R_B(\mathbf{p}_0, \widehat{\mathbf{p}}) := \mathbb{E}_{\mathcal{D}_n, \mathbf{X}} \left[\mathrm{KL}_B \left(\mathbf{p}_0(\mathbf{X}), \widehat{\mathbf{p}}(\mathbf{X}) \right) \right],$$

where

$$\mathrm{KL}_B \left(\mathbf{p}_0(\mathbf{X}), \widehat{\mathbf{p}}(\mathbf{X}) \right) := \mathbf{p}_0(\mathbf{X})^{\top} \left(B \wedge \log \left(\frac{\mathbf{p}_0(\mathbf{X})}{\widehat{\mathbf{p}}(\mathbf{X})} \right) \right).$$

The loss can be shown to be nonnegative whenever $B \geq 2$, see Lemma 2.3.4 below. The threshold B becomes void if the estimator $\widehat{\mathbf{p}}$ is constrained to be in $[e^{-B}, 1]^K$. If the estimator underestimates one of the true conditional class probabilities by a large factor, the logarithm becomes large and the threshold B kicks in. For $B = \infty$, we recover the Kullback-Leibler risk.

The idea of truncation is not new. [158] truncates the log-likelihood ratio to avoid problems with this ratio becoming infinite. Their risk rates, however, are in terms of the Hellinger distance and the truncation does not appear in the statement of their results. For the truncated Kullback-Leibler risk the truncation plays a much more prominent role and appears as a multiplicative factor in the risk bounds. Lemma 2.3.4 provides insight in this difference: it shows that any upper bound for any B-truncated Kullback-Leibler divergence with $B \geq 2$ provides an upper bound for the Hellinger distance.

As we are interested in the multiclass classification problem in the context of neural networks, the function class $\mathcal{F}$ is not convex. Due to this non-convexity, the training of neural networks does typically not yield a neural network achieving the global minimum. We therefore do not assume that the estimator is the MLE and use a parameter to quantify the difference between the achieved empirical risk and the global minimum: For any estimator $\widehat{\mathbf{p}}$ taking values in a function class $\mathcal{F}$, we denote the difference between $\widehat{\mathbf{p}}$ and the global minimum of the empirical risk over that entire class by

$$\Delta_n(\mathbf{p}_0, \widehat{\mathbf{p}}) := \mathbb{E}_{\mathcal{D}_n}\Big[- \frac{1}{n} \sum_{i=1}^n \mathbf{Y}_i^\top \log(\widehat{\mathbf{p}}(\mathbf{X}_i)) - \min_{\mathbf{p} \in \mathcal{F}} -\frac{1}{n} \sum_{i=1}^n \mathbf{Y}_i^\top \log(\mathbf{p}(\mathbf{X}_i)) \Big]. \quad (2.2.2)$$

2.2.1   Deep ReLU networks

In this chapter we study deep ReLU networks with softmax output layer. Recall that the rectified linear unit (ReLU) activation function is $\sigma(x) := \max\{x, 0\}$. For any vectors $\mathbf{v} = (v_1, \cdots, v_r)^\top, \mathbf{y} = (y_1, \cdots, y_r)^\top \in \mathbb{R}^r$, write $\sigma_{\mathbf{v}}\mathbf{y} := (\sigma(y_1 - v_1), \ldots, \sigma(y_r - v_r))^\top$. To ensure that the output of the network is a probability vector over the K classes, it is standard to apply the softmax function

$$\boldsymbol{\Phi} = \left(\frac{e^{x_1}}{\sum_{j=1}^K e^{x_j}}, \ldots, \frac{e^{x_K}}{\sum_{j=1}^K e^{x_j}} \right) : \mathbb{R}^K \to \mathcal{S}^K$$

in the last layer. We use L to denote the number of hidden layers or depth of the neural network, and $\mathbf{m} = (m_0, \cdots, m_{L+1}) \in \mathbb{N}^{L+2}$ to denote the widths, that is, the number of nodes in each layer of the network. A (ReLU) network architecture with output function $\psi : \mathbb{R}^{m_{L+1}} \to \mathbb{R}^{m_{L+1}}$ is a pair $(L, \mathbf{m})_\psi$ and a network with network architecture $(L, \mathbf{m})_\psi$ is any function of the form

$$\mathbf{f} : \mathbb{R}^{m_0} \to \mathbb{R}^{m_{L+1}}, \quad \mathbf{x} \mapsto \mathbf{f}(\mathbf{x}) = \psi W_L \sigma_{\mathbf{v}_L} W_{L-1} \sigma_{\mathbf{v}_{L-1}} \cdots W_1 \sigma_{\mathbf{v}_1} W_0 \mathbf{x}, \quad (2.2.3)$$

where W_j is a $m_j \times m_{j+1}$ weight matrix and $\mathbf{v}_j \in \mathbb{R}^{m_j}$ is a shift vector. Throughout this chapter we use the convention that $\mathbf{v}_0 := (0, \ldots, 0)^\top \in \mathbb{R}^{m_0}$.

First we define neural network classes with the additional property that all network parameters are bounded in absolute value by one via

$$\mathcal{F}_{\boldsymbol{\psi}}(L, \mathbf{m}) := \left\{ \mathbf{f} \text{ is of the form of } (2.2.3) : \max_{j \in \{0, \cdots, L\}} (\|W_j\|_\infty \vee \|\mathbf{v}_j\|_\infty) \leq 1 \right\},$$

with the maximum entry norm $\| \cdot \|_\infty$ as defined in the notation section above. As in previous work, we study estimation over s-sparse ReLU networks. Those are function classes of the form

$$\mathcal{F}_{\boldsymbol{\psi}}(L, \mathbf{m}, s) := \left\{ \mathbf{f} \in \mathcal{F}(L, \mathbf{m}) : \sum_{j=0}^L \|W_j\|_0 + \|\mathbf{v}_j\|_0 \leq s \right\},$$

where the counting norm $\| \cdot \|_0$ denotes the number of nonzero vector/matrix entries.

All neural network classes in this chapter have either softmax output activation $\boldsymbol{\psi} = \boldsymbol{\Phi}$ or identity output activation $\boldsymbol{\psi} = \mathrm{id}$.

2.3 Main Results

Interesting phenomena occur if the conditional class probabilities are close to zero or one. We now introduce a notion measuring the size of the set on which the conditional class probabilities are small. The index α will later appear in the convergence rate.

Definition 2.3.1. (Small Value Bound) Let $\alpha \geq 0$ and $\mathcal{H}$ be a function class. We say that $\mathcal{H}$ is α-small value bounded (or α-SVB) if there exists a constant $C > 0$, such that for all $\mathbf{p} = (p_1, \ldots, p_K) \in \mathcal{H}$ it holds that

$$\mathbb{P}_{\mathbf{X}}(p_k(\mathbf{X}) \leq t) \leq Ct^\alpha, \quad \text{for all } t \in (0, 1] \text{ and all } k \in \{1, \ldots, K\}.$$

The condition always holds for $\alpha = 0$ and $C = 1$. If $\mathbb{P}_{\mathbf{X}}(p_k(\mathbf{X}) = 0) > 0$, the condition does not hold for $\alpha > 0$. If all functions in a class are lower bounded by a constant B_0, the class is α-SVB for any α with constant $C = B_0^{-\alpha}$. More generally, the index α is completely determined by the behaviour near zero: If for some function class there exists some $0 < \tau \ll 1$, so that the bound holds for α and for all $t \in (0, \tau]$, then replacing C by $C' = \max\{C, \tau^{-\alpha}\}$ guarantees that $C'\tau^\alpha \geq 1$, which in turn implies that the function class is α-SVB. Moreover, if a function class is α-SVB, then it is also α^*-SVB for all $\alpha^* \leq \alpha$. This follows immediately by noticing that $t^{\alpha^*} \geq t^\alpha$ for all $t \in (0, 1]$. Increasing the index makes the small value bound condition thus more restrictive.

To show that the definition of the small value bound makes sense, we have to check that for any $\alpha > 0$, there exist conditional class probabilities that are α-SVB for that

α, but are not α^*-SVB for any larger $\alpha^* > \alpha$. To see this, consider the case that X is uniformly distributed on $[0,1]$, and that there are three classes $K = 3$. For given $\alpha > 0$, define the function $\mathbf{p}_\alpha : [0,1] \to \mathcal{S}^3$ as $p_1(x) = \min\{x^{1/\alpha}, 1/3\}$, $p_2(x) = 1/3$ and $p_3(x) = 1 - p_1(x) - p_2(x) = 2/3 - \min\{x^{1/\alpha}, 1/3\}$. Since $p_2(x), p_3(x) \geq 1/3$, we have for $k = 2, 3$ that $\mathbb{P}_X(p_k(X) \leq t) \leq (3t)^\alpha$. When $k = 1$, it holds for $t \leq 1/3$ that $\mathbb{P}_X(p_1(X) \leq t) = \mathbb{P}_X(X^{1/\alpha} \leq t) = \mathbb{P}_X(X \leq t^\alpha) = t^\alpha$. Hence $\mathbb{P}_X(p_k(X) \leq t) \leq (3t)^\alpha$ for $k = 1, 2, 3$, so $\mathbf{p}_\alpha$ is α-SVB with constant 3^α. Now we show that this function is not α^*-SVB for any $\alpha^* > \alpha$. Let $\alpha^* > \alpha$, then for every constant $C > 0$, there exists a $\tau_C \in (0, 1/3)$ such that $C(\tau_C)^{\alpha^*} < (\tau_C)^\alpha = \mathbb{P}_X(p_1(X) \leq \tau_C)$. Since C is arbitrary, $\mathbf{p}_\alpha$ is not α^*-SVB.

The following example provides some insights into the relation between the conditional class probabilities and the distribution of $\mathbf{X}$. Consider the binary case $K = 2$, with input domain $[0,1]^2$, $p_1(\mathbf{x}) = (3|x_1 + x_2 - 1|^8)/4$, and $p_2(\mathbf{x}) = 1 - p_1(\mathbf{x})$, see Figure 2.1. Observe that $0 \leq p_1(\mathbf{x}) \leq 3/4$ for all $\mathbf{x} \in [0,1]^2$, so $p_1(\mathbf{x})$ and $p_2(\mathbf{x})$ indeed define conditional class probabilities. Furthermore, $p_2(\mathbf{x}) \geq 1/4$, in other words, $p_2(\mathbf{x})$ is bounded away from zero. Thus, to determine the SVB index α, it remains to consider $p_1(\mathbf{x})$. If $\mathbf{X}$ is the uniform distribution on $[0,1]^2$, Proposition 2.C.7 tells us that

$$\mathbb{P}_{\mathbf{X}}\left(p_1(\mathbf{X}) \leq t\right) = 2\left(\frac{4t}{3}\right)^{\frac{1}{8}} - \left(\frac{4t}{3}\right)^{\frac{1}{4}}$$

and hence the small value bound is satisfied for α at most $1/8$. Now suppose that instead of the uniform design, the distribution of $\mathbf{X}$ is given by the density $(x_1, x_2) \mapsto 3|x_1 + x_2 - 1|$, see Figure 2.1 for a plot. Thus, the design density is zero if $p_1(\mathbf{x})$ is zero. In this case, Proposition 2.C.7 gives

$$\mathbb{P}_{\mathbf{X}}\left(p_1(\mathbf{X}) \leq t\right) = 3\left(\frac{4t}{3}\right)^{\frac{1}{4}} - 2\left(\frac{4t}{3}\right)^{\frac{3}{8}},$$

and the SVB index α is at most $1/4$.

The following theorem shows the influence of the index α in the small value bound on the approximation rates.

Theorem 2.3.2. *If the function class is α-SVB with constant C, then, for any approximating function $\mathbf{p} = (p_1, \ldots, p_k) : [0,1]^d \to \mathcal{S}^K$ satisfying $\|\mathbf{p} - \mathbf{p}_0\|_\infty \leq C_1/M$, and $\min_k \inf_{\mathbf{x} \in [0,1]^d} p_k(\mathbf{x}) \geq 1/M$, for some constant C_1, it holds that*

$$\mathbb{E}_{\mathbf{X}}\left[(\mathbf{p}_0(\mathbf{X}))^\top \log\left(\frac{\mathbf{p}_0(\mathbf{X})}{\mathbf{p}(\mathbf{X})}\right)\right] \leq CK\frac{(C_1 + 1)^{2+(\alpha \wedge 1)}}{M^{1+(\alpha \wedge 1)}}\left(1 + \frac{\mathbb{1}_{\{\alpha < 1\}}}{1 - \alpha} + \log(M)\right).$$

The proof for this result bounds the Kullback-Leibler divergence by the χ^2-divergence and then distinguishes the cases where the conditional class probabilities are

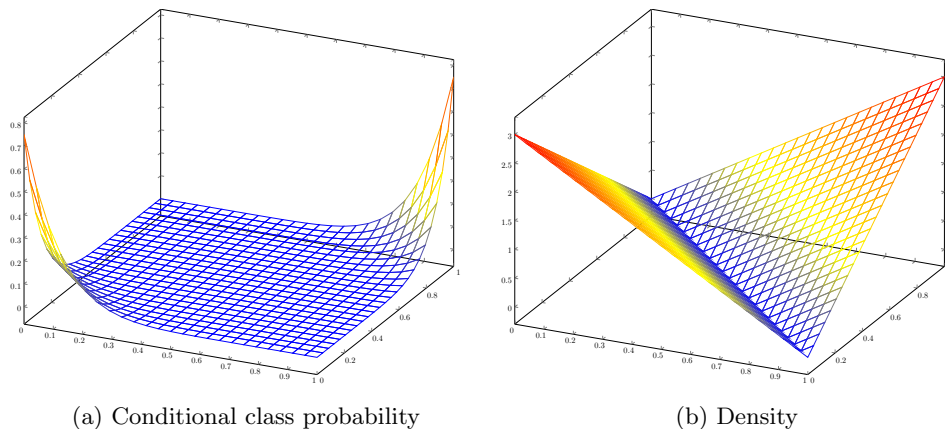(a) Conditional class probability   (b) Density

Figure 2.1: Plot of the conditional class probability $p_1(\mathbf{x}) = (3|x_1 + x_2 - 1|^8)/4$ on the left and of the density $(x_1, x_2) \mapsto 3|x_1 + x_2 - 1|$ on the right.

smaller and larger than $1/M$. Both terms can be controlled via the α-SVB condition. The convergence rate becomes faster in M up to $\alpha = 1$ and is $\log(M)/M^2$ for all $\alpha \geq 1$.

The small value bound provides a flexible framework that allows the conditional class probabilities to be close to zero and therefore generalizes the standard assumption in the nonparametric classification literature that the conditional class probabilities are bounded away from zero. Here, we argue that the regime of small conditional class probabilities is of particular relevance for classification tasks where most of the information about the class label is contained in the covariates. Indeed, if $\mathbf{X}$ contains all information about the class label Y, then $Y|\mathbf{X}$ is deterministic and the conditional class probability is either zero or one. On the contrary, in situations where the covariates/input variable $\mathbf{X}$ does not contain the full information about the class label, $Y|\mathbf{X}$ is random, and the conditional class probabilities are bounded away from zero or one. The case of small conditional class probabilities corresponds to a scenario where the covariates contain most of the information about the class label. These are classification tasks for which small misclassification errors can be achieved, but perfect classification is impossible. This is also the regime for which the SVB index α should be strictly larger than zero. For instance, for the widely used Breast Cancer Wisconsin (Diagnostic) dataset and Heart Disease dataset from the UCI machine learning repository [40] the covariates do not contain the full relevant information about the disease but small misclassification can be achieved. It is therefore conceivable

that these are prototypical examples for the case $\alpha > 0$.

The small value bound has a similar flavor as Tsybakov's margin condition, which can be stated as $\mathbb{P}_{\mathbf{X}}(0 < |p_0(\mathbf{X}) - 1/2| \leq t) \leq Ct^{\gamma}$ for binary classification [5]. The margin condition provides a control on the number of data points that are close to the decision boundary $\{\mathbf{x} : p_0(\mathbf{x}) = 1/2\}$ and that are therefore hard to classify correctly. Differently speaking, the problem becomes easier if the conditional class probabilities are either close to zero or one. This is in contrast with the small value bound, which will lead to faster convergence rates when the true conditional class probabilities are mostly away from zero. This difference is due to the loss: the 0-1 loss only cares about predicting the class membership, while the CE loss measures how well the conditional class probabilities are estimated and puts additional emphasis on small conditional class probabilities by considering the ratio between prediction and truth.

To obtain estimation rates, we further assume that the underlying true conditional class probability function $\mathbf{p}_0$ belongs to the class of Hölder-smooth functions. For $\beta > 0$ and $D \subset \mathbb{R}^m$, the ball of β-Hölder functions with radius Q is defined as

$$
C^{\beta}(D, Q) := \left\{ f : D \to \mathbb{R} : \right.
$$

$$
\left. \sum_{\boldsymbol{\gamma} : \|\boldsymbol{\gamma}\|_1 < \beta} \|\partial^{\boldsymbol{\gamma}} f\|_{\infty} + \sum_{\boldsymbol{\gamma} : \|\boldsymbol{\gamma}\|_1 = \lfloor \beta \rfloor} \sup_{\mathbf{x}, \mathbf{y} \in D, \mathbf{x} \neq \mathbf{y}} \frac{|\partial^{\boldsymbol{\gamma}} f(\mathbf{x}) - \partial^{\boldsymbol{\gamma}} f(\mathbf{y})|}{\|\mathbf{x} - \mathbf{y}\|_{\infty}^{\beta - \lfloor \beta \rfloor}} \leq Q \right\},
$$

where $\partial^{\boldsymbol{\gamma}} = \partial^{\gamma_1} \dots \partial^{\gamma_m}$, with $\boldsymbol{\gamma} = (\gamma_1, \dots, \gamma_m) \in \mathbb{N}^m$. The function class $\mathcal{G}(\beta, Q)$ of β-smooth conditional class probabilities is then defined as

$$
\mathcal{G}(\beta, Q) = \left\{ \mathbf{p} = (p_1, \cdots, p_K)^{\top} : [0, 1]^d \to \mathcal{S}^K : \right.
$$

$$
\left. p_k \in C^{\beta}([0, 1]^d, Q), k = 1, \dots, K \right\}.
$$

If $Q < 1/K$, then, $\|\mathbf{p}\|_{\infty} \leq Q$ implies $\sum_{k=1}^{K} p_k \leq KQ < 1$, so we need Hölder radius $Q \geq 1/K$ for this class to be non-empty. Combining the smoothness and the small value bound, we write $\mathcal{G}_{\alpha}(\beta, Q) = \mathcal{G}_{\alpha}(\beta, Q, C)$ for all functions in $\mathcal{G}(\beta, Q)$ that satisfy the α-SVB condition with constant C. For large enough radius Q and constant C, the class $\mathcal{G}_{\alpha}(\beta, Q)$ is non-empty. For example, the constant function $\mathbf{p} = (1/K, \dots, 1/K)$ is in $\mathcal{G}_{\alpha}(\beta, Q)$ for any $\beta > 0$ and $\alpha > 0$ when $Q \geq 1/K$ and $C \geq K^{\alpha}$.

For $0 \leq \alpha \leq 1$ the index from the SVB condition and β the smoothness index, we introduce the rate

$$
\phi_n = K^{\frac{(1+\alpha)\beta + (3+\alpha)d}{(1+\alpha)\beta + d}} n^{-\frac{(1+\alpha)\beta}{(1+\alpha)\beta + d}}.
$$

Theorem 2.3.3 (Main Risk Bound). *Consider the multiclass classification model with* $\mathbf{p}_0 \in \mathcal{G}_\alpha(\beta, Q)$, $0 \leq \alpha \leq 1$, *and* $n > 1$. *Let* $\widehat{\mathbf{p}}$ *be an estimator taking values in the network class* $\mathcal{F}_{\mathbf{\Phi}}(L, \mathbf{m}, s)$ *satisfying*

 (i) $A(d, \beta) \log_2(n) \leq L \lesssim n\phi_n$,
 (ii) $\min_{i=1,\cdots,L} m_i \gtrsim n\phi_n$,
 (iii) $s \asymp n\phi_n \log(n)$

for a suitable constant $A(d, \beta)$. *If* n *is sufficiently large, then, there exist constants* C', C'' *only depending on* α, C, β, d, *such that whenever* $\Delta_n(\widehat{\mathbf{p}}, \mathbf{p}_0) \leq C'' B\phi_n L \log^2(n)$ *then*

$$R_B(\mathbf{p}_0, \widehat{\mathbf{p}}) \leq C' B\phi_n L \log^2(n).$$

An explicit expression for the constant $A(d, \beta)$ can be derived from the proof. The risk bound depends linearly on B. Choosing, for instance, $B = O(\log(n))$ leads only to an additional logarithmic factor in the convergence rate. The risk bound grows with $K^{\frac{(1+\alpha)\beta+(3+\alpha)d}{(1+\alpha)\beta+d}}$ in the number of classes. Thus for large β, we obtain a near linear dependence on K. The worst behavior occurs for $\alpha = 1$ and d large. Then the dependence on the number of classes is essentially of the order K^4.

When the estimator $\widehat{\mathbf{p}}$ is guaranteed to have output in $[e^{-B}, 1]^K$, the truncation parameter B in the risk has no effect. The proof of the approximation properties is done by the construction of a softmax-network $\widehat{\mathbf{g}}$ with the property that $\widehat{\mathbf{g}}(\mathbf{x}) \gtrsim K^{\frac{-(2+\alpha)\beta}{(1+\alpha)\beta+d}} n^{-\frac{\beta}{(1+\alpha)\beta+d}}$, for all $\mathbf{x} \in [0, 1]^d$. This means that we can pick $B \asymp \log(n)$ such that $\widehat{\mathbf{g}}(\mathbf{x}) \geq e^{-B}$ and restrict the class $\mathcal{F}_{\mathbf{\Phi}}(L, \mathbf{m}, s)$ to networks that are guaranteed to have output in $[e^{-B}, 1]^K$. The proof of Theorem 2.3.3 can be extended for this setting and implies a risk bound for the Kullback-Leibler risk of the form

$$\mathbb{E}_{\mathcal{D}_n, \mathbf{X}} \big[\mathrm{KL}\big(\mathbf{p}_0(\mathbf{X}), \widehat{\mathbf{p}}(\mathbf{X})\big) \big] \leq C''' \phi_n L \log^3(n),$$

for some constant C'''. Thus Theorem 2.3.3 provides us with rates for the Kullback-Leibler risk when the networks outputs are guaranteed to be sufficiently large, while still providing a bound for the truncated Kullback-Leibler risk when no such guarantee can be given.

When the input dimension d is large, the obtained convergence rates become slow. A possibility to circumvent this curse of dimensionality is to assume additional structure on $\mathbf{p}_0$. For nonparametric regression, [62, 72, 11, 127, 75] show that under a composition assumption on the regression function, neural networks can exploit this structure to obtain fast convergence rates that are unaffected by the curse of dimensionality. It is conceivable that for various classification problems, such an underlying composition structure is present. For instance to classify an email as spam, the hierarchical structure is important and decision trees that are adapted to such structures work well, see Section 9.2.5 in [56]. In image classification it is often

assumed that an image can be constructed from compositions of simpler features; for example a square is built from lines and can itself be used as component of more complicated shapes.

It is possible to incorporate a composition assumption on the conditional class probabilities within the considered framework. As our approximation result already depends on Theorem 5 of [127] it is relatively straightforward to sketch how the additional composition assumption can help to deal with the curse of dimensionality. Consider the class of functions that satisfy the composition assumption in [127]

$$\mathcal{G}_{\text{comp}}(r, \mathbf{d}, \mathbf{t}, \boldsymbol{\beta}, Q) :=$$
$$\left\{ f = \mathbf{g}_r \circ \cdots \circ \mathbf{g}_0 : \mathbf{g}_i = (g_{ij})_j : [a_i, b_i]^{d_i} \to [a_{i+1}, b_{i+1}]^{d_{i+1}}, \right.$$
$$\left. g_{ij} \in \mathcal{C}^{\beta_i}([a_i, b_i]^{t_i}, Q) \text{ for some } |a_i|, |b_i| \le Q \right\}.$$

Here t_i is the maximal number of variables on which each of the component functions g_{ij} may depend on. For specific structural assumptions, such as generalized additive models and sparse tensor decompositions, t_i can be much smaller than the input dimension d, [127].

In our setting the composition constraint can be incorporated by assuming that each of the conditional class probabilities $p_1, \ldots, p_K$ lies in the class $\mathcal{G}_{\text{comp}}(r, \mathbf{d}, \mathbf{t}, \boldsymbol{\beta}, Q)$. Define the effective smoothness indices as $\beta_i^* := \beta_i \prod_{\ell=i+1}^r (\beta_\ell \wedge 1)$. By approximating these composition functions by neural networks as in the proof of Theorem 1 of [127] in place of Theorem 5 of the same article, one can then obtain the rate

$$\phi_n = \max_{i=0,\cdots,r} K^{\frac{(1+\alpha)\beta_i^* + (3+\alpha)t_i}{(1+\alpha)\beta_i^* t_i}} n^{-\frac{(1+\alpha)\beta_i^*}{(1+\alpha)\beta_i^* + t_i}}.$$

Let us briefly summarize the related literature. Convergence rates for neural networks in (binary) classification have recently been studied in [67, 74, 76, 134, 111] in various settings. [67] derives convergence rates for the $0-1$ loss based on different surrogate losses and assumptions. For the hinge loss as surrogate loss, the margin condition in combination with smoothness conditions on the decision boundary as well as smoothness conditions on the conditional class probabilities are studied. Moreover, the logistic loss is analyzed under a condition that requires the conditional class probabilities to be near zero or one combined with smoothness conditions on the decision boundary. Convergence rates for the $0-1$ loss for convolutional neural networks are studied in [74, 76]. Both papers assume smoothness conditions on the conditional class probabilities and impose a max-pooling structure assumption for the conditional class probability that is related to the structure of convolutional networks. In [74] the least squares loss is used as a surrogate loss, while [76] uses the logistic loss as surrogate loss. More recently, [134] studied the convergence rates for convex Lipschitz

losses of convolutional neural networks in binary classification under a submanifold condition. The framework includes least squares loss, hinge loss, truncated logistic loss and truncated exponential loss. In the truncated cases, the minimizers are also truncated. Furthermore, [111] studies convergence rates for the $0-1$ loss with the hinge loss as surrogate loss, in the case that the model is deterministic and that the decision boundary is Barron regular.

2.3.1 Relationship with Hellinger distance

The multiclass classification problem can be written as statistical model ($Q_{\mathbf{p}}, \mathbf{p} \in \mathcal{F}$), where $\mathcal{F}$ is the parameter space, $\mathbf{p}$ is the unknown vector of conditional class probabilities and $Q_{\mathbf{p}}$ denotes the data distribution if the data are generated from the conditional class probabilities $\mathbf{p}$. The squared Hellinger distance $H(P,Q)^2 = \frac{1}{2}\int(\sqrt{dP} - \sqrt{dQ})^2$, with P and Q probability measures on the same probability space, induces in a natural way a loss function on such a statistical model by associating to the two parameters $\mathbf{p}$ and $\mathbf{p}'$ the loss $H(Q_{\mathbf{p}}, Q_{\mathbf{p}'})$. The Hellinger loss function has been widely studied in the context of nonparametric variations of the maximum likelihood principle, mainly for the related nonparametric density estimation problem, [158, 147, 148]. The log-likelihood is closely related to the Kullback-Leibler divergence, which in turn is related to the Hellinger distance by the inequality $H(P,Q)^2 \leq \mathrm{KL}(P,Q)$, see for example [146]. The Kullback-Leibler divergence cannot be upper bounded by the squared Hellinger distance in general, although there exists conditions under which such a bound can be established, see for example Theorem 5 of [158] and Lemma 2.3.4 below.

In density estimation, the nonparametric MLE achieves in some regimes optimal rates with respect to the Hellinger distance for convex estimator classes or if the densities (or sieve estimators) are uniformly bounded away from zero, see [157, 158] and Chapters 7 and 10 in [147]. Neural network function classes are not convex and, as argued before, there are many applications in the deep learning literature, where the conditional class probabilities are very small or even zero. Thus, these general results are not applicable in our setting.

On the contrary, the convergence rates established above for the truncated Kullback-Leibler divergence imply convergence with respect to the Hellinger loss. This relationship is made precise in the next result.

Lemma 2.3.4. *Let P and Q be two probability measures defined on the same measurable space. For any $B \geq 2$,*

$$H^2(P,Q) \leq \frac{1}{2}\mathrm{KL}_2(P,Q) \leq \frac{1}{2}\mathrm{KL}_B(P,Q) \leq 2e^{B/2}H^2(P,Q).$$

For the proof see Appendix 2.C. The upper bound on the truncated Kullback-Leibler divergence is related to the inequalities that bound the Kullback-Leibler divergence by the squared Hellinger distance under the assumption of a bounded likelihood ratio, such as (7.6) in [20] or Lemma 4 in [57].

Combining the previous lemma and Theorem 2.3.3 with $B = 2$ gives

$$\mathbb{E}_{\mathcal{D}_n} \Big[\int_{[0,1]^d} \sum_{j=1}^{K} \Big(\sqrt{p_j^0(\mathbf{x})} - \sqrt{\widehat{p}_j(\mathbf{x})} \Big)^2 \, d\mathbb{P}_{\mathbf{X}}(\mathbf{x}) \Big] \leq 2C' \phi_n L \log^2(n), \qquad (2.3.1)$$

whenever $\Delta_n(\widehat{\mathbf{p}}, \mathbf{p}_0) \leq C'' B \phi_n L \log^2(n)$.

We can also use the relation with the Hellinger distance to show that for $\alpha = 1$, we obtain a near minimax optimal convergence rate. Indeed $n^{-\frac{2\beta}{2\beta+d}}$ is the optimal rate for the squared Hellinger distance. For references see for instance Example 7.4.1 of [147] for univariate densities bounded away from zero; the entropy bounds in Theorem 2.7.1. together with Proposition 1 of [159] for densities bounded away from zero; or the entropy bounds in Theorem 2.7.1. and Equation (3.4.5) of [148] together with Chapter 2.3. of [159] for densities p for which $\int \frac{1}{p}$ is bounded. Since the squared Hellinger distance can be upper bounded by the Kullback-Leibler divergence, the rate $n^{-\frac{2\beta}{2\beta+d}}$ is also a lower bound for the Kullback-Leibler risk. Since this rate is achieved for $\alpha = 1$, it is clear that no further gain in the convergence rate can be expected for $\alpha > 1$. For $\alpha \geq 1$, the rate of convergence is up to $\log(n)$-factors the same as in Theorem 5 of [158] and also the conditions are comparable.

It is instructive to relate the global convergence rates to pointwise convergence. Recall that for real numbers a, b, we have $(\sqrt{a} - \sqrt{b})^2 = (a - b)^2 / (\sqrt{a} + \sqrt{b})^2$. If $\mathbb{P}_{\mathbf{X}}$ has a Lebesgue density that is bounded on $[0,1]^d$ from below and above and if we choose L of the order $O(\log n)$, (2.3.1) indicates that on a large subset of $[0,1]^d$, we can expect a pointwise distance

$$\Big| p_j^0(\mathbf{x}) - \widehat{p}_j(\mathbf{x}) \Big| \lesssim \Big| \sqrt{p_j^0(\mathbf{x})} + \sqrt{\widehat{p}_j(\mathbf{x})} \Big| K^{\frac{(1+\alpha/2)d}{(1+\alpha)\beta+d}} n^{-\frac{(1+\alpha)\beta}{2(1+\alpha)\beta+2d}} \log^{3/2}(n).$$

The pointwise convergence rate gets therefore faster if the conditional class probabilities are small. In the most extreme case, $p_j^0(\mathbf{x}) = 0$, the previous bound becomes

$$\Big| p_j^0(\mathbf{x}) - \widehat{p}_j(\mathbf{x}) \Big| \lesssim K^{\frac{(2+\alpha)d}{(1+\alpha)\beta+d}} n^{-\frac{(1+\alpha)\beta}{(1+\alpha)\beta+d}} \log^3(n).$$

Since $n^{-(1+\alpha)\beta/((1+\alpha)\beta+d)} \ll n^{-\beta/(2\beta+d)}$, this rate can be much faster than the classical nonparametric rate for pointwise estimation $n^{-\beta/(2\beta+d)}$. The gain gets accentuated as the index α increases. A large index α in the SVB bound can be chosen if the conditional class probabilities are rarely small or zero. Hence there is a trade-off and the regions on which a faster rate can be obtained are thus smaller.

2.3.2   Oracle Inequality

The risk bound of Theorem 2.3.3 relies on an oracle-type inequality. Before we can state this inequality we first need some definitions. Given a function class of conditional class probabilities $\mathcal{F}$, we denote by $\log(\mathcal{F})$ the function class containing all functions that can be obtained by applying the logarithm coefficient-wise to functions from $\mathcal{F}$, that is,

$$\log(\mathcal{F}) = \big\{\mathbf{g} = \log(\mathbf{f}) : \mathbf{f} \in \mathcal{F}\big\}.$$

Next we define a family of pseudometrics. Recall that a pseudometric is a metric without the condition that $d(f, g) = 0$ implies $f = g$. For a real number τ and $\mathbf{f}, \mathbf{g} : \mathcal{D} \to \mathbb{R}^K$, set

$$d_\tau(\mathbf{f}, \mathbf{g}) := \sup_{\mathbf{x} \in \mathcal{D}} \max_{k=1,\cdots,K} |(\tau \vee f_k(\mathbf{x})) - (\tau \vee g_k(\mathbf{x}))|.$$

Lemma 2.C.3 in the appendix verifies that this indeed defines a pseudometric. For $\tau = -\infty$, $d_\tau(\mathbf{f}, \mathbf{g})$ coincides with the L^∞-norm as defined in the notation section.

Denote by $\mathcal{N}(\delta, \mathcal{F}, d(\cdot, \cdot))$ the δ interior covering number of a function class $\mathcal{F}$ with respect to a (pseudo)metric $d(., .)$. For interior coverings, the centers of the balls of any cover are required to be inside the function class $\mathcal{F}$. Triangle inequality shows that any (exterior) δ-cover can be used to construct an interior cover with the same number of balls, but with radius 2δ instead of δ.

Theorem 2.3.5 (Oracle Inequality)**.** *Let $\mathcal{F}$ be a class of conditional class probabilities and $\widehat{\mathbf{p}}$ be any estimator taking values in $\mathcal{F}$. If $B \geq 2$ and $\mathcal{N}_n = \mathcal{N}(\delta, \log(\mathcal{F}), d_\tau(\cdot, \cdot)) \geq 3$ for $\tau = \log(C_n e^{-B}/n)$, then*

$$R_B(\mathbf{p}_0, \widehat{\mathbf{p}}) \leq (1 + \epsilon)\left(\inf_{\mathbf{p} \in \mathcal{F}} R(\mathbf{p}_0, \mathbf{p}) + \Delta_n(\mathbf{p}_0, \widehat{\mathbf{p}}) + 3\delta\right)$$

$$+ \frac{(1+\epsilon)^2}{\epsilon} \cdot \frac{68B \log(\mathcal{N}_n) + 272B + (3/2)C_n K\big(\log\big(\frac{n}{C_n}\big) + B\big)}{n},$$

for all $\delta, \epsilon \in (0, 1]$, $0 < C_n \leq ne^{-1}$ and $\Delta_n(\mathbf{p}_0, \widehat{\mathbf{p}})$ as defined in (2.2.2).

The proof of this oracle inequality is a non-trivial variation of the proof for the oracle inequality in the regression model [127]. The statement seems to suggest to pick a small C_n. Then, however, also τ will be small, and d_τ becomes a stronger metric possibly leading to an increase of the covering number $\mathcal{N}_n$.

We can also replace the covering number of $\log(\mathcal{F})$ by the covering number of $\mathcal{F}$ in the oracle inequality:

Corollary 2.3.6. *Denote* $\widetilde{\mathcal{N}}_n := \mathcal{N}(\delta C_n e^{-B}/n, \mathcal{F}, d_\tau(\cdot, \cdot))$, *with* $\tau = C_n e^{-B}/n$. *Under the conditions of Theorem 2.3.5, it holds that*

$$R_B(\mathbf{p}_0, \widehat{\mathbf{p}}) \leq (1+\epsilon)\left(\inf_{\mathbf{p} \in \mathcal{F}} R(\mathbf{p}_0, \mathbf{p}) + \Delta_n(\mathbf{p}_0, \widehat{\mathbf{p}}) + 3\delta\right)$$
$$+ \frac{(1+\epsilon)^2}{\epsilon} \cdot \frac{68B\log(\widetilde{\mathcal{N}}_n) + 272B + (3/2)C_n K(\log(n/C_n) + B)}{n},$$

for all $\delta, \epsilon \in (0,1]$, $0 < C_n \leq ne^{-1}$ *and* $\Delta_n(\mathbf{p}_0, \widehat{\mathbf{p}})$ *as defined in (2.2.2).*

Let us briefly discuss some ideas underlying the proof of the oracle inequality. For simplicity, assume that $\widehat{\mathbf{p}}$ is the MLE over a class $\mathcal{F}$ and that $\mathbf{p}_0 \in \mathcal{F}$. By the definition of the MLE $\widehat{\mathbf{p}}$, we have that $-\frac{1}{n}\sum_{i=1}^n \mathbf{Y}_i^\top \log(\widehat{\mathbf{p}}(\mathbf{X}_i)) \leq -\frac{1}{n}\sum_{i=1}^n \mathbf{Y}_i^\top \log(\mathbf{p}_0(\mathbf{X}_i))$. Taking expectation on both sides, one can then show that for any $B \geq 0$,

$$\mathbb{E}_{\mathcal{D}_n}\left[\frac{1}{n}\sum_{i=1}^n \mathbf{p}_0(\mathbf{X}_i)^\top\left(B \wedge \log\left(\frac{\mathbf{p}_0(\mathbf{X}_i)}{\widehat{\mathbf{p}}(\mathbf{X}_i)}\right)\right)\right]$$
$$\leq \mathbb{E}_{\mathcal{D}_n}\left[\frac{1}{n}\sum_{i=1}^n \left(\mathbf{p}_0(\mathbf{X}_i) - \mathbf{Y}_i\right)^\top\left(B \wedge \log\left(\frac{\mathbf{p}_0(\mathbf{X}_i)}{\widehat{\mathbf{p}}(\mathbf{X}_i)}\right)\right)\right].$$

Using standard empirical process arguments, the right hand side can be roughly upper bounded by $\mathbb{E}_{\mathcal{D}_n}[\max_j \frac{1}{n}\sum_{i=1}^n(\mathbf{p}_0(\mathbf{X}_i) - \mathbf{Y}_i)^\top(B \wedge \log(\mathbf{p}_0(\mathbf{X}_i)/\mathbf{p}_j(\mathbf{X}_i)))]$, where the maximum is over all centers of an ε-covering of $\mathcal{F}$ for a sufficiently small ε. Since $\mathbb{E}_{\mathcal{D}_n}[\mathbf{Y}_i|\mathbf{X}_i] = \mathbf{p}_0(\mathbf{X}_i)$, this is the maximum over a centered process. Using empirical process theory a second time, the left hand side of the previous display can be shown to converge to the statistical risk $R_B(\mathbf{p}_0, \widehat{\mathbf{p}}) = \mathbb{E}_{\mathcal{D}_n, \mathbf{X}}[\mathrm{KL}_B(\mathbf{p}_0(\mathbf{X}), \widehat{\mathbf{p}}(\mathbf{X}))]$.

To apply Bernstein's inequality we need to bound the moments of the random variables in the empirical process. For that we have derived the following inequality that relates the m-th moment to the truncated Kullback-Leibler divergence and also shows the effect of the truncation level B.

Lemma 2.3.7. *If* $B > 1$ *and* $m = 2, 3, \ldots$, *then, for any two probability vectors* $(p_1, \ldots, p_K)$ *and* $(q_1, \ldots, q_K)$, *we have*

$$\sum_{k=1}^K p_k \left|B \wedge \log\left(\frac{p_k}{q_k}\right)\right|^m \leq \max\left\{m!, \frac{B^m}{B-1}\right\}\sum_{k=1}^K p_k\left(B \wedge \log\left(\frac{p_k}{q_k}\right)\right).$$

In order to use the oracle inequality for deep ReLU networks with softmax activation in the output layer, we now state a bound on the covering number of these classes. The bound and its proof are a slight modification of Lemma 5 in [127].

Lemma 2.3.8. *If $V := \prod_{\ell=0}^{L+1}(m_\ell + 1)$, then for every $\delta > 0$,*

$$\mathcal{N}\left(\delta, \log(\mathcal{F}_{\boldsymbol{\Phi}}(L, \mathbf{m}, s)), \|\cdot\|_\infty\right) \leq \left(4\delta^{-1}K(L+1)V^2\right)^{s+1},$$

and

$$\log \mathcal{N}\left(\delta, \log(\mathcal{F}_{\boldsymbol{\Phi}}(L, \mathbf{m}, s)), \|\cdot\|_\infty\right) \leq (s+1)\log(2^{2L+6}\delta^{-1}(L+1)K^3 d^2 s^L).$$

The second bound follows from the first by removing inactive nodes, Proposition 2.A.1, and taking the logarithm. The full proof can be found in Appendix 2.C.

The proof of the main risk bound in Theorem 2.3.3 is based on the oracle inequality derived above. To bound the individual error terms, we apply the approximation theory developed in Theorem 2.3.2 and Lemma 2.4.3 as well as the previous bound on the metric entropy. This shows that for any $M > 1$, the truncated Kullback-Leibler risk for a network class with depth L, width $\gtrsim KM^{d/\beta}$ and sparsity $s \lesssim KM^{d/\beta}$ can be bounded by

$$R_B(\mathbf{p}_0, \widehat{\mathbf{p}}) \lesssim K^{3+\alpha}\frac{\log(M)}{M^{1+\alpha}} + KM^{d/\beta}L\frac{\log^2(n)}{n} + \Delta_n(\widehat{\mathbf{p}}, \mathbf{p}_0).$$

Balancing the terms $K^{3+\alpha}/M^{1+\alpha}$ and $KM^{d/\beta}$ leads to $M \asymp K^{\frac{(2+\alpha)\beta}{(1+\alpha)\beta+d}} n^{\frac{\beta}{(1+\alpha)\beta+d}}$ and for small $\Delta_n(\widehat{\mathbf{p}}, \mathbf{p}_0)$, we get the rate

$$R_B(\mathbf{p}_0, \widehat{\mathbf{p}}) \lesssim K^{\frac{(1+\alpha)\beta+(3+\alpha)d}{(1+\alpha)\beta+d}} n^{-\frac{(1+\alpha)\beta}{(1+\alpha)\beta+d}} L\log^2(n)$$

in Theorem 2.3.3.

2.4 Proofs

Proof of Lemma 2.2.1. Consider the event $\mathcal{A}_n := \{(\mathbf{X}_i, \mathbf{Y}_i) \in ([0, 1/3]^d \times (1, 0)^\top) \cup ([2/3, 1]^d \times (0, 1)^\top), \text{ for all } i = 1, \ldots, n\}$. Recall that $0\log(0) = 0$. On the event $\mathcal{A}_n$, for any $\mathbf{p}(\mathbf{x}) = (p_1(\mathbf{x}), p_2(\mathbf{x}))^\top$ such that $p_1(\mathbf{x}) = 0$ for all $\mathbf{x} \in [0, 1/3]^d$ and $p_1(\mathbf{x}) = 1$ for all $\mathbf{x} \in [2/3, 1]^d$, we have that $\ell(\mathbf{p}, \mathcal{D}_n) = 0$, where $\ell(\mathbf{p}, \mathcal{D}_n)$ is the negative log-likelihood as defined in (2.2.1). Since the CE loss is nonnegative, any such $\mathbf{p}$ in the class $\mathcal{F}$ is a MLE on this event. Since $\mathbb{P}(\mathcal{A}_n) > 0$, it follows that

$$\mathbb{E}_{\mathcal{D}_n, \mathbf{X}}\left[\mathbf{p}_0(\mathbf{X})^\top \log\left(\frac{\mathbf{p}_0(\mathbf{X})}{\widehat{\mathbf{p}}(\mathbf{X})}\right)\right] \geq \mathbb{E}_{\mathcal{D}_n}\left[\mathbf{1}(\mathcal{A}_n)\int_{[0,1]^d} \mathbf{p}_0(\mathbf{u})^\top \log\left(\frac{\mathbf{p}_0(\mathbf{u})}{\widehat{\mathbf{p}}(\mathbf{u})}\right)d\mathbf{u}\right]$$
$$= \infty \cdot \mathbb{P}(\mathcal{A}_n)$$
$$= \infty.$$

$\square$

2.4.1   Approximation related results

This section is devoted to the proof of Theorem 2.3.3. First we construct a neural network that approximates $\mathbf{p}_0$ in terms of the L_∞-norm and is bounded away from zero. Afterwards we prove Theorem 2.3.2 relating the previously derived approximation theory to a bound on the approximation error in terms of the expected Kullback-Leibler divergence. We finish the proof combining this network with the new oracle inequality (Theorem 2.3.5) and an entropy bound for classes of neural networks with a softmax function in the output layer, Lemma 2.3.8. Recall that $\mathcal{F}_{\mathrm{id}}(L, \mathbf{m}, s)$ denotes the neural network class with L hidden layers, width vector $\mathbf{m}$, network sparsity s and identity activation function in the output layer.

Theorem 2.4.1. *For all $M \geq 2$ and $\beta > 0$ there exists a neural network $G \in \mathcal{F}_{\mathrm{id}}(L, \mathbf{m}, s)$, with*
   (i) $L = \lfloor 40(\beta + 2)^2 \log_2(M) \rfloor$,
   (ii) $\mathbf{m} = (1, \lfloor 48\lceil\beta\rceil^3 2^\beta M^{1/\beta} \rfloor, \cdots, \lfloor 48\lceil\beta\rceil^3 2^\beta M^{1/\beta} \rfloor, 1)$,
   (iii) $s \leq 4284(\beta + 2)^5 2^\beta M^{1/\beta} \log_2(M)$,
such that for any $x \in [0, 1]$,

$$\left| e^{G(x)} - x \right| \leq \frac{4}{M} \quad and \quad G(x) \geq \log\left(\frac{4}{M}\right).$$

The proof of this theorem can be found in Appendix 2.B. To approximate Hölder functions we use Theorem 5 from [127] with m equal to $\lceil \log_2(M))(d/\beta + 1) \rceil$. We state here a variation of that theorem in our notation using weaker upper bounds to simplify the expressions for the network size. These upper bounds can be deduced directly from the depth-synchronization and network enlarging properties of neural networks stated in Section 2.A.1. Set

$$C_{Q,\beta,d} := (2Q + 1)(1 + d^2 + \beta^2)6^d + Q3^\beta.$$

Theorem 2.4.2. *For every function $\mathbf{f} \in \mathcal{G}(\beta, Q)$ and every $M > (\beta+1)^\beta \vee (Q+1)^{\beta/d} e^\beta$, there exist neural networks $H_k \in \mathcal{F}_{\mathrm{id}}(L, \mathbf{m}, s)$ with*
   (i) $L = 3\lceil \log_2(M)(d/\beta + 1) \rceil(1 + \lceil \log_2(d \vee \beta) \rceil)$,
   (ii) $\mathbf{m} = (d, 6(d + \lceil\beta\rceil)\lfloor M^{d/\beta} \rfloor, \cdots, 6(d + \lceil\beta\rceil)\lfloor M^{d/\beta} \rfloor, 1)$,
   (iii) $s \leq 423(d + \beta + 1)^{3+d} M^{d/\beta} \log_2(M))(d/\beta + 1)$,
such that
$$\left\| H_k - f_k^0 \right\|_\infty \leq \frac{C_{Q,\beta,d}}{M}, \quad \forall k \in \{1, \cdots, K\}.$$

Here the M is chosen such that $M^{d/\beta} \asymp N$, where N is as defined in Theorem 5 of [127].

Without loss of generality we can assume that the output of the H_k networks lies in $[0,1]$. Indeed if this would not be the case, then the projection-layer that we use later on in our proof will guarantee that it is in this interval. This will not increase the error since the functions f_k^0 only take values in $[0,1]$.

To obtain a neural network with softmax output, the next lemma combines the neural network constructions from the previous two theorems and replaces the output with a softmax function.

Lemma 2.4.3. *For every function $\mathbf{f} \in \mathcal{G}(\beta, Q)$ and every $M > K(4 + C_{Q,\beta,d}) \vee (\beta + 1)^\beta \vee (Q+1)^{\beta/d} e^\beta$, there exists a neural network $\widetilde{\mathbf{q}} \in \mathcal{F}_{\mathbf{\Phi}}(L, \mathbf{m}, s)$, with*

(i) $L = 3\lceil \log_2(M)(d/\beta + 1)\rceil(1 + \lceil \log_2(d + \beta)\rceil) + \lfloor 40(\beta + 2)^2 \log_2(M)\rfloor + 2$,

(ii) $\mathbf{m} = \left(d, \lfloor 48K(d + \lceil\beta\rceil^3)2^\beta M^{d/\beta}\rfloor, \cdots, \lfloor 48K(d + \lceil\beta\rceil^3)2^\beta M^{d/\beta}\rfloor, K\right)$,

(iii) $s \leq 4707K(d + \beta + 1)^{4+d}2^\beta M^{d/\beta} \log_2(M))(d/\beta + 1)$,

such that,

$$\|\widetilde{\mathbf{q}}_k - \mathbf{p}_0\|_\infty \leq \frac{2K(4 + C_{Q,\beta,d})}{M},$$

and

$$\widetilde{q}_k(\mathbf{x}) \geq \frac{1}{M}, \quad \forall\, k \in \{1, \cdots, K\}, \forall\, \mathbf{x} \in [0,1]^d.$$

Proof. Composing the neural networks in Theorem 2.4.1 and Theorem 2.4.2 results in a neural network $\mathbf{G} = (G(H_1), \cdots, G(H_K))$ such that for any $k = 1, \cdots, K$,

$$\left\|e^{G(H_k)} - p_k^0\right\|_\infty \leq \left\|e^{G(H_k)} - H_k\right\|_\infty + \left\|H_k - p_k^0\right\|_\infty \leq \frac{4 + C_{Q,\beta,d}}{M}.$$

Define now the vector valued function $\widetilde{\mathbf{q}}$ component-wise by

$$\widetilde{q}_k(\mathbf{x}) = \frac{e^{G(H_k(\mathbf{x}))}}{\sum_{j=1}^K e^{G(H_j(\mathbf{x}))}}, \quad k = 1, \cdots, K.$$

Applying the composition (2.A.2), depth synchronization (2.A.3) and parallelization rules (2.A.4) it follows that $\widetilde{\mathbf{q}} \in \mathcal{F}_{\mathbf{\Phi}}(L, \mathbf{m}, s)$. To bound $\|\widetilde{q}_k - p_k^0\|_\infty$, we use that $\mathbf{p}_0 = (p_1^0, \cdots, p_K^0)$ is a probability vector, $e^{G(H_j)} \geq 0$ for $j = 1, \cdots, k$ and triangle

inequality, to obtain

$$\left\|\widetilde{q}_k - p_k^0\right\|_\infty \leq \left\|e^{G(H_k)}\left(\frac{1}{\sum_{j=1}^K e^{G(H_j)}} - 1\right)\right\|_\infty + \left\|e^{G(H_k)} - p_k^0\right\|_\infty$$

$$= \left\|e^{G(H_k)}\left(\frac{\sum_{\ell=1}^K p_\ell^0}{\sum_{j=1}^K e^{G(H_j)}} - \frac{\sum_{\ell=1}^K e^{G(H_\ell)}}{\sum_{j=1}^K e^{G(H_j)}}\right)\right\|_\infty + \left\|e^{G(H_k)} - p_k^0\right\|_\infty$$

$$\leq \left(\sum_{\ell=1}^K \left\|p_\ell^0 - e^{G(H_\ell)}\right\|_\infty\right)\left\|\frac{e^{G(H_k(\cdot))}}{\sum_{j=1}^K e^{G(H_j)}}\right\|_\infty + \left\|e^{G(H_k)} - p_k^0\right\|_\infty$$

$$\leq \frac{(K+1)(4+C_{Q,\beta,d})}{M} \leq \frac{2K(4+C_{Q,\beta,d})}{M}.$$

For the second bound of the lemma, notice that from the first bound of the lemma and the second bound of Theorem 2.4.1 it follows that

$$\widetilde{q}_k(\mathbf{x}) \geq \frac{\frac{4}{M}}{\sum_{j=1}^K e^{G(H_j(\mathbf{x}))}} \geq \frac{\frac{4}{M}}{1 + K\frac{(4+C_{Q,\beta,d})}{M}} = \frac{4}{M + K(4+C_{Q,\beta,d})} \geq \frac{1}{M},$$

where for the second inequality we used that $p_j(\mathbf{x}) \leq 1$, so $e^{G(H_j(\mathbf{x}))} \leq p_j^0(\mathbf{x}) + (4 + C_{Q,\beta,d})/M$ and for the last inequality we used that $M \geq K(4+C_{Q,\beta,d})$. $\square$

The Kullback-Leibler divergence can be upper bounded by the χ^2-divergence, see for instance Lemma 2.7 in [146]. Thus,

$$\mathbb{E}_{\mathbf{X}}\left[(\mathbf{p}_0(\mathbf{X}))^\top \log\left(\frac{\mathbf{p}_0(\mathbf{X})}{\widetilde{\mathbf{q}}(\mathbf{X})}\right)\right] \leq \mathbb{E}_{\mathbf{X}}\left[\sum_{k=1}^K \frac{(p_k^0(\mathbf{X}) - \widetilde{q}_k(\mathbf{X}))^2}{\widetilde{q}_k(\mathbf{X})}\right].$$

To control the approximation error, we can combine this bound with the first bound of Lemma 2.4.3 to conclude that if $p_k^0(\mathbf{X}) > 2K(4+C_{Q,\beta,d})/M$, then

$$\frac{(p_k^0(\mathbf{X}) - \widetilde{q}_k(\mathbf{X}))^2}{\widetilde{q}_k(\mathbf{X})} \leq \frac{4K^2(4+C_{Q,\beta,d})^2}{M^2}\left(p_k^0(\mathbf{X}) - \frac{2K(4+C_{Q,\beta,d})}{M}\right)^{-1}.$$

On the other hand, combining the bound with the second inequality from the same lemma yields

$$\frac{(p_k^0(\mathbf{X}) - \widetilde{q}_k(\mathbf{X}))^2}{\widetilde{q}_k(\mathbf{X})}$$

$$\leq \sum_{k=1}^K \frac{4K^2(4+C_{Q,\beta,d})^2}{M^2}\left(\max\left\{p_k^0(\mathbf{X}) - \frac{2K(4+C_{Q,\beta,d})}{M}, \frac{1}{M}\right\}\right)^{-1},$$

which is valid for all possible values of $\mathbf{p}_0(\mathbf{x}) \in [0,1]^k$. As M tends to infinity, $p_k^0(\mathbf{x}) - 2K(4 + C_{Q,\beta,d})/M$ tends to $p_k^0(\mathbf{x})$, while $1/M$ tends to zero. Without any further conditions on $p_k^0(\mathbf{X})$ this bound is thus of order M^{-1}. The small value bound, however, allows us to obtain an upper bound with better behaviour in M. The following proposition employs the small value bound to control the expectation of $(p_k^0(\mathbf{x}))^{-1}$ on the set that $p_k^0(\mathbf{x})$ exceeds some threshold value H.

Proposition 2.4.4. *Assume there exists an $\alpha \geq 0$ and a finite constant $C < \infty$, such that for $\mathbf{p} = (p_1, \ldots, p_K) : \mathcal{D} \to \mathcal{S}^K$ we have $\mathbb{P}_{\mathbf{X}}(p_k(\mathbf{X}) \leq t) \leq Ct^\alpha$ for all $t \geq 0$ and $k \in \{1, \ldots, K\}$. Let $H \in [0,1]$. Then it holds that*

$$\int_{\{p_k(\mathbf{x}) \geq H\}} \frac{1}{p_k(\mathbf{x})} \, d\mathbb{P}_{\mathbf{X}}(\mathbf{x}) \leq \begin{cases} C \frac{H^{\alpha-1}}{1-\alpha}, & \text{if } \alpha \in [0,1), \\ C(1 - \log(H)), & \text{if } \alpha \geq 1. \end{cases}$$

Proof. Observe that $p_k(\mathbf{X})$ is a probability. Therefore, $p_k(\mathbf{X}) \leq 1$ and consequently $C \geq 1$. For any nonnegative function h and random variable $Z \sim \mathbb{P}_Z$, we have $\int h(Z) \, d\mathbb{P}_Z = \mathbb{E}[h(Z)] = \int_0^\infty \mathbb{P}_Z(h(Z) \geq u) \, du$. Hence

$$\int_{\{p_k(\mathbf{x}) \geq H\}} \frac{1}{p_k(\mathbf{x})} \, d\mathbb{P}_{\mathbf{X}}(\mathbf{x}) = \int_0^\infty \mathbb{P}_{\mathbf{X}}\left(\frac{1}{p_k(\mathbf{X})} \mathbb{1}_{\{p_k(\mathbf{X}) \geq H\}} \geq u \right) du$$

$$\leq \int_0^{\frac{1}{H}} \mathbb{P}_{\mathbf{X}}\left(p_k(\mathbf{X}) \leq \frac{1}{u} \right) du,$$

where the inequality follows from observing that $\frac{1}{p_k(\mathbf{X})} \mathbb{1}_{\{p_k(\mathbf{X}) \geq H\}} \geq u$ implies $H \leq p_k(\mathbf{X}) \leq \frac{1}{u}$ and $u \leq 1/H$.

If $\alpha = 0$, we use the trivial bound $\mathbb{P}_{\mathbf{X}}(p_k(\mathbf{x}) \leq t) \leq 1$, for all $t \in [0,1]$, and obtain

$$\int_0^{\frac{1}{H}} \mathbb{P}_{\mathbf{X}}\left(p_k(\mathbf{X}) \leq \frac{1}{u} \right) du \leq \int_0^{\frac{1}{H}} 1 \, du = \frac{1}{H}.$$

If $0 < \alpha < 1$, we can invoke the assumption of this proposition to obtain

$$\int_0^{\frac{1}{H}} \mathbb{P}_{\mathbf{X}}\left(p_k(\mathbf{X}) \leq \frac{1}{u} \right) du \leq C \int_0^{\frac{1}{H}} u^{-\alpha} \, du = \frac{CH^{\alpha-1}}{1-\alpha}.$$

For $\alpha \geq 1$, we have $\mathbb{P}_{\mathbf{X}}(p_k(\mathbf{X}) \leq t) \leq Ct$ for all $0 \leq t \leq 1$. If moreover $C \leq H^{-1}$, the inequality $\mathbb{P}_{\mathbf{X}}(p_k(\mathbf{X}) \leq t) \leq \min\{1, Ct\}$ leads to

$$\int_0^{\frac{1}{H}} \mathbb{P}_{\mathbf{X}}\left(p_k(\mathbf{X}) \leq \frac{1}{u} \right) du \leq \int_0^C 1 \, du + C \int_C^{\frac{1}{H}} \frac{1}{u} \, du$$

$$= C + C(-\log(H) - \log(C)).$$

If $\alpha \geq 1$ and $C \geq H^{-1}$, we can upper bound the integral by $\int_0^C 1 \, du = C$. The result of the proposition now follows from simplifying the expressions using that $C \geq 1$. $\qquad \square$

We can now state and prove the main approximation bound.

Proof of Theorem 2.3.2. The condition $\|\mathbf{p} - \mathbf{p}_0\|_\infty \leq C_1/M$ implies that $p_k(\mathbf{x}) \geq p_k^0(\mathbf{x}) - C_1/M$. Combined with $p_k(\mathbf{x}) \geq 1/M$, this gives

$$p_k(\mathbf{x}) \geq \left(p_k^0(\mathbf{x}) - \frac{C_1}{M} \right) \vee \frac{1}{M} \geq \frac{p_k^0(\mathbf{x})}{C_1 + 1} \vee \frac{1}{M},$$

where we used that $p_k^0(\mathbf{x}) \geq (C_1 + 1)/M = ((C_1 + 1)/C_1) \cdot (C_1/M)$ implies

$$p_k^0(\mathbf{x}) - \frac{C_1}{M} \geq p_k^0(\mathbf{x}) \left(1 - \frac{C_1}{C_1 + 1} \right) = \frac{p_k^0(\mathbf{x})}{C_1 + 1}.$$

This gives rise to the upper bound

$$\frac{(p_k^0(\mathbf{X}) - p_k(\mathbf{X}))^2}{p_k(\mathbf{X})} \leq \frac{C_1^2}{M} \mathbb{1}_{\{p_k^0(\mathbf{x}) \leq \frac{C_1+1}{M}\}} + \frac{C_1^2}{M^2} \cdot \frac{C_1 + 1}{p_k^0(\mathbf{x})} \mathbb{1}_{\{p_k^0(\mathbf{x}) \geq \frac{C_1+1}{M}\}}.$$

Taking the expectation over the right hand side yields

$$\frac{C_1^2}{M} \mathbb{P}_{\mathbf{X}} \left(p_k^0(\mathbf{x}) \leq \frac{C_1 + 1}{M} \right) + \frac{C_1^2(C_1 + 1)}{M^2} \int_{\{p_k^0(\mathbf{x}) \geq \frac{C_1+1}{M}\}} \frac{1}{p_k^0(\mathbf{x})} \, d\mathbb{P}_{\mathbf{X}}(\mathbf{x})$$

By the α-SVB condition the first term is upper bounded by

$$\frac{C_1^2}{M} \mathbb{P}_{\mathbf{X}} \left(p_k^0(\mathbf{x}) \leq \frac{C_1 + 1}{M} \right) \leq \frac{C_1^2 C}{M} \left(\frac{C_1 + 1}{M} \right)^{\alpha \wedge 1} \leq C \frac{(C_1 + 1)^{2 + (\alpha \wedge 1)}}{M^{1 + (\alpha \wedge 1)}}.$$

Applying Proposition 2.4.4 with $H = (C_1 + 1)/M$ to the second term yields the result. $\qquad \square$

Now we have all the ingredients to complete the proof of the main theorem.

Proof of Theorem 2.3.3. Take $\delta = n^{-1}$ and $\epsilon = C_n = 1$ in Theorem 2.3.5. Using that d_τ is upper bounded by the sup-norm distance together with Lemma 2.3.8 gives

$$R_B(\mathbf{p}_0, \widehat{\mathbf{p}}) \leq 2 \left(\inf_{\mathbf{p} \in \mathcal{F}} R(\mathbf{p}_0, \mathbf{p}) + \Delta_n(\mathbf{p}_0, \widehat{\mathbf{p}}) + \frac{3}{n} \right)$$

$$+ 4 \cdot \frac{68B(s+1) \log(2^{2L+6} n(L+1) K^3 d^2 s^L) + 272B + (3/2)K(\log(n) + B)}{n}. \quad (2.4.1)$$

Recall that $0 \leq \alpha \leq 1$ is the index from the SVB condition. We now choose $M = \lfloor cK^{\frac{(2+\alpha)\beta}{(1+\alpha)\beta+d}} n^{\frac{\beta}{(1+\alpha)\beta+d}} \rfloor$ for a small constant c chosen below. To apply Lemma 2.4.3, we need to show that $M \gg K$. To see this, observe that $R_B(\mathbf{p}_0, \widehat{\mathbf{p}}) \leq B$ and therefore the convergence rate becomes trivial if $\phi_n \geq 1$. Using that $\phi_n = K^{\frac{(1+\alpha)\beta+(3+\alpha)d}{(1+\alpha)\beta+d}} n^{-\frac{(1+\alpha)\beta}{(1+\alpha)\beta+d}}$, this implies $K \leq n^{\frac{(1+\alpha)\beta}{(1+\alpha)\beta+(3+\alpha)d}} \leq n^{\frac{\beta}{\beta+2d}} \leq n^{\frac{\beta}{2d}}$. Hence, $K^{d-\beta} \ll n^{\beta}$ and thus also $M \gg K$.

For this choice of M, the network $\widetilde{\mathbf{q}}$ from Lemma 2.4.3 is in the network class $\mathcal{F}_{\mathbf{\Phi}}(L, \mathbf{m}, s)$, where $L = 3\lceil \log_2(M)(d/\beta + 1) \rceil (1 + \lceil \log_2(d+\beta) \rceil) + \lfloor 40(\beta+2)^2 \log_2(M) \rfloor + 2$, the maximum width of the hidden layers is bounded by $\lesssim Kc^{d/\beta} M^{d/\beta} = c^{d/\beta} n\phi_n$ and similarly $s \lesssim Kc^{d/\beta} M^{d/\beta} \log_2(M) = c^{d/\beta} n\phi_n \log_2(M)$. In particular, by taking c sufficiently small and using the depth synchronization property (2.A.3), $\widetilde{\mathbf{q}} \in \mathcal{F}_{\mathbf{\Phi}}(L, \mathbf{m}, s)$, whenever $A(d, \beta) \log_2(n) \leq L \lesssim n\phi_n$, for a suitable constant $A(d, \beta)$, the maximum width is $\gtrsim n\phi_n$ and $s \asymp n\phi_n \log(n)$. We now apply Theorem 2.3.2 with $C_1 = 2K(4 + C_{Q,\beta,d})$. Using that $C_1 + 1 = 2K(4 + C_{Q,\beta,d}) + 1 \leq 2K(5 + C_{Q,\beta,d})$, we find

$$\inf_{\mathbf{p} \in \mathcal{F}} R(\mathbf{p}_0, \mathbf{p}) \leq 8CK^{3+\alpha} \frac{(5 + C_{Q,\beta,d})^3}{M^{1+\alpha}} \left(1 + \frac{\mathbb{1}_{\{\alpha<1\}}}{1-\alpha} + \log(M) \right) \lesssim \phi_n \log(n).$$

Together with (2.4.1) and $s \asymp n\phi_n \log(n)$, the statement of Theorem 2.3.3 follows. $\square$

2.4.2 Oracle inequality related results

In this section we prove Theorem 2.3.5. For $B > 0$, consider

$$R_{B,n}(\mathbf{p}_0, \widehat{\mathbf{p}}) := \mathbb{E}_{\mathcal{D}_n} \left[\frac{1}{n} \sum_{i=1}^{n} \mathbf{Y}_i^\top \left(B \wedge \log \left(\frac{\mathbf{p}_0(\mathbf{X}_i)}{\widehat{\mathbf{p}}(\mathbf{X}_i)} \right) \right) \right].$$

The next proposition shows how this risk is related to the approximation error and the quantity $\Delta_n(\mathbf{p}_0, \widehat{\mathbf{p}})$ defined in (2.2.2) that measures the empirical distance between an arbitrary estimator and an empirical risk minimizer.

Proposition 2.4.5. *For any estimator $\widehat{\mathbf{p}} \in \mathcal{F}$,*

$$R_{B,n}(\mathbf{p}_0, \widehat{\mathbf{p}}) \leq R_{\infty,n}(\mathbf{p}_0, \widehat{\mathbf{p}}) \leq \inf_{\mathbf{p} \in \mathcal{F}} R(\mathbf{p}_0, \mathbf{p}) + \Delta_n(\mathbf{p}_0, \widehat{\mathbf{p}}).$$

Proof. The first inequality follows from $a \geq \min(a, b)$, for all $a, b \in \mathbb{R}$. To prove the second inequality, fix a $\mathbf{p}^* \in \mathcal{F}$. Using that $\Delta_n(\mathbf{p}_0, \mathbf{p}^*) \geq 0$ and

$$\mathbb{E}_{\mathcal{D}_n} \left[\mathbf{Y}_i^\top \log(\mathbf{p}^*(\mathbf{X}_i)) \right] = \mathbb{E}_{\mathcal{D}_n} \left[\mathbb{E}_{\mathcal{D}_n} [\mathbf{Y}_i^\top | \mathbf{X}_i] \log(\mathbf{p}^*(\mathbf{X}_i)) \right]$$
$$= \mathbb{E}_{\mathcal{D}_n} \left[\mathbf{p}_0(\mathbf{X}_i)^\top \log(\mathbf{p}^*(\mathbf{X}_i)) \right],$$

we get

$$\mathbb{E}_{\mathcal{D}_n}\Big[-\frac{1}{n}\sum_{i=1}^{n}\mathbf{Y}_i^{\top}\log(\widehat{\mathbf{p}}(\mathbf{X}_i))\Big] \leq \mathbb{E}_{\mathcal{D}_n}\Big[-\frac{1}{n}\sum_{i=1}^{n}\mathbf{Y}_i^{\top}\log(\widehat{\mathbf{p}}(\mathbf{X}_i))\Big] + \Delta_n(\mathbf{p}_0,\mathbf{p}^*)$$

$$=\mathbb{E}_{\mathcal{D}_n}\Big[-\frac{1}{n}\sum_{i=1}^{n}\mathbf{Y}_i^{\top}\log(\mathbf{p}^*(\mathbf{X}_i))\Big] + \Delta_n(\mathbf{p}_0,\widehat{\mathbf{p}})$$

$$=\mathbb{E}_{\mathbf{X}}\Big[-\mathbf{p}_0^{\top}(\mathbf{X})\log(\mathbf{p}^*(\mathbf{X}))\Big] + \Delta_n(\mathbf{p}_0,\widehat{\mathbf{p}}).$$

As this holds for all $\mathbf{p}^* \in \mathcal{F}$, we can take on the right hand side also the infimum over all $\mathbf{p}^* \in \mathcal{F}$. To complete the proof for the second inequality, we add to both sides $\mathbb{E}_{\mathcal{D}_n}[\mathbf{Y}_i^{\top}\log(\mathbf{p}_0(\mathbf{X}_i))] = \mathbb{E}_{\mathcal{D}_n}[\mathbf{p}_0(\mathbf{X}_i)^{\top}\log(\mathbf{p}_0(\mathbf{X}_i))]$. □

The truncation level B allows us to split the statistical risk into multiple parts that can be controlled separately. The following lemma provides a bound on the event that $p_k^0(\mathbf{X})$ is small.

Lemma 2.4.6. *Let $\mathcal{F}$ be a class of conditional class probabilities, $\widehat{\mathbf{p}}$ be any estimator taking values in $\mathcal{F}$, $(\overline{\mathbf{X}},\overline{\mathbf{Y}})$ be a random pair with the same distribution as $(\mathbf{X}_1,\mathbf{Y}_1)$ and $C_n \in (0,n/e)$. Then, for any $i \in \{1,\cdots,n\}$, and any $k \in \{1,\cdots,K\}$, we have*

$$\left|\mathbb{E}_{\mathcal{D}_n,(\overline{\mathbf{X}},\overline{\mathbf{Y}})}\left[\overline{Y}_k\mathbb{1}_{\{p_k^0(\overline{\mathbf{X}})\leq\frac{C_n}{n}\}}\Big(B\wedge\log\Big(\frac{p_k^0(\overline{\mathbf{X}})}{\widehat{p}_k(\overline{\mathbf{X}})}\Big)\Big)\right]\right| \leq \frac{C_n\big(\log\big(\frac{n}{C_n}\big)+B\big)}{n}.$$

Proof. Since $\mathbf{p}_0,\widehat{\mathbf{p}} \in [0,1]^K$, we have

$$\log(p_k^0(\overline{\mathbf{X}})) \leq B \wedge \log\Big(\frac{p_k^0(\overline{\mathbf{X}})}{\widehat{p}_k(\overline{\mathbf{X}})}\Big) \leq B. \tag{2.4.2}$$

Using that $a \leq x \leq b$ implies $|x| \leq \max\{|a|,|b|\} \leq |a|+|b|$ and $Y_k \geq 0$, we can get an upper bound that does not depend on $\widehat{\mathbf{p}}$

$$\left|\mathbb{E}_{\mathcal{D}_n,(\overline{\mathbf{X}},\overline{\mathbf{Y}})}\left[\overline{Y}_k\mathbb{1}_{\{p_k^0(\overline{\mathbf{X}})\leq\frac{C_n}{n}\}}\Big(B\wedge\log\Big(\frac{p_k^0(\overline{\mathbf{X}})}{\widehat{p}_k(\overline{\mathbf{X}})}\Big)\Big)\right]\right|$$

$$\leq \mathbb{E}_{(\overline{\mathbf{X}},\overline{\mathbf{Y}})}\left[\overline{Y}_k\mathbb{1}_{\{p_k^0(\overline{\mathbf{X}})\leq\frac{C_n}{n}\}}\big|\log(p_k^0(\overline{\mathbf{X}}))\big|\right] + \mathbb{E}_{(\overline{\mathbf{X}},\overline{\mathbf{Y}})}\left[\overline{Y}_k\mathbb{1}_{\{p_k^0(\overline{\mathbf{X}})\leq\frac{C_n}{n}\}}B\right]$$

$$= \mathbb{E}_{\overline{\mathbf{X}}}\left[p_k^0(\overline{\mathbf{X}})\mathbb{1}_{\{p_k^0(\overline{\mathbf{X}})\leq\frac{C_n}{n}\}}\big|\log(p_k^0(\overline{\mathbf{X}}))\big|\right] + \mathbb{E}_{\overline{\mathbf{X}}}\left[p_k^0(\overline{\mathbf{X}})\mathbb{1}_{\{p_k^0(\overline{\mathbf{X}})\leq\frac{C_n}{n}\}}B\right],$$

where the last equality follows from conditioning on $\overline{\mathbf{X}}$. Using that the function $u \mapsto u|\log(u)|$ is monotone increasing on $(0,e^{-1})$ and $n \geq eC_n$, yields

$$\left|\mathbb{E}_{\mathcal{D}_n,(\overline{\mathbf{X}},\overline{\mathbf{Y}})}\left[\overline{Y}_k\mathbb{1}_{\{p_k^0(\overline{\mathbf{X}})\leq\frac{C_n}{n}\}}\Big(B\wedge\log\Big(\frac{p_k^0(\overline{\mathbf{X}})}{\widehat{p}_k(\overline{\mathbf{X}})}\Big)\Big)\right]\right| \leq \frac{C_n\big(\log\big(\frac{n}{C_n}\big)+B\big)}{n}.$$

$\square$

Corollary 2.4.7. *Under the conditions of Lemma 2.4.6 it holds that*

$$-\frac{C_n \log(n/C_n)}{n} \leq \mathbb{E}_{\mathcal{D}_n,(\overline{\mathbf{X}},\overline{\mathbf{Y}})} \left[\overline{Y}_k \mathbb{1}_{\{p_k^0(\overline{\mathbf{X}}) \leq \frac{C_n}{n}\}} \left(B \wedge \log \left(\frac{p_k^0(\overline{\mathbf{X}})}{\widehat{p}_k(\overline{\mathbf{X}})} \right) \right) \right] \leq \frac{C_n B}{n}.$$

Proof. The lower and upper bound can be obtained from (2.4.2), $\overline{Y}_k \geq 0$ and the fact that $u \mapsto u \log(u)$ is monotone decreasing on $(0, e^{-1})$. $\square$

Both Lemma 2.4.6 and Corollary 2.4.7 do not require that the random pair $(\overline{\mathbf{X}}, \overline{\mathbf{Y}})$ is independent of the data. Specifically, they also hold in the case that $(\overline{\mathbf{X}}, \overline{\mathbf{Y}}) = (\mathbf{X}_i, \mathbf{Y}_i)$ for some $i \in \{1, \cdots, n\}$.

Proof of Theorem 2.3.5. For ease of notation set

$$\left(B \wedge \log \left(\frac{\mathbf{p}_0(\mathbf{X}_i)}{\widehat{\mathbf{p}}(\mathbf{X}_i)} \right) \right)_{\geq C_n/n}$$

to denote the vector with coefficients

$$\mathbb{1}_{\{p_k^0(\mathbf{X}_i) \geq \frac{C_n}{n}\}} \left(B \wedge \log \left(\frac{p_k^0(\mathbf{X}_i)}{\widehat{p}_k(\mathbf{X}_i)} \right) \right), \quad k = 1, \ldots, K.$$

For i.i.d. random pairs $(\widetilde{\mathbf{X}}_i, \widetilde{\mathbf{Y}}_i)$, $i = 1, \cdots, n$ with joint distribution $\mathbb{P}$ that are generated independently of the data sample define $\mathcal{D}'_n := \{(\mathbf{X}_i, \mathbf{Y}_i)_i, (\widetilde{\mathbf{X}}_i, \widetilde{\mathbf{Y}}_i)_i\}$. Then, for any $C_n > 0$,

$$\begin{aligned}
&|R_B(\mathbf{p}_0, \widehat{\mathbf{p}}) - R_{B,n}(\mathbf{p}_0, \widehat{\mathbf{p}})| \\
&= \left| \mathbb{E}_{\mathcal{D}'_n} \left[\frac{1}{n} \sum_{i=1}^{n} \sum_{k=1}^{K} \widetilde{Y}_{i,k} \left(B \wedge \log \left(\frac{p_k^0(\widetilde{\mathbf{X}}_i)}{\widehat{p}_k(\widetilde{\mathbf{X}}_i)} \right) \right) \right. \right. \\
&\qquad \left. \left. - \frac{1}{n} \sum_{i=1}^{n} \sum_{k=1}^{K} Y_{i,k} \left(B \wedge \log \left(\frac{p_k^0(\mathbf{X}_i)}{\widehat{p}_k(\mathbf{X}_i)} \right) \right) \right] \right| \\
&\leq (I) + (II) + (III),
\end{aligned} \tag{2.4.3}$$

where

$$(I) = \left| \mathbb{E}_{\mathcal{D}'_n} \left[\frac{1}{n} \sum_{i=1}^{n} \left(\widetilde{\mathbf{Y}}_i^{\top} \left(B \wedge \log \left(\frac{\mathbf{p}_0(\widetilde{\mathbf{X}}_i)}{\widehat{\mathbf{p}}(\widetilde{\mathbf{X}}_i)} \right) \right)_{\geq C_n/n} \right. \right. \right.$$

$$- \mathbf{Y}_i^\top \Big(B \wedge \log \Big(\frac{\mathbf{p}_0(\mathbf{X}_i)}{\widehat{\mathbf{p}}(\mathbf{X}_i)} \Big) \Big)_{\geq C_n/n} \Big) \Big] \Big|$$

$$(II) = \Big| \mathbb{E}_{\mathcal{D}_n'} \Big[\frac{1}{n} \sum_{i=1}^n \sum_{k=1}^K \widetilde{Y}_{i,k} \mathbb{1}_{\{p_k^0(\widetilde{\mathbf{X}}_i) \leq \frac{C_n}{n}\}} \Big(B \wedge \log \Big(\frac{p_k^0(\widetilde{\mathbf{X}}_i)}{\widehat{p}_k(\widetilde{\mathbf{X}}_i)} \Big) \Big) \Big] \Big|$$

$$(III) = \Big| \mathbb{E}_{\mathcal{D}_n'} \Big[\frac{1}{n} \sum_{i=1}^n \sum_{k=1}^K Y_{i,k} \mathbb{1}_{\{p_k^0(\mathbf{X}_i) \leq \frac{C_n}{n}\}} \Big(B \wedge \log \Big(\frac{p_k^0(\mathbf{X}_i)}{\widehat{p}_k(\mathbf{X}_i)} \Big) \Big) \Big] \Big|.$$

First we bound the terms (II) and (III). Applying Lemma 2.4.6 in total nK times with $(\overline{\mathbf{X}}, \overline{\mathbf{Y}}) = (\widetilde{\mathbf{X}}_i, \widetilde{\mathbf{Y}}_i)$, yields

$$(II) \leq \frac{1}{n} \sum_{i=1}^n \sum_{k=1}^K \frac{C_n \big(\log \big(\frac{n}{C_n} \big) + B \big)}{n} = \frac{C_n K \big(\log \big(\frac{n}{C_n} \big) + B \big)}{n}, \qquad (2.4.4)$$

while taking $(\overline{\mathbf{X}}, \overline{\mathbf{Y}}) = (\mathbf{X}_i, \mathbf{Y}_i)$ in Lemma 2.4.6 yields

$$(III) \leq \frac{1}{n} \sum_{i=1}^n \sum_{k=1}^K \frac{C_n \big(\log \big(\frac{n}{C_n} \big) + B \big) + B \big)}{n} = \frac{C_n K \big(\log \big(\frac{n}{C_n} \big) + B \big)}{n}. \qquad (2.4.5)$$

Now we deal with the term (I). Due to the bound B and the indicator function

$$\mathbb{1}_{\{p_k^0(\mathbf{X}_i) \geq \frac{C_n}{n}\}} \Big(B \wedge \log \Big(\frac{p_k^0(\mathbf{X}_i)}{\widehat{p}_k(\mathbf{X}_i)} \Big) \Big)$$

$$= \mathbb{1}_{\{p_k^0(\mathbf{X}_i) \geq \frac{C_n}{n}\}} \Big(B \wedge \log \Big(\frac{p_k^0(\mathbf{X}_i)}{(C_n e^{-B}/n) \vee \widehat{p}_k(\mathbf{X}_i)} \Big) \Big). \quad (2.4.6)$$

Given a minimal (internal) δ-covering of $\log(\mathcal{F})$ with respect to the pseudometric d_τ, with $\tau = \log(C_n e^{-B}/n)$, denote the centers of the balls by $\mathbf{p}_\ell$. Then there exists a random ℓ^* such that

$$\Big\| \log \Big(\frac{C_n e^{-B}}{n} \Big) \vee \log(\widehat{\mathbf{p}}) - \log \Big(\frac{C_n e^{-B}}{n} \Big) \vee \log(\mathbf{p}_{\ell^*}) \Big\|_\infty \leq \delta.$$

This together with (2.4.6) and using that $\mathbf{Y}$ is one of the K-dimensional standard basis vectors yields

$$(I) \leq \mathbb{E}_{\mathcal{D}_n'} \Big[\Big| \frac{1}{n} \sum_{i=1}^n G_{\ell^*}(\widetilde{\mathbf{X}}_i, \widetilde{\mathbf{Y}}_i, \mathbf{X}_i, \mathbf{Y}_i) \Big| \Big] + 2\delta, \qquad (2.4.7)$$

where

$$G_{\ell^*}(\widetilde{\mathbf{X}}_i, \widetilde{\mathbf{Y}}_i, \mathbf{X}_i, \mathbf{Y}_i) :=$$

$$\widetilde{\mathbf{Y}}_i^\top \left(B \wedge \log \left(\frac{\mathbf{p}_0(\widetilde{\mathbf{X}}_i)}{\mathbf{p}_{\ell^*}(\widetilde{\mathbf{X}}_i)} \right) \right)_{\geq C_n/n} - \mathbf{Y}_i^\top \left(B \wedge \log \left(\frac{\mathbf{p}_0(\mathbf{X}_i)}{\mathbf{p}_{\ell^*}(\mathbf{X}_i)} \right) \right)_{\geq C_n/n}. \quad (2.4.8)$$

For all $\ell \in \{1, \cdots, \mathcal{N}_n\}$ define G_ℓ in the same way. Moreover, write

$$\mathbf{Z}_i := (\widetilde{\mathbf{X}}_i, \widetilde{\mathbf{Y}}_i, \mathbf{X}_i, \mathbf{Y}_i).$$

In a next step, we apply Bernstein's inequality (Proposition 2.C.1) to $(G_\ell(\mathbf{Z}_i))_{i=1}^n$. Using that $(\mathbf{X}_i, \mathbf{Y}_i)$ and $(\widetilde{\mathbf{X}}_i, \widetilde{\mathbf{Y}}_i)$ have the same distribution, we get for the expectation of G_ℓ that

$$\mathbb{E}_{\mathcal{D}_n'}[G_\ell(\mathbf{Z}_i)] = 0.$$

To verify the assumptions of Bernstein's inequality, it remains to prove that

$$\mathbb{E}|G_\ell(\mathbf{Z}_i)|^m \leq m!(2B)^{m-2} R_B(\mathbf{p}_0, \mathbf{p}_\ell) 32B 2^{-1}, \ \forall m \in \mathbb{N}_{\geq 2}, \quad (2.4.9)$$

such that, in the notation of Proposition 2.C.1, we have $v_i = R_B(\mathbf{p}_0, \mathbf{p}_\ell) 32B$ and $U = 2B$. To show this moment bound, observe that any real numbers a, b satisfy $|a + b|^m \leq 2^m(|a|^m + |b|^m)$. Using moreover that $(\mathbf{X}_i, \mathbf{Y}_i)$ and $(\widetilde{\mathbf{X}}_i, \widetilde{\mathbf{Y}}_i)$ have the same distribution, the m-th absolute moment of G_ℓ is given by

$$\mathbb{E}_{\mathcal{D}_n'}\left[|G_\ell(\mathbf{Z}_i)|^m\right]$$

$$= \mathbb{E}_{\mathcal{D}_n'}\left[\left| \widetilde{\mathbf{Y}}_i^\top \left(B \wedge \log \left(\frac{\mathbf{p}_0(\widetilde{\mathbf{X}}_i)}{\mathbf{p}_\ell(\widetilde{\mathbf{X}}_i)} \right) \right)_{\geq C_n/n} \right.\right.$$

$$\left.\left. - \mathbf{Y}_i^\top \left(B \wedge \log \left(\frac{\mathbf{p}_0(\mathbf{X}_i)}{\mathbf{p}_\ell(\mathbf{X}_i)} \right) \right)_{\geq C_n/n} \right|^m\right]$$

$$\leq 2^{m+1} \mathbb{E}_{\mathcal{D}_n}\left[\left| \mathbf{Y}_i^\top \left(B \wedge \log \left(\frac{\mathbf{p}_0(\mathbf{X}_i)}{\mathbf{p}_\ell(\mathbf{X}_i)} \right) \right)_{\geq C_n/n} \right|^m\right].$$

Triangle inequality gives

$$\mathbb{E}_{\mathcal{D}_n}\left[\left| \mathbf{Y}_i^\top \left(B \wedge \log \left(\frac{\mathbf{p}_0(\mathbf{X}_i)}{\mathbf{p}_\ell(\mathbf{X}_i)} \right) \right)_{\geq C_n/n} \right|^m\right]$$

$$\leq \mathbb{E}_{\mathcal{D}_n}\left[\left(\mathbf{Y}_i^\top \left| \left(B \wedge \log \left(\frac{\mathbf{p}_0(\mathbf{X}_i)}{\mathbf{p}_\ell(\mathbf{X}_i)} \right) \right)_{\geq C_n/n} \right| \right)^m\right],$$

where for a vector $\mathbf{v}$, $|\mathbf{v}|$ denotes the absolute value coefficient-wise. Since $\mathbf{Y}$ is one of the standard basis vectors, it holds that $Y_k \in \{0, 1\}$, and $Y_k Y_j$ is equal to 0 when $j \neq k$ and equal to Y_k when $k = j$. Using this observation together with conditioning on $\mathbf{X}_i$ yields

$$\mathbb{E}_{\mathcal{D}_n}\left[\left(\mathbf{Y}_i^\top \left|\left(B \wedge \log\left(\frac{\mathbf{p}_0(\mathbf{X}_i)}{\mathbf{p}_\ell(\mathbf{X}_i)}\right)\right)_{\geq C_n/n}\right|\right)^m\right]$$

$$= \mathbb{E}_{\mathcal{D}_n}\left[\mathbf{Y}_i^\top \left|\left(B \wedge \log\left(\frac{\mathbf{p}_0(\mathbf{X}_i)}{\mathbf{p}_\ell(\mathbf{X}_i)}\right)\right)_{\geq C_n/n}\right|^m\right]$$

$$= \mathbb{E}_{\mathbf{X}_i}\left[\mathbf{p}_0^\top(\mathbf{X}_i) \left|\left(B \wedge \log\left(\frac{\mathbf{p}_0(\mathbf{X}_i)}{\mathbf{p}_\ell(\mathbf{X}_i)}\right)\right)_{\geq C_n/n}\right|^m\right]$$

$$\leq \mathbb{E}_{\mathbf{X}_i}\left[\mathbf{p}_0^\top(\mathbf{X}_i) \left|B \wedge \log\left(\frac{\mathbf{p}_0(\mathbf{X}_i)}{\mathbf{p}_\ell(\mathbf{X}_i)}\right)\right|^m\right],$$

where we used for the last inequality that for every set Ω, each $A \subseteq \Omega$, every function $\theta : \Omega \to \mathbb{R}$ and every $m \in \mathbb{N}_{\geq 2}$ it holds that $|\mathbb{1}_A \theta|^m = (\mathbb{1}_A)^m |\theta|^m = \mathbb{1}_A |\theta|^m \leq |\theta|^m$. Combining the previous displays and applying Lemma 2.3.7, we get that

$$\mathbb{E}_{\mathcal{D}_n'}[|G_\ell(\mathbf{Z}_i)|^m]$$

$$\leq 2^{m+1} \mathbb{E}_{\mathbf{X}_i}\left[\mathbf{p}_0^\top(\mathbf{X}_i) \left|B \wedge \log\left(\frac{\mathbf{p}_0(\mathbf{X}_i)}{\mathbf{p}_\ell(\mathbf{X}_i)}\right)\right|^m\right]$$

$$\leq 2^{m+1} C_{m,B} \mathbb{E}_{\mathbf{X}_i}\left[\mathbf{p}_0^\top(\mathbf{X}_i) \left(B \wedge \log\left(\frac{\mathbf{p}_0(\mathbf{X}_i)}{\mathbf{p}_\ell(\mathbf{X}_i)}\right)\right)\right] = 2^{m+1} C_{m,B} R_B(\mathbf{p}_0, \mathbf{p}_\ell),$$
$$(2.4.10)$$

where $C_{m,B}$ is given by

$$C_{m,B} = \max\left\{m!, \frac{B^m}{B-1}\right\}.$$

Since $B \geq 2$, we get that $B/(B-1) \leq 2$ and $C_{m,B} \leq \max\left\{m!, 2B^{m-1}\right\} \leq 2m!B^{m-1}$. Together with (2.4.10) this yields

$$\mathbb{E}_{\mathcal{D}_n'}[|G_\ell(\mathbf{Z}_i)|^m] \leq 2^{m+1} C_{m,B} R_B(\mathbf{p}_0, \mathbf{p}_\ell) \leq m!(2B)^{m-2} R_B(\mathbf{p}_0, \mathbf{p}_\ell) 32B2^{-1},$$

completing the proof for the moment bound (2.4.9).

Now define $z_\ell := \sqrt{n^{-1}68B \log(\mathcal{N}_n)} \vee \sqrt{\mathbb{E}_{(\mathbf{X},\mathbf{Y})}[\mathbf{Y}^\top (B \wedge \log(\mathbf{p}_0(\mathbf{X})/\mathbf{p}_\ell(\mathbf{X})))]}$. Since $B \geq 2$, Lemma 2.3.4 guarantees that the truncated Kullback-Leibler risk is always nonnegative, so z_ℓ is well defined. Define $z^* = z_{\ell^*}$, that is,

$$z^* = \sqrt{\frac{68B \log(\mathcal{N}_n)}{n}} \vee \sqrt{\mathbb{E}_{\mathcal{D}_N,(\mathbf{X},\mathbf{Y})}\left[\mathbf{Y}^\top \left(B \wedge \log\left(\frac{\mathbf{p}_0(\mathbf{X})}{\mathbf{p}_{\ell^*}(\mathbf{X})}\right)\right)\middle|\mathcal{D}_n\right]},$$

where we also condition on the dataset $\mathcal{D}_n$. To upper bound z^*, we split the truncated empirical risk

$$
\mathbb{E}_{\mathcal{D}_N,(\mathbf{X},\mathbf{Y})}\left[\mathbf{Y}^\top\left(B\wedge\log\left(\frac{\mathbf{p}_0(\mathbf{X})}{\mathbf{p}_{\ell^*}(\mathbf{X})}\right)\right)\Bigg|\mathcal{D}_n\right]
$$

$$
=\mathbb{E}_{\mathcal{D}_N,(\mathbf{X},\mathbf{Y})}\left[\sum_{k=1}^K Y_k\mathbb{1}_{\{p_k^0(\mathbf{X})\leq\frac{C_n}{n}\}}\left(B\wedge\log\left(\frac{p_k^0(\mathbf{X})}{p_{\ell^*,k}(\mathbf{X})}\right)\right)\Bigg|\mathcal{D}_n\right]
$$

$$
+\mathbb{E}_{\mathcal{D}_N,(\mathbf{X},\mathbf{Y})}\left[\sum_{k=1}^K Y_k\mathbb{1}_{\{p_k^0(\mathbf{X})\geq\frac{C_n}{n}\}}\left(B\wedge\log\left(\frac{p_k^0(\mathbf{X})}{p_{\ell^*,k}(\mathbf{X})}\right)\right)\Bigg|\mathcal{D}_n\right].
$$

Using the property of the δ-cover, Equation (2.4.6) and the fact that $\mathbf{Y}$ is a standard basis vector, it holds that

$$
\mathbb{E}_{\mathcal{D}_N,(\mathbf{X},\mathbf{Y})}\left[\sum_{k=1}^K Y_k\mathbb{1}_{\{p_k^0(\mathbf{X})\geq\frac{C_n}{n}\}}\left(B\wedge\log\left(\frac{p_k^0(\mathbf{X})}{p_{\ell^*,k}(\mathbf{X})}\right)\right)\Bigg|\mathcal{D}_n\right]
$$

$$
\leq\mathbb{E}_{\mathcal{D}_N,(\mathbf{X},\mathbf{Y})}\left[\sum_{k=1}^K Y_k\mathbb{1}_{\{p_k^0(\mathbf{X})\geq\frac{C_n}{n}\}}\left(B\wedge\log\left(\frac{p_k^0(\mathbf{X})}{\widehat{p}_k(\mathbf{X})}\right)\right)\Bigg|\mathcal{D}_n\right]+\delta.
$$

On the other hand, applying Corollary 2.4.7, with $(\overline{\mathbf{X}},\overline{\mathbf{Y}})=(\mathbf{X},\mathbf{Y})$, K times for $\widehat{\mathbf{p}}$ and K times with $\widehat{\mathbf{p}}$ replaced by $\mathbf{p}_{\ell^*}$, yields

$$
\mathbb{E}_{\mathcal{D}_N,(\mathbf{X},\mathbf{Y})}\left[\sum_{k=1}^K Y_k\mathbb{1}_{\{p_k^0(\mathbf{X})\leq\frac{C_n}{n}\}}\left(B\wedge\log\left(\frac{p_k^0(\mathbf{X})}{p_{\ell^*,k}(\mathbf{X})}\right)\right)\Bigg|\mathcal{D}_n\right]
$$

$$
\leq\mathbb{E}_{\mathcal{D}_N,(\mathbf{X},\mathbf{Y})}\left[\sum_{k=1}^K Y_k\mathbb{1}_{\{p_k^0(\mathbf{X})\leq\frac{C_n}{n}\}}\left(B\wedge\log\left(\frac{p_k^0(\mathbf{X})}{\widehat{p}_k(\mathbf{X})}\right)\right)\Bigg|\mathcal{D}_n\right]
$$

$$
+\frac{C_nK\left(\log\left(\frac{n}{C_n}\right)+B\right)}{n}.
$$

Define

$$
V:=\sqrt{\mathbb{E}_{\mathcal{D}_N,(\mathbf{X},\mathbf{Y})}\left[\mathbf{Y}^\top\left(B\wedge\log\left(\frac{\mathbf{p}_0(\mathbf{X})}{\widehat{\mathbf{p}}(\mathbf{X})}\right)\right)\Bigg|\mathcal{D}_n\right]}.
$$

Combining the previous inequalities, we get that

$$
\sqrt{\mathbb{E}_{\mathcal{D}_N,(\mathbf{X},\mathbf{Y})}\left[\mathbf{Y}^\top\left(B\wedge\log\left(\frac{\mathbf{p}_0(\mathbf{X})}{\mathbf{p}_{\ell^*}(\mathbf{X})}\right)\right)\Bigg|\mathcal{D}_n\right]}\leq V+\sqrt{\delta+\frac{C_nK\left(\log\left(\frac{n}{C_n}\right)+B\right)}{n}},
$$

where we also used the elementary inequality $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$ for all $a, b \geq 0$. Hence,

$$z^* \leq \sqrt{\frac{68B \log(\mathcal{N}_n)}{n}} + V + \sqrt{\delta + \frac{C_n K \left(\log\left(\frac{n}{C_n}\right) + B\right)}{n}}. \tag{2.4.11}$$

The term $\sqrt{n^{-1} 68B \log(\mathcal{N}_n)}$ is chosen such that in (2.4.13) and (2.4.14) below the equations balance out. Now define

$$T := \max_{\ell} \left| \sum_{i=1}^{n} \frac{G_\ell(\mathbf{Z}_i)}{z_\ell} \right|.$$

The Cauchy-Schwarz inequality gives us that $\mathbb{E}_{\mathcal{D}'_n}[VT] \leq \sqrt{\mathbb{E}_{\mathcal{D}'_n}[V^2] \mathbb{E}_{\mathcal{D}'_n}[T^2]}$. Noticing that $\mathbb{E}_{\mathcal{D}'_n}[V^2] = R_B(\mathbf{p}_0, \widehat{\mathbf{p}})$, we get from (2.4.3), (2.4.4), (2.4.5), (2.4.7) and (2.4.11) that

$$
\begin{aligned}
&|R_B(\mathbf{p}_0, \widehat{\mathbf{p}}) - R_{B,n}(\mathbf{p}_0, \widehat{\mathbf{p}})| \\
&\leq \frac{1}{n} \sqrt{R_B(\mathbf{p}_0, \widehat{\mathbf{p}})} \sqrt{\mathbb{E}_{\mathcal{D}'_n}[T^2]} \\
&+ \frac{1}{n} \left(\sqrt{\frac{68B \log(\mathcal{N}_n)}{n}} + \sqrt{\delta + \frac{C_n K\left(\log\left(\frac{n}{C_n}\right) + B\right)}{n}} \right) \mathbb{E}_{\mathcal{D}'_n}[T] \\
&+ 2\delta + \frac{2C_n K\left(\log\left(\frac{n}{C_n}\right) + B\right)}{n}.
\end{aligned} \tag{2.4.12}
$$

The next step in the proof derives bounds on $\mathbb{E}_{\mathcal{D}'_n}[T]$ and $\mathbb{E}_{\mathcal{D}'_n}[T^2]$. Using an union bound it holds that

$$
\begin{aligned}
\mathbb{P}\left(T \geq t\right) &= \mathbb{P}\left(\max_{\ell} \left| \sum_{i=1}^{n} \frac{G_\ell(\mathbf{Z}_i)}{z_\ell} \right| \geq t \right) = \mathbb{P}\left(\bigcup_{\ell=1}^{\mathcal{N}_n} \left(\left| \sum_{i=1}^{n} \frac{G_\ell(\mathbf{Z}_i)}{z_\ell} \right| \geq t \right) \right) \\
&\leq \sum_{\ell=1}^{\mathcal{N}_n} \mathbb{P}\left(\left| \sum_{i=1}^{n} G_\ell(\mathbf{Z}_i) \right| \geq t z_\ell \right).
\end{aligned}
$$

We already showed that $G_\ell(\mathbf{Z}_i)$ satisfies the conditions of Bernstein's inequality (Proposition 2.C.1) with $v_i = R_B(\mathbf{p}_0, \mathbf{p}_\ell) 32B$ and $U = 2B$. Bernstein's inequality

applied to the last term gives

$$\mathbb{P}\left(T \geq t\right) \leq \sum_{\ell=1}^{\mathcal{N}_n} \mathbb{P}\left(\left|\sum_{i=1}^{n} G_\ell(\mathbf{Z}_i)\right| \geq t z_\ell\right)$$

$$\leq \sum_{\ell=1}^{\mathcal{N}_n} 2\exp\left(-\frac{(tz_\ell)^2}{2n R_B(\mathbf{p}_0,\mathbf{p}_\ell)32B + 4Bt z_\ell}\right)$$

$$= 2\mathcal{N}_n \exp\left(-\frac{t^2}{2n\frac{R_B(\mathbf{p}_0,\mathbf{p}_\ell)32B}{z_\ell^2} + 4B\frac{t}{z_\ell}}\right).$$

Since $z_\ell \geq \sqrt{R_B(\mathbf{p}_0,\mathbf{p}_\ell)}$ it holds that $z_\ell^2 \geq R_B(\mathbf{p}_0,\mathbf{p}_\ell)$. As probabilities are in the interval $[0,1]$, this gives us that

$$\mathbb{P}\left(T \geq t\right) \leq 1 \wedge 2\mathcal{N}_n \exp\left(-\frac{t^2}{64Bn + 4B\frac{t}{z_\ell}}\right).$$

If $t \geq \sqrt{68Bn\log(\mathcal{N}_n)}$, then since $z_\ell \geq \sqrt{n^{-1}68B\log(\mathcal{N}_n)}$ it holds that

$$\exp\left(-\frac{t^2}{64Bn + 4B\frac{t}{z_\ell}}\right) \leq \exp\left(-\frac{t\sqrt{\log(\mathcal{N}_n)}}{\sqrt{68Bn}}\right).$$

For every nonnegative random variable X with finite expectation one has $\mathbb{E}[X] = \int_0^\infty \mathbb{P}(X \geq t)\, dt$. Therefore,

$$\mathbb{E}_{\mathcal{D}_n'}[T] \leq \sqrt{68Bn\log(\mathcal{N}_n)} + \int_{\sqrt{68Bn\log(\mathcal{N}_n)}}^\infty 2\mathcal{N}_n \exp\left(-\frac{t\sqrt{\log(\mathcal{N}_n)}}{\sqrt{68Bn}}\right) dt$$

$$= \sqrt{68Bn\log(\mathcal{N}_n)} + \sqrt{\frac{272Bn}{\log(\mathcal{N}_n)}}.$$

$$(2.4.13)$$

Since T is nonnegative, $\mathbb{P}(T^2 \geq u) = \mathbb{P}(T \geq \sqrt{u})$, so using the same arguments as before we get that

$$\mathbb{E}_{\mathcal{D}_n'}[T^2] \leq 68Bn\log(\mathcal{N}_n) + \int_{68Bn\log(\mathcal{N}_n)}^\infty 2\mathcal{N}_n \exp\left(-\sqrt{\frac{u\log(\mathcal{N}_n)}{68Bn}}\right) du.$$

Substitution $s = \sqrt{u}$ and integration by parts gives us that $(1/2)\int_a^\infty e^{-\sqrt{u}b}\, du = \int_{\sqrt{a}}^\infty se^{-sb}\, ds = (\sqrt{a}b+1)e^{-\sqrt{a}b}/b^2$ and consequently

$$\mathbb{E}_{\mathcal{D}_n'}[T^2] \leq 68Bn\log(\mathcal{N}_n) + 544Bn, \qquad (2.4.14)$$

where we also used that $\mathcal{N}_n \geq e$ and thus $(\log(\mathcal{N}_n) + 1)/\log(\mathcal{N}_n) \geq 2$.

Combining (2.4.13), (2.4.14) with (2.4.12), using twice that $2xy \leq x^2 + y^2$ for all real numbers x, y, and using that $\log(\mathcal{N}_n) \geq 1$, we get that

$$|R_B(\mathbf{p}_0, \widehat{\mathbf{p}}) - R_{B,n}(\mathbf{p}_0, \widehat{\mathbf{p}})| \leq \sqrt{R_B(\mathbf{p}_0, \widehat{\mathbf{p}})}\sqrt{\frac{68B\log(\mathcal{N}_n) + 544B}{n}} + 3\delta$$
$$+ \frac{102B\log(\mathcal{N}_n) + 272B}{n} + \frac{3C_nK\left(\log\left(\frac{n}{C_n}\right) + B\right)}{n}.$$
$$(2.4.15)$$

Setting $a = R_B(\mathbf{p}_0, \widehat{\mathbf{p}})$, $b = R_{B,n}(\mathbf{p}_0, \widehat{\mathbf{p}})$,

$$c = \sqrt{\frac{17B\log(\mathcal{N}_n) + 134B}{n}},$$

and

$$d = \frac{102B\log(\mathcal{N}_n) + 272B + 3C_nK\left(\log\left(\frac{n}{C_n}\right) + B\right)}{n} + 3\delta,$$

we get from (2.4.15) that $|a - b| \leq 2\sqrt{ac} + d$. Since the excess risk is always nonnegative we can apply Proposition 2.C.2. This gives us for any $0 < \epsilon \leq 1$

$$R_B(\mathbf{p}_0, \widehat{\mathbf{p}}) \leq (1 + \epsilon)\left(R_{B,n}(\mathbf{p}_0, \widehat{\mathbf{p}}) + 3\delta\right)$$
$$+ (1 + \epsilon)\left(\frac{102B\log(\mathcal{N}_n) + 272B + 3C_nK\left(\log\left(\frac{n}{C_n}\right) + B\right)}{n}\right)$$
$$+ \frac{(1 + \epsilon)^2}{\epsilon} \cdot \frac{17B\log(\mathcal{N}_n) + 136B}{n}.$$

Proposition 2.4.5 gives $R_{B,n}(\mathbf{p}_0, \widehat{\mathbf{p}}) \leq \inf_{\mathbf{p} \in \mathcal{F}} R(\mathbf{p}_0, \mathbf{p}) + \Delta_n(\mathbf{p}_0, \widehat{\mathbf{p}})$. Substituting this in the previous equation and observing that $(1 + \epsilon)/\epsilon \geq 2$, $1/\epsilon \geq 1$ and $0 < 1 - \epsilon \leq 1$ for $\epsilon \in (0, 1]$ yields the assertion of the theorem. $\qquad\square$

Appendix Chapter 2

2.A Basic network properties and operations

In this section we state elementary properties of network classes and introduce small networks that are capable of approximating multiplication operations based on similar results in [127].

2.A.1 Embedding properties of neural network function classes

This section extends the results in [127] to arbitrary output activation function.

Enlarging: Let $\mathbf{m}$ and $\mathbf{m}'$ be two width-vectors of the same length and let $s, s' > 0$. If $\mathbf{m} \leq \mathbf{m}'$ component-wise, $m_{L+1} = m'_{L+1}$ and $s \leq s'$, then

$$\mathcal{F}_{\boldsymbol{\psi}}(L, \mathbf{m}, s) \subseteq \mathcal{F}_{\boldsymbol{\psi}}(L, \mathbf{m}', s'). \tag{2.A.1}$$

This rule allows us to simplify the neural network architectures. For example we can simplify a network class by embedding it in a class for which all hidden layers have the same width.

Composition: Let $\mathbf{f} \in \mathcal{F}_{\mathrm{id}}(L, \mathbf{m}, s_1)$ and let $\mathbf{g}$ be a network in $\mathcal{F}_{\boldsymbol{\psi}}(L', \mathbf{m}', s_2)$, with $m_{L+1} = m'_0$. For a vector $\mathbf{v} \in \mathbb{R}^{m_{L+1}}$, define the composed network $\mathbf{g} \circ \sigma_{\mathbf{v}}(\mathbf{f})$. Then

$$\mathbf{g} \circ \sigma_{\mathbf{v}}(\mathbf{f}) \in \mathcal{F}_{\boldsymbol{\psi}}\big(L + L' + 1, (m_0, \cdots, m_{L+1}, m'_1, \cdots, m'_{L'+1}), s_1 + s_2 + |\mathbf{v}|_0\big). \tag{2.A.2}$$

The following rule allows us to synchronize the depths of neural networks.

Depth synchronization: For any positive integer a,

$$\mathcal{F}_{\boldsymbol{\psi}}(L, \mathbf{m}, s) \subset \mathcal{F}_{\boldsymbol{\psi}}(L + a, (\underbrace{m_0, \cdots, m_0}_{a \text{ times}}, \mathbf{m}), s + am_0). \tag{2.A.3}$$

To identify simple neural network architectures, we can combine the depth synchronization and enlarging properties. When there exist $c \geq m_0$ and $b > 0$, such that

$s = cL + b$, and L^* is an upper bound on L, combining the previous two properties yields

$$\mathcal{F}_{\psi}(L, \mathbf{m}, s) \subset \mathcal{F}_{\psi}(L^*, \mathbf{m}', cL + m_0(L^* - L) + b) \subset \mathcal{F}_{\psi}(L^*, \mathbf{m}', cL^* + b),$$

where the width vector $\mathbf{m}'$ has length $L^* + 2$ and can be chosen as $(m_0, m', m', \cdots m', m_{L+1})$ with m' equal to the largest coefficient of $\mathbf{m}$.

Parallelization: Let $\mathbf{m}$, $\mathbf{m}'$ be two width vectors such that $m_0 = m_0'$ and let $\mathbf{f} \in \mathcal{F}_{\mathrm{id}}(L, \mathbf{m})$ and $\mathbf{g} \in \mathcal{F}_{\mathrm{id}}(L, \mathbf{m}')$. Define the parallelized network $\mathbf{h}$ as $\mathbf{h} := (\mathbf{f}, \mathbf{g})$. Then

$$\mathbf{h} \in \mathcal{F}_{\mathrm{id}}(L, (m_0, m_1 + m_1', \cdots, m_{L+1} + m_{L+1}')). \tag{2.A.4}$$

Proposition 2.A.1 (Removal of inactive nodes)**.** *It holds that*

$$\mathcal{F}_{\psi}(L, \mathbf{m}, s) = \mathcal{F}_{\psi}(L, (m_0, m_1 \wedge s, \cdots, m_L \wedge s, m_{L+1}), s).$$

For this property, the output function plays no role and the proof in [127] carries over.

The following equation gives the number of parameters in a fully connected network in $\mathcal{F}_{\psi}(L, \mathbf{m})$:

$$\sum_{j=0}^{L} (m_j + 1) m_{j+1} - m_{L+1}. \tag{2.A.5}$$

This will be used further on as an upper bound on the number of active parameters in sub-networks.

2.A.2   Scaling numbers

We constrain all neural network parameters to be bounded in absolute value by one. To build neural networks with large output values we construct small rescaling networks.

Proposition   2.A.2. *For   any   real   number C there   exists   a   network* $\mathrm{Scale}_C \in \mathcal{F}_{\mathrm{id}}(\lceil \log_2(|C|) \rceil + (\lceil \log_2(|C|) \rceil - 1), (1, 2, 1, 2, 1, \cdots, 1, 2, 1), 4 \lceil \log_2(|C|) \rceil)$ *such that* $\mathrm{Scale}_C(x) = C(x)_+$.

Proof. Set

$$W_0 = \begin{pmatrix} 1 \\ 1 \end{pmatrix}, \quad \mathbf{v}_1 = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \quad \text{and} \quad W_1 = (1, 1).$$

The network $W_1 \sigma_{\mathbf{v}_1} W_0 x$ computes $x \mapsto 2(x)_+$. This network has exactly one hidden layer, one input node, one output node and two nodes in the hidden layer. It

uses four nonzero-parameters. Composing $\lceil \log_2(|C|) \rceil$ of these networks, using the composition rule (2.A.2), where we take the output layer of one network to be the input layer of the next one with shift vector zero, yields a network in the right network class computing $x \mapsto 2^{\lceil \log_2(|C|) \rceil}(x)_+$. Replacing the last weight matrix by $(C2^{-\lceil \log_2(|C|) \rceil}, C2^{-\lceil \log_2(|C|) \rceil})$ yields the result. $\qquad \square$

2.A.3 Negative numbers

For negative input, the ReLU activation without shift returns zero. As a result, many network constructions output zero for negative input. Using that $x = \sigma(x) - \sigma(-x)$, the next result shows existence of a neural network function that extends the original network function as an even (or odd) function to negative input values.

Proposition 2.A.3. *Assume $f \in \mathcal{F}_{\mathrm{id}}(L, (m_0, m_1, \cdots, m_L, 1), s)$ and $f(\mathbf{x}) = 0$ whenever $x_j \leq 0$ for some index $j \in \{1, \cdots, m_0\}$. Then there exist neural networks*

$$f^{\pm} \in \mathcal{F}_{\mathrm{id}}(L, (m_0, 2m_2, \cdots, 2m_L, 1), 2s),$$

such that $x_j \mapsto f^+(\mathbf{x})$ is an even function, $x_j \mapsto f^-(\mathbf{x})$ is an odd function and $f^{\pm}(\mathbf{x}) = f(\mathbf{x})$ for all $\mathbf{x}$ with $x_j \geq 0$.

Proof. Take two neural networks in the class $\mathcal{F}_{\mathrm{id}}(L, (m_0, m_1, \cdots, m_L, 1), s)$ in parallel: The original network f to deal with the positive part and the second network to deal with the negative part. This second network can be build from the first network f by multiplying the j-th column vector of W_0 by -1 and multiplying the output of the network by ± 1. The parallelized network computes then $f^{\pm}$. $\qquad \square$

The extension to more than one output is straightforward. Following the same construction as in the previous section, all that has to be done is multiplying the corresponding rows of the weight matrix in the output layer of the neural network by either -1, 1 of 0 depending on how we wish to extend the function. More precisely, if we have $m_0^- \leq m_0$ input coefficients x_j for which $x_j \leq 0$ implies $f(\mathbf{x}) = 0$, we can find neural networks

$$\mathbf{f}^{\pm} \in \mathcal{F}_{\mathrm{id}}(L, (m_0, 2^{m_0^-}m_2, \cdots, 2^{m_0^-}m_L, m_{L+1}), 2^{m_0^-}s),$$

such that $x_j \mapsto \mathbf{f}^+(\mathbf{x})$ is an even function and $x_j \mapsto \mathbf{f}^-(\mathbf{x})$ is an odd function for all of the m_0^- indices j. This network can be constructed using $2^{m_0^-}$ parallel networks.

2.B Neural networks approximating the logarithm

Theorem 2.4.1 assumes $M \geq 2$. We use this throughout the proof without further mentioning.

2.B.1   Taylor approximation

Set

$$T_c^\kappa(x) = \log(c) + \sum_{\gamma=0}^{\kappa} x^\gamma \sum_{\alpha=\gamma\vee 1}^{\kappa} \binom{\alpha}{\gamma} \frac{c^{-\gamma}(-1)^{1-\gamma}}{\alpha} = \sum_{\gamma=0}^{\kappa} x^\gamma c_\gamma.$$

Proposition 2.B.1. *For all $\kappa = 0, 1, \ldots$ and every $c > 0$, we have that*

$$\left| \log(x) - T_c^\kappa(x) \right| \leq \frac{1}{\kappa+1} \left| \frac{x-c}{x \wedge c} \right|^{\kappa+1},$$

where the sum in T_c^κ is defined as zero if $\kappa = 0$. Moreover, if $0 < x \leq c$, we also have that $T_c^\kappa(x) \leq \log(c)$.

Proof. We claim that T_c^κ is equal to the k-th order Taylor approximation of the logarithm. First we show that from this claim the statements of the proposition follow. The α-th derivative of the logarithm is $\log^{(\alpha)}(x) = (\alpha - 1)!(-1)^{\alpha+1}x^{-\alpha}$. Thus, the k-th order Taylor approximation of the logarithm around the point c is given by

$$\log(c) + \sum_{\alpha=1}^{\kappa} \frac{(x-c)^\alpha(-1)^{\alpha+1}}{\alpha c^\alpha}. \tag{2.B.1}$$

By the mean value theorem, the remainder is bounded by

$$\frac{1}{\kappa+1} \left| \frac{x-c}{s} \right|^{\kappa+1},$$

for some s between x and c. Now since the function $1/s$ on $(0, \infty)$ is decreasing, its maximum is obtained at the left boundary, that is, $x \wedge c$, which yields the first claim of the proposition. Now we show that $T_c^\kappa \leq \log(c)$ whenever $0 < x \leq c$. When $\kappa = 0$, the sum in (2.B.1) disappears and the result follows immediately. When $\kappa \geq 1$, notice that $(x - c)$ is always negative. Hence the product $(x - c)^\alpha(-1)^{\alpha+1}$ is negative for all α, so together with the case $\kappa = 0$ this yields $T_c^\kappa(x) \leq \log(c)$, for $0 < x \leq c$.

It remains to prove that T_c^κ is the k-th order Taylor approximation of the logarithm around the point c. Writing the Taylor approximation as a linear combination of monomials gives

$$\log(c) + \sum_{\alpha=1}^{\kappa} \frac{(x-c)^\alpha(-1)^{\alpha+1}}{\alpha c^\alpha} = \sum_{\gamma=0}^{\kappa} x^\gamma \bar{c}_\gamma,$$

for suitable coefficients $\bar{c}_\gamma$. Using this expression we can obtain the coefficients $\bar{c}_\gamma$ for $\gamma \geq 1$ by evaluating the derivatives at $x = 0$:

$$\left. \frac{d^\gamma}{dx^\gamma} \log(c) + \sum_{\alpha=1}^{\kappa} \frac{(x-c)^\alpha (-1)^{\alpha+1}}{\alpha c^\alpha} \right|_{x=0} = \gamma! \bar{c}_\gamma.$$

This gives us that

$$\bar{c}_\gamma = \sum_{\alpha=\gamma}^{\kappa} \frac{(\alpha-1)!(-c)^{\alpha-\gamma}(-1)^{\alpha+1}}{\gamma!(\alpha-\gamma)!c^\alpha} = \sum_{\alpha=\gamma}^{\kappa} \binom{\alpha}{\gamma} \frac{c^{-\gamma}(-1)^{1-\gamma}}{\alpha}.$$

For $\bar{c}_0$ we get

$$\bar{c}_0 = \log(c) + \sum_{\alpha=1}^{\kappa} \frac{(\alpha-1)!(-c)^\alpha(-1)^{\alpha+1}}{(\alpha)! c^\alpha} = \log(c) + \sum_{\alpha=1}^{\kappa} \frac{(-1)}{\alpha}.$$

Hence $\sum_\gamma^\kappa x^\gamma \bar{c}_\gamma = \sum_\gamma^\kappa x^\gamma c_\gamma = T_c^\kappa(x)$, proving the claim. $\qquad \square$

Next we establish a bound on the sum of the coefficients c_γ of T_c^κ in the case $c \leq e$. For $\gamma \geq 1$, we bound c_γ by

$$|c_\gamma| \leq \sum_{\alpha=\gamma}^{\kappa} \binom{\alpha}{\gamma} \frac{(1 \wedge c)^{-\gamma}}{\alpha} \leq (1 \wedge c)^{-\kappa} \sum_{\alpha=\gamma}^{\kappa} \binom{\alpha}{\gamma}.$$

Since also

$$|c_0| \leq |\log(c)| + \sum_{\alpha=1}^{\kappa} \frac{1}{\alpha} \leq |\log(c)| + \sum_{\alpha=1}^{\kappa} \binom{\alpha}{0},$$

this shows that the sum of the coefficients is bounded by

$$\sum_{\gamma=0}^{\kappa} |c_\gamma| \leq |\log(c)| + (1 \wedge c)^{-\kappa} \sum_{\gamma=0}^{\kappa} \sum_{\alpha=1 \wedge \gamma}^{\kappa} \binom{\alpha}{\gamma} \leq |\log(c)| + (1 \wedge c)^{-\kappa} \sum_{\gamma=0}^{\kappa} \sum_{\alpha=\gamma}^{\kappa} \binom{\alpha}{\gamma}.$$

The double sum can be rewritten as the sum of all the entries in the rows $0, \cdots, \kappa$ of Pascal's triangle. From the binomial theorem we know that summing over the α-th row of Pascal's triangle gives 2^α. Combined with $|\log(c)| \leq (1 \wedge c)^{-1}$ for $0 < c \leq e$, this gives

$$\sum_{\gamma=0}^{\kappa} |c_\gamma| \leq (\kappa+1)2^{\kappa+1}(1 \wedge c)^{-(\kappa \vee 1)} \leq (\kappa+1)2^{\kappa+1}(1 \wedge c)^{-\kappa-1}, \text{ for all } 0 < c \leq e. \quad (2.B.2)$$

Applying the softmax function to an approximation g of the logarithm involves the exponential function and requires a bound for $|e^{g(x)} - x|$ with $x > 0$. By the mean value theorem $|e^{g(x)} - e^{\log(x)}| = e^s|g(x) - \log(x)|$ for a suitable s between $\log(x)$ and $g(x)$. The next proposition provides such a bound.

Proposition 2.B.2. *For all $\lambda \geq 1$, define*

$$\mathcal{D}_\lambda := \left[\frac{\lambda^{\lceil \beta \rceil}}{2^{\lceil \beta \rceil^2} \lceil \beta \rceil^{\lfloor \beta \rfloor} M}, \frac{(\lambda + 1)^{\lceil \beta \rceil}}{2^{\lceil \beta \rceil^2} \lceil \beta \rceil^{\lfloor \beta \rfloor} M} \right].$$

If $[a, b] \subset \mathcal{D}_\lambda$, then it holds for any $x \in [a, b]$ and any $\omega \leq \log\left(\frac{(\lambda+1)^{\lceil \beta \rceil}}{2^{\lceil \beta \rceil^2} \lceil \beta \rceil^{\lfloor \beta \rfloor} M} \right)$, that

$$e^\omega |T_b^{\lfloor \beta \rfloor}(x) - \log(x)| \leq \frac{1}{M}.$$

Proof. First notice that on $(0, \infty)$ the logarithm is strictly increasing and is infinitely times continuously differentiable. For real numbers a, b and a positive integer j, $a^j - b^j = (a - b) \sum_{i=1}^j a^{j-i} b^{i-1}$. Applied to $a = \lambda + 1$ and $b = \lambda$, this gives $(\lambda + 1)^j - \lambda^j \leq j(\lambda + 1)^{j-1}$ and thus for $x \in [a, b] \subseteq \mathcal{D}_\lambda$, we get that

$$|x - b| \leq b - a \leq \frac{(\lambda + 1)^{\lceil \beta \rceil} - \lambda^{\lceil \beta \rceil}}{2^{\lceil \beta \rceil^2} \lceil \beta \rceil^{\lfloor \beta \rfloor} M} \leq b \frac{\lceil \beta \rceil}{\lambda + 1}.$$

Substituting this in the bound from Proposition 2.B.1 and using that $x \geq a$ gives

$$|T_b^{\lfloor \beta \rfloor}(x) - \log(x)| \leq \frac{1}{\lceil \beta \rceil} \left| \frac{\lceil \beta \rceil (\lambda + 1)^{\lfloor \beta \rfloor}}{a 2^{\lceil \beta \rceil^2} \lceil \beta \rceil^{\lfloor \beta \rfloor} M} \right|^{\lceil \beta \rceil}.$$

Since $a \in \mathcal{D}_\lambda$,

$$|T_b^{\lfloor \beta \rfloor}(x) - \log(x)| \leq \frac{1}{\lceil \beta \rceil} \left| \frac{\lceil \beta \rceil (\lambda + 1)^{\lfloor \beta \rfloor}}{2^{\lceil \beta \rceil^2} \lceil \beta \rceil^{\lfloor \beta \rfloor} M} \cdot \frac{2^{\lceil \beta \rceil^2} \lceil \beta \rceil^{\lfloor \beta \rfloor} M}{\lambda^{\lceil \beta \rceil}} \right|^{\lceil \beta \rceil}$$

$$= \lceil \beta \rceil^{\lfloor \beta \rfloor} \left| \frac{(\lambda + 1)^{\lfloor \beta \rfloor})}{\lambda^{\lceil \beta \rceil}} \right|^{\lceil \beta \rceil}.$$

Multiplying both sides with an exponential, noticing that the exponential function is strictly increasing, and applying the upper bound on ω given in the statement of the proposition yields

$$e^\omega |T_b^{\lfloor \beta \rfloor}(x) - \log(x)| \leq \frac{(\lambda + 1)^{\lceil \beta \rceil} \lceil \beta \rceil^{\lfloor \beta \rfloor}}{2^{\lceil \beta \rceil^2} \lceil \beta \rceil^{\lfloor \beta \rfloor} M} \left| \frac{(\lambda + 1)^{\lfloor \beta \rfloor})}{\lambda^{\lceil \beta \rceil}} \right|^{\lceil \beta \rceil}$$

$$= \frac{1}{2^{\lceil \beta \rceil^2} M} \left(\frac{\lambda + 1}{\lambda} \right)^{\lceil \beta \rceil^2}.$$

Since $(\lambda+1)\lambda^{-1}$ is positive and decreasing for $\lambda \geq 1$, we can upper bound the last display by $1/M$. □

2.B.2 Partition of unity

So far we have bounded the approximation error on subintervals. As we work with ReLU functions, indicator functions of intervals are impractical to use, because they are discontinuous. Instead we create a partition of unity consisting of continuous piecewise linear functions for an interval that contains the interval $[M^{-1}, 1 - M^{-1}]$.

Define R as the smallest integer sucht that

$$\frac{(\frac{R}{2} + 2^{\lceil \beta \rceil}\lceil \beta \rceil^{\lfloor \beta \rfloor/\lceil \beta \rceil} - \frac{3}{4})^{\lceil \beta \rceil}}{2^{\lceil \beta \rceil^2}\lceil \beta \rceil^{\lfloor \beta \rfloor}M} \geq 1 - \frac{1}{M}.$$

Rewriting this equation yields

$$R = \lceil * \rceil 2^{\lceil \beta \rceil+1}\lceil \beta \rceil^{\lfloor \beta \rfloor/\lceil \beta \rceil}(M-1)^{\frac{1}{\lceil \beta \rceil}} - 2\left(2^{\lceil \beta \rceil}\lceil \beta \rceil^{\lfloor \beta \rfloor/\lceil \beta \rceil} - \frac{3}{4}\right)$$

$$\leq 2^{\lceil \beta \rceil+1}\lceil \beta \rceil^{\lfloor \beta \rfloor/\lceil \beta \rceil}M^{\frac{1}{\lceil \beta \rceil}}.$$

Now we define sequences $(a_r)_{r=1,\cdots,R}$ and $(b_r)_{r=1,\cdots,R-1}$ as follows

$$a_r := \frac{(2^{\lceil \beta \rceil}\lceil \beta \rceil^{\lfloor \beta \rfloor/\lceil \beta \rceil} + \frac{r}{2} - \frac{3}{4})^{\lceil \beta \rceil}}{2^{\lceil \beta \rceil^2}\lceil \beta \rceil^{\lfloor \beta \rfloor}M},$$

$$b_r := \frac{(2^{\lceil \beta \rceil}\lceil \beta \rceil^{\lfloor \beta \rfloor/\lceil \beta \rceil} + \frac{r}{2} - \frac{1}{2})^{\lceil \beta \rceil}}{2^{\lceil \beta \rceil^2}\lceil \beta \rceil^{\lfloor \beta \rfloor}M},$$

and for ease of notation define $b_0 = a_1$ and $b_R = a_R$. Notice that $[M^{-1}, 1 - M^{-1}] \subseteq [a_1, a_R] \subseteq [M^{-1}, 1 + M^{-1}]$.

Next we define a family of functions $(F_r)_{r=2,3,\cdots,R}$ and $(H_r)_{r=1,2,\cdots,R}$ on the interval $[a_1, a_R]$. For $r = 2,\cdots,R$ define the function F_r to be zero outside of the interval $[a_{r-1}, a_r]$ and to be a linear interpolation between the value one at the point b_{r-1} and the value zero at the boundaries of this interval. In the same way define for $r = 2,\cdots,R-1$ the function H_r, but with support on the interval $[b_{r-1}, b_r]$ and with interpolation point a_r. Define H_1 to be the linear interpolation between the value one at the point a_1 and the value zero at b_1 and let it be zero outside this interval. Finally define H_R as the linear interpolation between the value one at the point b_R and the value zero at b_{R-1} and set it to zero outside of this interval.
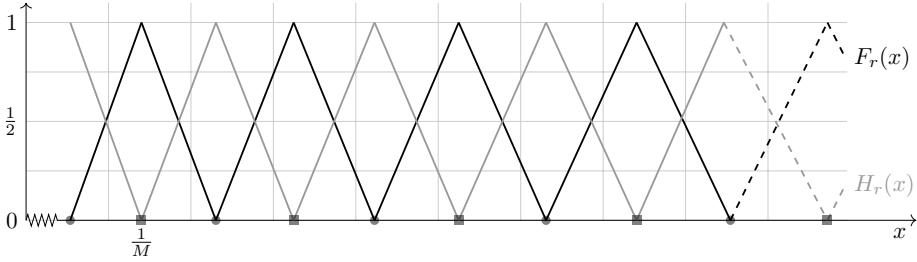
Figure 2.B.1: The first few functions $F_r(x)$ and $H_r(x)$ when $\beta \in (1, 2]$. The points a_r are marked with circles, while the points b_r are denoted by squares.

By construction it holds that

$$\sum_{r=2}^{R} F_r(x) + \sum_{r=1}^{R} H_r(x) = 1, \quad \text{for all } x \in [a_1, a_R].$$

Figure 2.B.1 gives the first few functions F_r and H_r in the case that $\beta \in (1, 2]$.

We can construct a ReLU network that exactly represents the functions F_r and H_r. This construction is a modification of the construction of continuous piecewise linear functions as used in [160]. This modification assures that the parameters are bounded by one.

Proposition 2.B.3. *For each function F_r and H_r their exists a network $U_{F_r}, U_{H_r} \in \mathcal{F}_{\text{id}}(L, \mathbf{m}, s)$, with $L = 3((1 + \lceil \beta \rceil)^2 + \lfloor \log_2(M \lceil \beta \rceil^{\lfloor \beta \rfloor}) \rfloor)$, $\mathbf{m} = (1, 3, 3, \cdots, 3, 1)$ and $s = 8((1 + \lceil \beta \rceil)^2 + \log_2(M \lceil \beta \rceil^{\lfloor \beta \rfloor}))$, such that $F_r(x) = U_{F_r}(x)$ and $H_r(x) = U_{H_r}(x)$ for all $x \in [a_1, a_R]$.*

Proof. The functions F_r and H_r, $r = 2, \cdots, R$, are piecewise linear functions, consisting of four pieces each. This means that these function can be perfectly represented as a linear combination of three ReLU functions. The interpolation points provide the values of the shift vectors. Writing this out for F_r gives

$$F_r(x) = \frac{\sigma(x - a_{r-1})}{b_{r-1} - a_{r-1}} + \left(\frac{1}{b_{r-1} - a_{r-1}} + \frac{1}{a_r - b_{r-1}} \right) \sigma(x - a_{r-1}) + \frac{\sigma(x - a_{r-1})}{a_r - b_{r-1}}.$$

For H_r, $r = 2, \cdots, R$ this can be done in a similar way. For H_1 and H_R we actually only need one ReLU function. The networks weights in this construction are greater than one. The difference between two consecutive points a_r and b_r can be lower

bounded by using that for $x, y \geq 0$: $(x+y)^{\lceil \beta \rceil} - x^{\lceil \beta \rceil} \geq y^{\lceil \beta \rceil}$. Because of

$$\frac{(2^{\lceil \beta \rceil} \lceil \beta \rceil^{\lfloor \beta \rfloor / \lceil \beta \rceil})^{\lceil \beta \rceil}}{2^{\lceil \beta \rceil^2} \lceil \beta \rceil^{\lfloor \beta \rfloor} M} - \frac{(2^{\lceil \beta \rceil} \lceil \beta \rceil^{\lfloor \beta \rfloor / \lceil \beta \rceil} - \frac{1}{4})^{\lceil \beta \rceil}}{2^{\lceil \beta \rceil^2} \lceil \beta \rceil^{\lfloor \beta \rfloor} M} \geq \frac{(\frac{1}{4})^{\lceil \beta \rceil}}{2^{\lceil \beta \rceil^2} \lceil \beta \rceil^{\lfloor \beta \rfloor} M},$$

we can upper bound all the network weights by

$$2^{1 + 2\lceil \beta \rceil + \lceil \beta \rceil^2} \lceil \beta \rceil^{\lfloor \beta \rfloor} M, \tag{2.B.3}$$

which is the inverse of the lower bound on the smallest difference between two consecutive points multiplied by two. Dividing the multiplicative constants by this bound and combining (2.A.2) the resulting network with the $\mathrm{Scale}_C(x)$ network from Proposition 2.A.2 with C equal to (2.B.3) yields a network with the required output and parameters bounded by one. The network class is simplified by using the depth-synchronization (2.A.3) followed by the enlarging property of neural networks (2.A.1). $\qquad \square$

The previous partition yields an approximation $T^\beta : [a_1, a_R] \to \mathbb{R}$ of the logarithm on the entire interval $[a_1, a_R]$ via

$$T^\beta(x) := \sum_{r=2}^{R} F_r(x) T_{a_r}^{\lfloor \beta \rfloor}(x) + \sum_{r=1}^{R} H_r(x) T_{b_r}^{\lfloor \beta \rfloor}(x). \tag{2.B.4}$$

This function depends on M through the sequence of points a_r and b_r.

We can now derive the same type of error bound as in Lemma 2.B.2 for all $x \in [0, 1]$. For this, define the projection $\pi : [0, 1] \to [a_1, a_R]$, that maps $x \in [0, 1]$ to itself, if it is already in the interval $[a_1, a_R]$, and to the closest boundary point otherwise.

Lemma 2.B.4. *For all $x \in [0, 1]$, we have $|e^{T^\beta(\pi(x))} - x| \leq M^{-1}$.*

Proof. First consider $x \in (a_1, a_R)$. By construction there exists a unique $r^* \in \{2, 3, \cdots, R\}$ and a unique $\bar{r} \in \{1, \cdots, R\}$ such that $x \in (a_{r^*-1}, a_{r^*}]$, and $x \in (b_{\bar{r}-1}, b_{\bar{r}}]$. By the mean value theorem and (2.B.4),

$$\left| e^{T^\beta(x)} - x \right| \leq e^\xi \left| T^\beta(x) - \log(x) \right|$$

$$= e^\xi \left| \sum_{r=2}^{R} F_r(x) T_{a_r}^{\lfloor \beta \rfloor}(x) + \sum_{r=1}^{R} H_r(x) T_{b_r}^{\lfloor \beta \rfloor}(x) - \log(x)(F_{r^*}(x) + H_{\bar{r}}(x)) \right|$$

$$\leq F_{r^*}(x) e^\xi \left| T_{a_{r^*}}^{\lfloor \beta \rfloor}(x) - \log(x) \right| + H_{\bar{r}}(x) e^\xi \left| T_{b_{\bar{r}}}^{\lfloor \beta \rfloor}(x) - \log(x) \right|,$$

where ξ is some number between $T^\beta(x)$ and $\log(x)$. We now want to apply Proposition 2.B.2. For this we need to find a $\lambda \geq 1$ such that $[a_{r^*-1}, a_{r^*}] \cup [b_{\bar{r}-1}, b_{\bar{r}}] \in \mathcal{D}_\lambda$ and

$\xi \leq \max_{y \in \mathcal{D}_\lambda} \log(y)$, with $\mathcal{D}_\lambda$ as defined by that proposition. Because of our choice of the sequences of points a_r and b_r,

$$\lambda := \max \left\{ \frac{r^*}{2} + 2^{\lceil \beta \rceil} \lceil \beta \rceil^{\lfloor \beta \rfloor / \lceil \beta \rceil} - \frac{3}{4}, \frac{\bar{r}}{2} + 2^{\lceil \beta \rceil} \lceil \beta \rceil^{\lfloor \beta \rfloor / \lceil \beta \rceil} - \frac{1}{2} \right\} - 1$$

satisfies $\lambda \geq 1$, since $r^* \geq 2$ and $\bar{r} \geq 1$. Furthermore this choice of λ guarantees that $[a_{r^*-1}, a_{r^*}] \cup [b_{\bar{r}-1}, b_{\bar{r}}] \subseteq \mathcal{D}_\lambda$. For the bound on ξ, notice that $x \in [a_{r^*-1}, a_{r^*}] \cup [b_{\bar{r}-1}, b_{\bar{r}}]$ and that $T^\beta(x) = F_{r^*}(x) T_{a_{r^*}}^{\lfloor \beta \rfloor}(x) + H_{\bar{r}}(x) T_{b_{\bar{r}}}^{\lfloor \beta \rfloor}(x)$. Combined with the second statement of Proposition 2.B.1, that is $T_c^\kappa \leq \log(c)$ for $0 < c \leq x$, and together with $F_{r^*}(x) + H_{\bar{r}}(x) = 1$, this yields $\xi \leq \max\{\log(a_{r^*}), \log(b_{\bar{r}})\}$. Thus we can apply Proposition 2.B.2 and obtain

$$F_{r^*}(x) e^\xi \left| T_{a_{r^*}}^{\lfloor \beta \rfloor}(x) - \log(x) \right| + H_{\bar{r}}(x) e^\xi \left| T_{b_{\bar{r}}}^{\lfloor \beta \rfloor}(x) - \log(x) \right|$$
$$\leq F_{r^*}(x) \frac{1}{M} + H_{\bar{r}}(x) \frac{1}{M} = \frac{1}{M},$$

completing the proof for $x \in [a_1, a_R]$.

When $x \in [0, a_1]$, notice that $0 < a_1 < M^{-1}$ and $T^\beta(\pi(x)) = T_{b_1}^{\lfloor \beta \rfloor}(a_1)$. Hence by Proposition 2.B.1 together with $b_1 = M^{-1}$, we get that $T^\beta(\pi(x)) \leq \log(M^{-1})$ proving that both x and $e^{T^\beta(\pi(x))}$ are in $[0, M^{-1}]$. Thus the conclusion also holds for $x \in [0, a_1]$.

For $a_R \geq 1$, the proof follows from $[0, 1] \subseteq ([0, a_1] \cup [a_1, a_R])$. Thus it remains to study $a_R < 1$. Consider $x \in [a_R, 1]$. Using that $1 - M^{-1} \leq a_R < 1$ and that $T^\beta(\pi(x)) = T_{b_R}^{\lfloor \beta \rfloor}(a_R) = T_{a_R}^{\lfloor \beta \rfloor}(a_R)$ yields $T^\beta(\pi(x)) = \log(a_R)$. This gives us that both x and $e^{T^\beta(\pi(x))}$ are in $[a_R, 1] \subset [1 - M^{-1}, 1]$, which immediately yields the required bound. $\square$

Network Construction

The following result shows how to approximate multiplications with deep ReLU networks. This is required later to construct neural networks mimicking the Taylor-approximation T^β considered in the previous section.

Lemma 2.B.5 (Lemma A.3. of [127])**.** *For every $\eta \in \mathbb{N}_{\geq 1}$ and $D \in \mathbb{N}_{\geq 1}$, there exists a network $Mult_\eta^D \in \mathcal{F}_{\mathrm{id}}((\eta + 5)\lceil \log_2(D) \rceil, (D, 6D, 6D, \cdots, 6D, 1))$, such that $Mult_\eta^D \in [0, 1]$ and*

$$\left| Mult_\eta^D(x_1, \cdots, x_D) - \prod_{i=1}^D x_i \right| \leq 3^D 2^{-\eta}, \quad \text{for all } (x_1, \cdots, x_D) \in [0, 1]^D.$$

Moreover $\text{Mult}_\eta^D(x) = 0$ *if one of the coefficients of* $\mathbf{x}$ *is zero.*

Remark 2.B.6. Using (2.A.5) the number of parameters in the neural network Mult_η^D is bounded by $((\eta + 5)\lceil \log_2(D) \rceil + 1)42D^2 \leq (\eta + 5)126D^2 \log_2(D)$.

We now have all the required ingredients to finish the proof of Theorem 2.4.1:

Proof of Theorem 2.4.1. Since $a_1 = \sigma(0 \cdot x + a_1)$, the projection π can be written in terms of ReLU functions as

$$\pi(x) = \max\big(a_1, \min(x, a_R)\big) = \sigma(0 \cdot x + a_1) + \sigma(x - a_1) - \sigma(x - a_R).$$

For $a_R \leq 1$, all network parameters are bounded by one and this defines a neural network in $\mathcal{F}_{\text{id}}(1, (1, 3, 1), 8)$. When $a_R > 1$, we replace $\sigma(x - a_R)$ with $\sigma(x - 1)$ as we are only interested in input in the interval $[0, 1]$. Having thus obtained a value in the interval $[a_1, a_R]$, we can, for any $r \in \{1, \cdots, R\}$, apply the network U_{F_r} from Proposition 2.B.3 to it. Using depth synchronization (2.A.3) and parallelization (2.A.4), we can combine the network U_{F_r} with a parallel network that forwards the input value to obtain a network in the network class $\mathcal{F}_{\text{id}}(L, \mathbf{m}, s)$, with $L = 4\big((1 + \lceil \beta \rceil)^2 + \log_2(M\lceil \beta \rceil^{\lfloor \beta \rfloor})\big)$, $\mathbf{m} = (1, 3, 1, 4, \cdots, 4, 2)$ and $s = 13\big((1 + \lceil \beta \rceil)^2 + \log_2(M\lceil \beta \rceil^{\lfloor \beta \rfloor})\big)$, that maps $x \in [0, 1]$ to $(F_r(\pi(x)), \pi(x))$. The next step is to construct a network that approximates $F_r(x)T_{a_r}^\beta(x)$. Since $a_r \in [M^{-1}, 1 + M^{-1}]$, (2.B.2) allows us, for $\gamma = 1, \cdots, \lfloor \beta \rfloor$, to use the network $\text{Mult}_\eta^{\gamma+1}$ with input vector $(F_r(\pi(x)), \pi(x), \cdots, \pi(x))$ to compute approximately the function $F_r(\pi(x))\pi(x)^\gamma$, and multiply its output with $c_\gamma / \lceil \beta \rceil 2^{\lfloor \beta \rfloor + 1} M^{\lceil \beta \rceil}$. For each $\gamma \in \{1, \cdots, \lfloor \beta \rfloor\}$ we have a network that approximately computes the function $x \mapsto F_r(\pi(x))\pi(x)^\gamma c_\gamma / \lceil \beta \rceil 2^{\lfloor \beta \rfloor + 1} M^{\lceil \beta \rceil}$. We now consider the network that computes these functions in parallel and combines this with a single shallow hidden node network to approximately compute $F_r(\pi(x))c_0 / \lceil \beta \rceil 2^{\lfloor \beta \rfloor + 1} M^{\lceil \beta \rceil}$. Making use of parallelization (2.A.4), depth synchronization (2.A.3) and Remark 2.B.6, this yields a network $G_{F_r} \in \mathcal{F}_{\text{id}}(L^*, (1, 6(\lceil \beta \rceil)^2, \cdots, 6(\lceil \beta \rceil)^2, 1), s^*)$, with

$$L^* = 4((1 + \lceil \beta \rceil)^2 + \log_2(M\lceil \beta \rceil^{\lfloor \beta \rfloor})) + 2(\eta + 5)\log_2(\lceil \beta \rceil)$$
$$s^* = 13((1 + \lceil \beta \rceil)^2 + \log_2(M\lceil \beta \rceil^{\lfloor \beta \rfloor})) + (\eta + 5)\log_2(\lceil \beta \rceil)126(\lceil \beta \rceil)^3$$

such that

$$\left| G_{F_r}(x) - F_r(\pi(x)) \sum_{\gamma=0}^{\lfloor \beta \rfloor} \frac{c_\gamma}{\lceil \beta \rceil 2^{\lfloor \beta \rfloor + 1} M^{\lceil \beta \rceil}} \pi(x)^\gamma \right| \leq 3^{\lceil \beta \rceil} 2^{-\eta}.$$

Due to the normalization constant $\lceil \beta \rceil 2^{\lfloor \beta \rfloor + 1} M^{\lceil \beta \rceil}$ it holds that $G_{F_r}(x) \in [-1, 1]$ when $\pi(x)$ is in the support of F_r. If $\pi(x)$ is outside the support of F_r, then Lemma 2.B.5

guarantees that $G_{F_r}(x) = 0$. Similarly for F_r replaced by H_r, we can construct deep ReLU networks G_{H_r} with the same properties.

Using the R networks G_{H_r} and $R - 1$ networks G_{F_r} in parallel together with the observation that each x can be in the support of at most one F_r and one H_r, this yields a deep ReLU network with output $\sum_{r=2}^{R} G_{F_r}(x) + \sum_{r=1}^{R} G_{H_r}(x)$, such that

$$\left| \sum_{r=2}^{R} G_{F_r}(x) + \sum_{r=1}^{R} G_{H_r}(x) - \frac{T^{\beta}(\pi(x))}{\lceil \beta \rceil 2^{\lfloor \beta \rfloor + 1} M^{\lceil \beta \rceil}} \right| \leq 3^{\lceil \beta \rceil} 2^{-\eta + 1}.$$

In the next step we compose the network construction with a scaling network. For this we use the scaling network from Proposition 2.A.2 with constant $C = \lceil \beta \rceil 2^{\lfloor \beta \rfloor + 1} M^{\lceil \beta \rceil}$. Since the input can be negative we use two of those networks in parallel as described in Proposition 2.A.3. This gives us a network

$$\widetilde{G} \in \mathcal{F}_{\mathrm{id}} \left(L^* + 4 \log_2 \left(\lceil \beta \rceil 2^{\lfloor \beta \rfloor + 1} M^{\lceil \beta \rceil} \right), \mathbf{m}^*, 2Rs^* + 16 \log_2 \left(\lceil \beta \rceil 2^{\lfloor \beta \rfloor + 1} M^{\lceil \beta \rceil} \right) \right),$$

where $\mathbf{m}^* = (1, 12R(\lceil \beta \rceil)^2, \cdots, 12R(\lceil \beta \rceil)^2, 1)$, such that

$$\left| \widetilde{G}(x) - T^{\beta}(\pi(x)) \right| \leq \lceil \beta \rceil 2^{\lfloor \beta \rfloor + 2} M^{\lceil \beta \rceil} 3^{\lceil \beta \rceil} 2^{-\eta}.$$

Setting $\eta = \lceil \log_2(\lceil \beta \rceil 2^{\lfloor \beta \rfloor + 2} M^{\lceil \beta \rceil + 1} 3^{\lceil \beta \rceil}) \rceil$, this is upper bounded by M^{-1}. Applying the triangle inequality, the mean value theorem and Lemma 2.B.4 yields

$$\left| e^{\widetilde{G}(x)} - x \right| \leq \left| e^{\widetilde{G}(x)} - e^{T^{\beta} \log(\pi(x))} \right| + \left| e^{T^{\beta} \log(\pi(x))} - x \right| \leq \frac{e^{2/M}}{M} + \frac{1}{M} \leq \frac{4}{M}, \quad (2.B.5)$$

where the term $e^{2/M}$ comes from noticing that $|\widetilde{G}(x) - T^{\beta} \log(\pi(x))| \leq M^{-1}$, $|T^{\beta} \log(\pi(x)) - \log(1)| \leq M^{-1}$ and triangle inequality.

To derive the lower bound $G(x) \geq \log(4/M)$, we construct a network that computes the maximum between $\widetilde{G}(x)$ and $\log(4/M)$. Since $M \geq 1$ implies $|\log(4/M)|/\lceil \beta \rceil 2^{\lfloor \beta \rfloor + 1} M^{\lceil \beta \rceil} \leq 1$, we can achieve this by adding one additional layer before the scaling. This layer can be written as

$$\sigma \left(x - \frac{\log(4/M)}{\lceil \beta \rceil 2^{\lfloor \beta \rfloor + 1} M^{\lceil \beta \rceil}} \right) + \frac{\log(4/M)}{\lceil \beta \rceil 2^{\lfloor \beta \rfloor + 1} M^{\lceil \beta \rceil}} \sigma(1). \quad (2.B.6)$$

Applying the scaling as before yields a network $G(x) = \max\{\widetilde{G}(x), \log(4/M)\}$ that is in the same network class as $\widetilde{G}(x)$. For the upper bound notice that if $G(x) = \widetilde{G}(x)$,
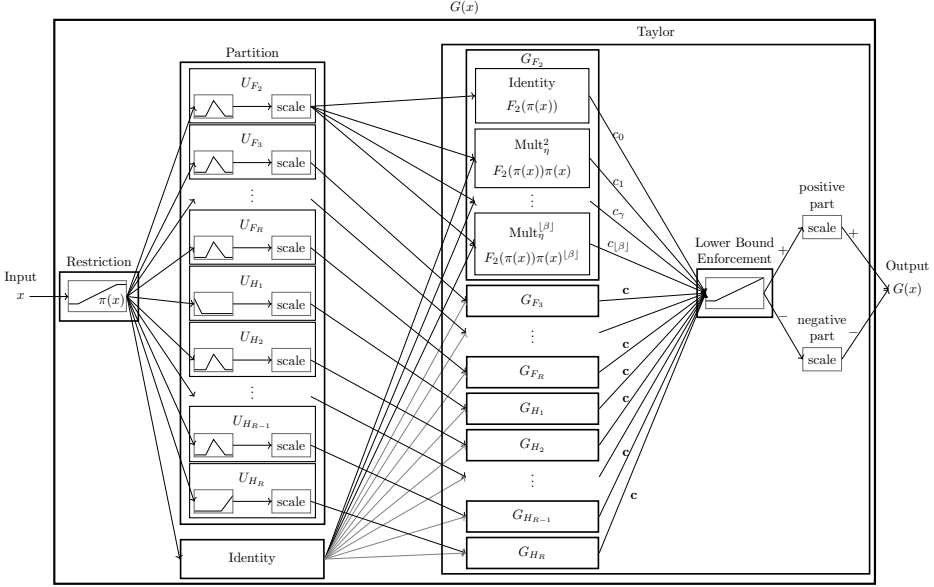
Figure 2.B.2: The construction of the logarithm approximation network G of Theorem 2.4.1 from subnetworks. The difference between the networks G and $\widetilde{G}$ is the single layer which enforces the lower bound, which is not present in the network $\widetilde{G}$.

then the bound follows from (2.B.5). When $G(x) = \log(4/M)$, then $\widetilde{G}(x) \le \log(4/M)$, so (2.B.5) implies that $x \le 8/M$. Hence

$$\left| e^{G(x)} - x \right| = \left| \frac{4}{M} - x \right| \le \frac{4}{M}.$$

The network size as given in the theorem is an upper bound on the network size obtained here, which is allowed by the depth-synchronization followed by the enlarging property, and is done in order to simplify the expressions. $\qquad\square$

Figure 2.B.2 shows the main substructures of the deep ReLU network construction in this proof.

2.C   Further technicalities

Proposition 2.C.1 (Bernstein's inequality). *For independent random variables* $(Z_i)_{i=1}^n$ *with zero mean and moment bounds* $\mathbb{E}|Z_i|^m \leq \frac{1}{2}m!U^{m-2}v_i$ *for* $m = 2, 3, \ldots$ *and* $i = 1, \ldots, n$ *for some constants* U *and* v_i, *we have*

$$\mathbb{P}\left(\left|\sum_{i=1}^n Z_i\right| > x\right) \leq 2e^{-\frac{x^2}{2v+2Ux}}, \quad \text{for } v \geq \sum_{i=1}^n v_i.$$

This formulation of Bernstein's inequality is based on the formulation in Lemma 2.2.11 of [148]. The proof can be found in [16].

The next elementary inequality generalizes Lemma 10 of [127].

Lemma 2.C.2. *If* a, b, c, d *are real numbers,* $a \geq 0$, *such that* $|a - b| \leq 2\sqrt{ac} + d$, *then, for each* $\epsilon \in (0, 1]$,

$$(1 - \epsilon)(b - d) - \frac{(1-\epsilon)^2}{\epsilon}c^2 \leq a \leq (1 + \epsilon)(b + d) + \frac{(1+\epsilon)^2}{\epsilon}c^2.$$

Proof. First notice that $|a-b| \leq 2\sqrt{ac}+d$ if and only if $-2\sqrt{ac}-d \leq a-b \leq 2\sqrt{ac}+d$. Using that $2xy \leq x^2 + y^2$ for all $x, y \in \mathbb{R}$, we get for $x := \sqrt{a}\sqrt{\epsilon}/\sqrt{1+\epsilon}$ and $y := c\sqrt{1+\epsilon}/\sqrt{\epsilon}$, that

$$2\sqrt{ac} = 2xy \leq x^2 + y^2 = \frac{\epsilon a}{1+\epsilon} + \frac{(1+\epsilon)c^2}{\epsilon}$$

and therefore

$$a - b \leq \frac{\epsilon a}{1+\epsilon} + \frac{(1+\epsilon)c^2}{\epsilon} + d.$$

Rearranging the terms yields the upper bound of the lemma. For the lower bound notice that if $\epsilon = 1$, then the lower bound is zero, and holds since $a \geq 0$. For $\epsilon \in (0, 1)$ using the same argument but now with $x = \sqrt{a}\sqrt{\epsilon}/\sqrt{1-\epsilon}$ and $y = c\sqrt{1-\epsilon}/\sqrt{\epsilon}$, gives

$$a - b \geq -\frac{\epsilon a}{1-\epsilon} - \frac{(1-\epsilon)c^2}{\epsilon} - d.$$

Rearranging the terms yields the lower bound of the proposition. $\square$

The number a is required to be nonnegative as otherwise $\sqrt{a}$ would not be a real number. In the statement in [127] the constants a, b, c, d are all required to be positive. However since the inequality $2xy \leq x^2+y^2$ holds for all real numbers x, y the positivity constraint is not necessary. However, when c and d are negative the term $2\sqrt{ac} + d$ is negative, and no pair a, b exists such that the condition is satisfied.

Recall that $d_\tau(\mathbf{f}, \mathbf{g}) := \sup_{\mathbf{x} \in \mathcal{D}} \max_{k=1,\cdots,K} |(\tau \vee f_k(\mathbf{x})) - (\tau \vee g_k(\mathbf{x}))|$. Observe that $d_\tau(\mathbf{f}, \mathbf{g}) = 0$ does not imply $\mathbf{f} = \mathbf{g}$, which is why d_τ is not a metric. The next lemma shows that this, however, defines a pseudometric.

Lemma 2.C.3. *Let $\mathbf{f}, \mathbf{g}, \mathbf{h} : \mathcal{D} \to \mathbb{R}^K$, then for every $\tau \in \mathbb{R}$:*
  (i) $d_\tau(\mathbf{f}, \mathbf{g}) \geq 0$
  (ii) $d_\tau(\mathbf{f}, \mathbf{f}) = 0$
  (iii) $d_\tau(\mathbf{f}, \mathbf{g}) = d_\tau(\mathbf{g}, \mathbf{f})$
  (iv) $d_\tau(\mathbf{f}, \mathbf{g}) \leq d_\tau(\mathbf{f}, \mathbf{h}) + d_\tau(\mathbf{h}, \mathbf{g})$.

Proof. (i), (ii) and (iii) follow immediately. (iv) follows from applying triangle inequality to the $\|\cdot\|_\infty$ norm,

$$
\begin{aligned}
d_\tau(\mathbf{f}, \mathbf{g}) &= \big\| \max_{k=1,\cdots,K} |(\tau \vee f_k(\cdot)) - (\tau \vee g_k(\cdot))| \big\|_\infty \\
&\leq \big\| \max_{k=1,\cdots,K} |(\tau \vee f_k(\cdot)) - (\tau \vee h_k(\cdot))| \big\|_\infty \\
&\quad + \big\| \max_{k=1,\cdots,K} |(\tau \vee h_k(\cdot)) - (\tau \vee g_k(\cdot))| \big\|_\infty \\
&= d_\tau(\mathbf{f}, \mathbf{h}) + d_\tau(\mathbf{h}, \mathbf{g}).
\end{aligned}
$$

$\square$

Lemma 2.C.4. *If $\mathcal{G}$ is a function class of functions from $\mathcal{D}$ to $[0, \infty)^K$, then for all $\delta > 0$ and $\tau > 0$*

$$
\mathcal{N}\big(\delta, \log(\mathcal{G}), d_{\log(\tau)}(\cdot, \cdot)\big) \leq \mathcal{N}\big(\delta\tau, \mathcal{G}, d_\tau(\cdot, \cdot)\big).
$$

Proof. Let $\delta > 0$. Denote by $(\mathbf{g}_j)_{j=1}^{\mathcal{N}_n}$ the centers of a minimal internal $\delta\tau$-covering of $\mathcal{G}$ with respect to d_τ and let $\mathbf{g} \in \mathcal{G}$. By the cover property, there exist a $j \in \{1, \cdots, \mathcal{N}_n\}$ such that $d_\tau(\mathbf{g}, \mathbf{g}_j) \leq \delta\tau$.

The derivative of $\log(u)$ is $1/u$, so the logarithm is Lipschitz on $[\tau, \infty)$ with Lipschitz constant τ^{-1}. Applying this to $d_{\log(\tau)}(\log(\mathbf{g}), \log(\mathbf{g}_j))$, noticing that $\max\{\log(\tau), \log(x)\} \in [\log(\tau), \infty)$ for $x \in [0, \infty)$, yields

$$
\begin{aligned}
\max_{\mathbf{x} \in \mathcal{D}} &\max_{k=1,\cdots,K} |(\log(\tau) \vee \log(g_k(\mathbf{x}))) - (\log(\tau) \vee \log(g_{j,k}(\mathbf{x})))| \\
&\leq \tau^{-1} \max_{\mathbf{x} \in \mathcal{D}} \max_{k=1,\cdots,K} |(\tau \vee g_k(\mathbf{x})) - (\tau \vee g_{j,k})(\mathbf{x})| \\
&\leq \tau^{-1} \delta\tau = \delta.
\end{aligned}
$$

Since $\mathbf{g} \in \mathcal{G}$ was arbitrary, this means that for all $\mathbf{g} \in \mathcal{G}$ there exists a $j \in \{1, \cdots, \mathcal{N}_n\}$ such that $d_{\log(\tau)}(\log(\mathbf{g}), \log(\mathbf{g}_j)) \leq \delta$. Hence $(\log(\mathbf{g}_j))_{j=1}^{\mathcal{N}_n}$ is a δ-cover for $\log(\mathcal{G})$ with

respect to $d_{\log(\tau)}$. Since the $\mathbf{g}_j$ are in $\mathcal{G}$, the $\log(\mathbf{g}_j)$ are in $\log(\mathcal{G})$, thus this cover is an internal cover. Since $\mathcal{N}(\delta, \log(\mathcal{G}), d_{\log(\tau)}(\cdot, \cdot))$ is the minimal number of balls with center in $\log(\mathcal{G})$ required to cover $\log(\mathcal{G})$. This proves the assertion. $\qquad\square$

Proof of Lemma 2.3.7. Let $\mathbf{p}, \mathbf{q} \in \mathcal{S}^k$. Thus, $\sum_{k=1}^{K} p_k = 1$, $\sum_{k=1}^{K} q_k = 1$ and

$$\sum_{k=1}^{K} p_k \left(B \wedge \log \left(\frac{p_k}{q_k} \right) \right) = \sum_{k=1}^{K} \left(p_k \left(B \wedge \log \left(\frac{p_k}{q_k} \right) \right) - p_k + q_k \right). \qquad (2.C.1)$$

Suppose for the moment that for any $k = 1, \cdots, K$,

$$p_k \left(B \wedge \log \left(\frac{p_k}{q_k} \right) \right) - p_k + q_k \geq \frac{1}{C_{m,B}} p_k \left| B \wedge \log \left(\frac{p_k}{q_k} \right) \right|^m, \qquad (2.C.2)$$

with $C_{m,B} := \max\{m!, B^m/(B-1)\}$. Applying this inequality to each term on the right hand side of (2.C.1) gives

$$\sum_{k=1}^{K} p_k \left(B \wedge \log \left(\frac{p_k}{q_k} \right) \right) \geq \sum_{k=1}^{K} \frac{1}{C_{m,B}} p_k \left| B \wedge \log \left(\frac{p_k}{q_k} \right) \right|^m.$$

Since $C_{m,B} > 0$, multiplying both sides of the inequality with $C_{m,B}$ yields the claim.

It remains to proof (2.C.2). First we consider the case that $p_k = 0$. By considering the limit we get that $0 \log^m(0) = 0$, for $m = 1, 2, \cdots$. Thus the right hand side of (2.C.2) is equal to 0, while the left hand side is equal to q_k. Since $q_k \geq 0$, this proves (2.C.2) for this case.

Assume now that $p_k > 0$. Dividing both sides by p_k yields

$$B \wedge \log \left(\frac{p_k}{q_k} \right) - 1 + \frac{q_k}{p_k} \geq \frac{1}{C_{m,B}} \left| B \wedge \log \left(\frac{p_k}{q_k} \right) \right|^m.$$

If $p_k/q_k \geq e^B$ the inequality follows immediately. It remains to study the case that $p_k/q_k < e^B$. In this case one can always replace $B \wedge \log(p_k/q_k)$ by $\log(p_k/q_k)$. Introducing the new variable $u = q_k/p_k$ and replacing $C_{m,B}$ by $C > 0$ gives rise to a function

$$H_{C,m}(u) = u - 1 - \log(u) - |\log(u)|^m/C.$$

It remains to show that $H_{C_{m,B},m}(u) \geq 0$ for all $u \geq e^{-B}$. Obviously, $H_{C,m}(1) = 0$ for all C, so we only have to consider $u \neq 1$. Consider first $u > 1$ and $C = m!$. Using the substitution $u = e^s$ gives

$$m! e^s - m!(s+1) - s^m.$$

Substituting the power series for the exponential function leads to

$$m! \sum_{n=0}^{\infty} \frac{s^n}{n!} - m!(1+s) - s^m = m! \sum_{n=2}^{m-1} \frac{s^n}{n!} + m! \sum_{n=m+1}^{\infty} \frac{s^n}{n!} > 0,$$

where the last strict inequality holds because $u > 1$ and thus $s > 0$. Thus $H_{m!,m}(u) \geq 0$ for $u > 1$.

For $u \in (e^{-b}, 1)$, dividing by $u - \log(u) - 1$ gives us the following constraint on the constant C :

$$C \geq \sup_{u \in (e^{-B}, 1)} \frac{|\log(u)|^m}{u - \log(u) - 1}. \tag{2.C.3}$$

This division can be done since $u - \log(u) - 1 > 0$ when $u > 0$, $u \neq 1$ and zero if and only if $u = 1$, which for example can be shown by observing the sign of the derivative.

Define $C_{<1}$ as $C_{<1} := B^m/(B - 1)$. Since $|\log(u)|^m/(u - \log(u) - 1)$ is strictly decreasing on $(0, 1)$, see Proposition 2.C.5 (II), it follows for $u \in [e^{-B}, 1)$ that $|\log(u)|^m/(u - \log(u) - 1) \leq B^m/(e^{-B} + B - 1)$. Now since $B > 1$, it follows that $B^m/(u + B - 1)$ is also strictly decreasing on $[0, 1]$. Hence on $[0, e^{-B}]$ we have $B^m/(e^{-B} + B - 1) \leq B^m/(u + B - 1) \leq C_{<1}$, thus $C_{<1}$ satisfies (2.C.3).

Now notice that $C_{m,B} = \max\{C_{<1}, m!\}$. Consequently $H_{C_{m,B},m}(u) \geq 0$, for all $u \geq e^{-B}$, proving (2.C.2). □

For all $m = 2, 3, \ldots$ define the function $F_m : (0, \infty) \to [0, \infty)$ as

$$F_m(u) := \frac{|\log^m(u)|}{u - \log(u) - 1}.$$

Since $u - \log(u) - 1 \geq 0$, this function indeed takes only positive values. Furthermore since $u - \log(u) - 1 = 0$ only when $u = 1$ this is the only possible singularity/discontinuity of this function. The next result derives some properties of the function $F_m(u)$.

Proposition 2.C.5. *If $m = 2, 3, \cdots$, then*
  (i) $\lim_{u \to 1} F_2(u) = 2$ and $\lim_{u \to 1} F_m(u) = 0$ for $m > 2$
  (ii) $F_m(u)$ is strictly decreasing on $(0, 1)$.

Proof. (i): For $u = 1$, it holds that $(u - \log(u) - 1) = 0$ and $|\log^m(u)| = 0$. Applying L'Hopital's rule twice yields the desired result.

(ii): The L'Hopital's like rule for monotonicity, see [112] or Lemma 2.2 in [4], states that a function f/g on an interval (a, b), satisfying $g' \neq 0$ and either $f(a) = 0 = g(a)$ or $f(b) = 0 = g(b)$, is strictly increasing/decreasing if f'/g' is strictly

increasing/decreasing on (a, b). For $f(u) = |\log^m(u)|$ and $g(u) = u - \log(u) - 1$, we have

$$\frac{f'(u)}{g'(u)} = \frac{m \log(u)|\log^{m-2}(u)|}{u - 1}$$

and for $\bar{f}(u) = m \log(u)|\log^{m-2}(u)|$ and $\bar{g}(u) = u - 1$, we obtain

$$\frac{\bar{f}'(u)}{\bar{g}'(u)} = \frac{(m-1)m|\log^{m-2}(u)|}{u}.$$

On $u \in (0, 1)$, $\bar{f}'(u)/\bar{g}'(u)$ is strictly decreasing. Applying the L'Hopital's like rule for monotonicity twice yields the statement. $\qquad\square$

Proof of Lemma 2.3.4. The inequality $\mathrm{KL}_2(P, Q) \leq \mathrm{KL}_B(P, Q)$ follows direct from the definition of the truncated Kullback-Leibler divergence. Write $P = P^a + P^s$ for the Lebesgue decomposition of P with respect to Q such that $P^a \ll Q$. The Lebesgue decomposition ensures existence of a set A with $P^a(A) = 0 = P^s(A^c)$. For $x \in A$, we define $dP/dQ(x) := +\infty$. For the dominating measure $\mu = (P + Q)/2$, denote by p, p^a, p^s, q the μ-densities of P, P^a, P^s, Q, respectively. Since $p^s q = 0$,

$$H^2(P, Q) = \int \left(p^a + p^s - \sqrt{p^a q} \right)$$

$$\leq \int_{0 < p^a/q \leq e^2} \left(p^a - \sqrt{p^a q} \right) + \int_{p^a/q > e^2} p^a + \int p^s.$$

For every $u \in \mathbb{R}$, we have $1 - u \leq e^{-u}$ and hence $e^u - 1 \leq u e^u$. Substituting $u = \log(\sqrt{y})$ yields $\sqrt{y} - 1 \leq \sqrt{y} \log(\sqrt{y})$ and therefore $y - \sqrt{y} \leq y \log(\sqrt{y}) = y \log(y)/2$. With $y = p^a/q$, we find,

$$H^2(P, Q) \leq \int_{0 < p^a/q \leq e^2} \frac{p^a}{2q} \log\left(\frac{p^a}{q}\right) q + \int_{p^a/q > e^2} p^a + \int p^s.$$

The other direction works similarly. Second order Taylor expansion around one gives for $y > 0$, $y \log(y) \leq y - 1 + \frac{1}{2}(y - 1)^2/(y \wedge 1)$. For $y = \sqrt{x}$, we find $x \log(x) = 2\sqrt{x} \cdot \sqrt{x} \log(\sqrt{x}) \leq 2(x - \sqrt{x}) + (1 \vee \sqrt{x})(\sqrt{x} - 1)^2$. Consequently, for each $B \geq 0$,

$$\mathrm{KL}_B(P, Q) = \int_{p^a/q \leq e^B} \frac{p^a}{q} \log\left(\frac{p^a}{q}\right) q + B \int_{dP/dQ > e^B} dP$$

$$\leq 2 e^{B/2} H^2(P, Q) + 2 \int_{p^a/q \leq e^B} p - \sqrt{pq} + B \int_{dP/dQ > e^B} dP.$$

If $\int_{p^a/q \leq e^B} p^a - \sqrt{p^a q} \leq 0$, we can use that $H^2(P,Q) \geq \frac{1}{2} \int_{p/q \geq e^B} (\sqrt{p} - \sqrt{q})^2 \geq \frac{1}{2} \int_{p/q \geq e^B} p(1 - e^{-B/2})^2$ and hence

$$\mathrm{KL}_B(P,Q) \leq 2\Big(e^{B/2} + (1 - e^{-B/2})^{-2}\Big) H^2(P,Q).$$

Otherwise, if $\int_{p^a/q \leq e^B} p^a - \sqrt{p^a q} > 0$, we can upper bound

$$\mathrm{KL}_B(P,Q) \leq 2e^{B/2} H^2(P,Q) + B(1 - e^{-B/2})^{-1} \int_{p^a/q \leq e^B} p - \sqrt{pq}$$

$$+ B \int_{dP/dQ > e^B} dP$$

$$\leq 2e^{B/2} H^2(P,Q) + B(1 - e^{-B/2})^{-1} \int p - \sqrt{pq}$$

$$= \Big(2e^{B/2} + B(1 - e^{-B/2})^{-1}\Big) H^2(P,Q).$$

The result now follows by observing that since $B \geq 2$, both $B(1 - e^{-B/2})^{-1}$ and $2(1 - e^{-B/2})^{-2}$ are less than $2e^{B/2}$. $\qquad\square$

Proposition 2.C.6. *Recall that Φ denotes the softmax function. The function $\log(\Phi(\cdot)) : \mathbb{R}^K \to \mathbb{R}^K$ satisfies $|\log(\Phi(\mathbf{x})) - \log(\Phi(\mathbf{y}))|_\infty \leq K\|\mathbf{x} - \mathbf{y}\|_\infty$.*

Proof. Consider the composition of the logarithm with the softmax function, that is,

$$\left(\log\left(\frac{e^{x_1}}{\sum_{j=1}^K e^{x_j}} \right), \cdots, \log\left(\frac{e^{x_K}}{\sum_{j=1}^K e^{x_j}} \right) \right).$$

It holds for $k, i \in \{1, \cdots, K\}$, $i \neq k$ that

$$\frac{\partial}{\partial x_k} \log\left(\frac{e^{x_k}}{\sum_{j=1}^K e^{x_j}} \right) = 1 - \frac{e^{x_k}}{\sum_{j=1}^K e^{x_j}},$$

$$\frac{\partial}{\partial x_k} \log\left(\frac{e^{x_i}}{\sum_{j=1}^K e^{x_j}} \right) = -\frac{e^{x_k}}{\sum_{j=1}^K e^{x_j}}.$$

The partial derivatives are bounded in absolute value by one. The combined log-softmax function is therefore Lipschitz continuous (w.r.t to $\|\cdot\|_\infty$ norm for vectors) with Lipschitz constant bounded by K. $\qquad\square$

Proof of Lemma 2.3.8. We start proving the first bound. Notice that $g \in \log(\mathcal{F}_{\mathbf{\Phi}}(L, \mathbf{m}, s))$ means that there exists a ReLU network $f_g \in \mathcal{F}_{\mathrm{id}}(L, \mathbf{m}, s)$ such that $g(\mathbf{x}) = \log(\mathbf{\Phi}(f_g(\mathbf{x})))$. By Lemma 5 of [127] it holds that $\mathcal{N}(\delta/(2K), \mathcal{F}_{\mathrm{id}}(L, \mathbf{m}, s), \|\cdot\|_\infty) \leq (4\delta^{-1}K(L+1)V^2)$. Let $\delta > 0$. Denote by $(\mathbf{f}_j)_{j=1}^{\mathcal{N}_n}$ the centers of a minimal $\delta/(2K)$-covering of $\mathcal{F}_{\mathrm{id}}(L, \mathbf{m}, s)$ with respect to $\|\cdot\|_\infty$. Triangle inequality gives that for each $\mathbf{f}_j$ there exists a $\widehat{\mathbf{f}}_j \in \mathcal{F}_{\mathrm{id}}(L, \mathbf{m}, s)$ such that $(\widehat{\mathbf{f}}_j)_{j=1}^{\mathcal{N}_n}$ is an interior δ/K-cover of $\mathcal{F}_{\mathrm{id}}(L, \mathbf{m}, s)$. Let $\mathbf{g} \in \log(\mathcal{F}_{\mathbf{\Phi}}(L, \mathbf{m}, s))$. By the cover property, there exists a $j \in \{1, \cdots, \mathcal{N}_n\}$ such that $\|f_g - \widehat{f}_j\| \leq \delta/K$. Proposition 2.C.6 yields:

$$\|\mathbf{g} - \log(\Phi(\widehat{\mathbf{f}}_j))\|_\infty = \|\log(\Phi(\mathbf{f}_g)) - \log(\Phi(\widehat{\mathbf{f}}_j))\|_\infty \leq K\|f_g - \widehat{f}_j\|_\infty \leq \delta.$$

Since $\mathbf{g} \in \log(\mathcal{F}_{\mathbf{\Phi}}(L, \mathbf{m}, s))$ was arbitrary and $\widehat{f}_j \in \mathcal{F}_{\mathrm{id}}(L, \mathbf{m}, s)$ for $j = 1, \cdots, \mathcal{N}_n$, this means that $(\log(\Phi(\widehat{\mathbf{f}}_j))$ is an internal δ-cover for $\log(\mathcal{F}_{\mathbf{\Phi}}(L, \mathbf{m}, s))$ with respect to $\|\cdot\|_\infty$. Hence

$$\mathcal{N}(\delta, \log(\mathcal{F}_{\mathbf{\Phi}}(L, \mathbf{m}, s)), \|\cdot\|_\infty) \leq \mathcal{N}(\delta/(2K), \mathcal{F}_{\mathrm{id}}(L, \mathbf{m}, s), \|\cdot\|_\infty)$$
$$\leq (4\delta^{-1}K(L+1)V^2).$$

Now we consider the second bound of the lemma. Using that $m_0 = d$, $m_{L+1} = K$ and by removing inactive nodes, Proposition 2.A.1, we get that $m_\ell \leq s$ for $s = 1, \cdots, L$, and thus

$$V \leq dKs^L 2^{L+2}.$$

Substituting this in the first bound and taking the logarithm yields the result. $\square$

Proposition 2.C.7. *Consider binary classification ($K = 2$) for the conditional class probabilities $p_1(\mathbf{x}) = (3|x_1 + x_2 - 1|^8)/4$ and $p_2(\mathbf{x}) = 1 - p_1(\mathbf{x})$. If $\mathbf{X}$ is uniformly distributed on $[0, 1]^2$, then*

$$\mathbb{P}_{\mathbf{X}}\left(p_1(\mathbf{X}) \leq t\right) = 2\left(\frac{4t}{3}\right)^{\frac{1}{8}} - \left(\frac{4t}{3}\right)^{\frac{1}{4}}.$$

If the distribution of $\mathbf{X}$ is given by the density $(x_1, x_2) \mapsto 3|x_1 + x_2 - 1|$, then

$$\mathbb{P}_{\mathbf{X}}\left(p_1(\mathbf{X}) \leq t\right) = 3\left(\frac{4t}{3}\right)^{\frac{1}{4}} - 2\left(\frac{4t}{3}\right)^{\frac{3}{8}}.$$

Proof. By rewriting the inequality $p_1(\mathbf{X}) \leq t$, we get for both cases that

$$\mathbb{P}_{\mathbf{X}}\left(p_1(\mathbf{X}) \leq t\right) = \mathbb{P}_{\mathbf{X}}\left((3|x_1 + x_2 - 1|^8)/4 \leq t\right)$$

$$= \mathbb{P}_{\mathbf{X}}\left(1 - \left(\frac{4t}{3}\right)^{\frac{1}{8}} \le x_1 + x_2 \le 1 + \left(\frac{4t}{3}\right)^{\frac{1}{8}}\right).$$

First we consider the case of uniform design. In this case, we find

$$\mathbb{P}_{\mathbf{X}}\left(1 - \left(\frac{4t}{3}\right)^{\frac{1}{8}} \le x_1 + x_2 \le 1 + \left(\frac{4t}{3}\right)^{\frac{1}{8}}\right)$$

$$= \int_0^1 \int_{1-\left(\frac{4t}{3}\right)^{\frac{1}{8}}-x_2}^{1+\left(\frac{4t}{3}\right)^{\frac{1}{8}}-x_2} 1 dx_1 dx_2 - \int_0^{\left(\frac{4t}{3}\right)^{\frac{1}{8}}} \int_1^{1+\left(\frac{4t}{3}\right)^{\frac{1}{8}}-x_2} 1 dx_1 dx_2$$

$$- \int_{1-\left(\frac{4t}{3}\right)^{\frac{1}{8}}}^1 \int_{1-\left(\frac{4t}{3}\right)^{\frac{1}{8}}-x_2}^0 1 dx_1 dx_2$$

$$= 2\left(\frac{4t}{3}\right)^{\frac{1}{8}} - \frac{1}{2}\left(\frac{4t}{3}\right)^{\frac{1}{4}} - \frac{1}{2}\left(\frac{4t}{3}\right)^{\frac{1}{4}}.$$

Here, the second and third double integral are correction terms that compensate for the regions where the first double integral integrates over values outside $[0,1]^2$.

To prove the second part of the statement, consider the case that the distribution of $\mathbf{X}$ is given by the density $(x_1, x_2) \mapsto 3|x_1 + x_2 - 1|$. In this case we have that

$$\mathbb{P}_{\mathbf{X}}\left(1 - \left(\frac{4t}{3}\right)^{\frac{1}{8}} \le x_1 + x_2 \le 1 + \left(\frac{4t}{3}\right)^{\frac{1}{8}}\right)$$

$$= \int_0^1 \int_{1-\left(\frac{4t}{3}\right)^{\frac{1}{8}}-x_2}^{1+\left(\frac{4t}{3}\right)^{\frac{1}{8}}-x_2} 3|x_1 + x_2 - 1| dx_1 dx_2$$

$$- \int_0^{\left(\frac{4t}{3}\right)^{\frac{1}{8}}} \int_1^{1+\left(\frac{4t}{3}\right)^{\frac{1}{8}}-x_2} 3|x_1 + x_2 - 1| dx_1 dx_2$$

$$- \int_{1-\left(\frac{4t}{3}\right)^{\frac{1}{8}}}^1 \int_{1-\left(\frac{4t}{3}\right)^{\frac{1}{8}}-x_2}^0 3|x_1 + x_2 - 1| dx_1 dx_2$$

$$= \int_0^1 \int_{1-\left(\frac{4t}{3}\right)^{\frac{1}{8}}-x_2}^{1-x_2} 3(-x_1 - x_2 + 1) dx_1 dx_2$$

$$+ \int_0^1 \int_{1-x_2}^{1+\left(\frac{4t}{3}\right)^{\frac{1}{8}}-x_2} 3(x_1 + x_2 - 1) dx_1 dx_2$$

$$- \int_0^{\left(\frac{4t}{3}\right)^{\frac{1}{8}}} \int_1^{1+\left(\frac{4t}{3}\right)^{\frac{1}{8}}-x_2} 3(x_1 + x_2 - 1) dx_1 dx_2$$

$$- \int_{1-\left(\frac{4t}{3}\right)^{\frac{1}{8}}}^{1} \int_{1-\left(\frac{4t}{3}\right)^{\frac{1}{8}}-x_2}^{0} 3(-x_1 - x_2 + 1)dx_1 dx_2$$

$$= \frac{3}{2}\left(\frac{4t}{3}\right)^{\frac{1}{4}} + \frac{3}{2}\left(\frac{4t}{3}\right)^{\frac{1}{4}} - \left(\frac{4t}{3}\right)^{\frac{3}{8}} - \left(\frac{4t}{3}\right)^{\frac{3}{8}}.$$

Again, the correction terms occur because we integrate over values outside $[0,1]^2$.   □

Acknowledgments