

# Risk bounds for aggregated shallow neural networks using Gaussian priors

**Laura Tinsi**

*CREST, ENSAE Paris, Institut Polytechnique de Paris  
EDFLab, Electricité de France R&D*

TINSILAURA@GMAIL.COM

**Arnak Dalalyan**

*CREST, ENSAE Paris, Institut Polytechnique de Paris*

ARNAK.DALALYAN@ENSAE.FR

**Editors:** Po-Ling Loh and Maxim Raginsky

## Abstract

Analysing statistical properties of neural networks is a central topic in statistics and machine learning. However, most results in the literature focus on the properties of the neural network minimizing the training error. The goal of this paper is to consider aggregated neural networks using a Gaussian prior. The departure point of our approach is an arbitrary aggregate satisfying the PAC-Bayesian inequality. The main contribution is a precise nonasymptotic assessment of the estimation error appearing in the PAC-Bayes bound. We also review available bounds on the error of approximating a function by a neural network. Combining bounds on estimation and approximation errors, we establish risk bounds which are sharp enough to lead to minimax rates of estimation over Sobolev smoothness classes.

**Keywords:** shallow neural networks, PAC-Bayes, mirror averaging, exponential weights

## 1. Introduction

Neural networks are the most widely used parameterised functions for solving machine learning tasks. Given a neural network architecture, the parameters of the neural network are then learned from data. Assessing the error of the learned network on new, unobserved examples is a central topic in statistics and learning theory (Bartlett et al., 2021; Fan et al., 2021). The most popular approach for estimating the parameters of the network from data, referred to as weights and biases, is the minimization of the (regularized) training error. This is usually done by the stochastic gradient descent algorithm, or a version of it. For binary classification, risk bounds for neural network classifiers are based on Vapnik-Chervonenkis dimension (Bartlett et al., 1998; Anthony and Bartlett, 1999; Bartlett et al., 2019). Even for simple networks containing only one hidden layer, these risk bounds are rather involved (Xie et al., 2017; Zhong et al., 2017; Cao and Gu, 2019; Ba et al., 2020).

A well-known alternative to minimizing the regularized training error is to use a prediction rule based on a posterior distribution. Typical example is the network obtained by sampling its weights from the posterior, or the convex combination of the networks averaged using the posterior distribution. Surprisingly, little is known about risk bounds of posterior-based prediction rules in the context of neural networks. The goal of the present work is to do the first step in filling this gap by focusing on one-hidden-layer feedforward neural networks and Gibbs posteriors using Gaussian priors. An attractive feature of posterior-based methods is that their analysis can be carried out using the PAC-Bayes theory McAllester (1999, 2003) as a substitute to the Vapnik-Chervonenkis dimension or the Rademacher complexity. We refer the reader to Catoni (2007); Guedj (2019);

Alquier (2021); Germain et al. (2009) for an enlightening account of the PAC-Bayesian approach in statistics and learning.

PAC-Bayes theory with data-dependent prior Rivasplata et al. (2018); Lever et al. (2013) has been already used in the framework of neural networks, mainly for providing data-driven bounds on the generalisation error of trained (stochastic) networks and prior selection based on these bounds Dziugaite and Roy (2017); Neyshabur et al. (2017); Zhou et al. (2019); Letarte et al. (2019); Biggs and Guedj (2021); Perez-Ortiz et al. (2021a,b); Pérez-Ortiz et al. (2021). A derandomized approach of such data-driven bounds has also been investigated, including for (deterministic) shallow neural networks, in the setting of classification; by means of margins and sub-gaussian concentration in Biggs and Guedj (2022a), and with a majority vote approach in Biggs and Guedj (2022b). In this paper, we take a different route and propose to use in-expectation PAC-Bayes bounds — Leung and Barron (2006); Juditsky et al. (2008); Dalalyan and Tsybakov (2012a); Rigollet and Tsybakov (2012), for investigating the risk (or, the expected excess loss) of aggregated neural networks. To be more specific, let  $\mathcal{F}_W := \{f_w, w \in W\}$  be a parametric class of prediction rules, with a parameter  $w$  lying in a measurable space  $(W, \mathcal{W})$ . One can think of  $\mathcal{F}_W$  as a set of neural networks with a given architecture and of  $w$  as the vector of the weights and biases. Assume we are given a data sample of size  $n$  independently drawn from an unknown distribution  $\mathbf{P}$  over the data space, and we wish to “aggregate” elements of  $\mathcal{F}_W$  to obtain a prediction rule  $\hat{f}_n$  that mimics the Bayes predictor  $f_{\mathbf{P}}$ . This means that for a prescribed loss function  $\ell(\cdot, \cdot)$  taking real values, we wish  $\ell(\hat{f}_n, f_{\mathbf{P}})$  to be small. It turns out that under some general assumptions, for a given prior distribution  $\pi$  on  $W$  and a temperature parameter  $\beta > 0$ , there exists an aggregate  $\hat{f}_n$  such that

$$\mathbf{E}[\ell(\hat{f}_n, f_{\mathbf{P}})] \leq C_{\mathbf{PB}} \inf_p \left\{ \int_W \ell(f_w, f_{\mathbf{P}}) p(dw) + \frac{\beta}{n} D_{\text{KL}}(p||\pi) \right\}, \quad (1)$$

where  $C_{\mathbf{PB}}$  is some constant and the infimum is over all probability distributions  $p$  over  $W$ . We say then that  $\hat{f}_n$  satisfies a PAC-Bayes inequality in-expectation. The general specification of (1) applies, under some assumptions, to a wide range of statistical problems such as regression, classification or density estimation. For regression with fixed design, the Gibbs-posterior mean was shown to satisfy (1) with  $C_{\mathbf{PB}} = 1$  in Leung and Barron (2006) for Gaussian noise, and in (Dalalyan and Tsybakov, 2007) for more general noise distributions. In some other problems, including the random design regression and the density estimation, similar bounds were established for the mirror averaging (Yuditskii et al., 2005; Juditsky et al., 2008). PAC-Bayes bounds with  $C_{\mathbf{PB}} > 1$  for the prediction rule obtained by randomly drawing  $w$  from the Gibbs posterior were proved in (Catoni, 2007; Alquier and Biau, 2013). The recent papers (Biggs and Guedj, 2021; Fortier-Dubois et al., 2021) studied the problem of aggregation of neural networks with sign activation.

In the present work, we elaborate on (1) to get a tractable risk bound when  $\mathcal{F}_W$  is the set of neural networks with a single hidden layer. The tractability here should be understood as the property of showing clearly the dependence on the important problem characteristics (sample-size, input and output dimensions) and those of the learning algorithm (variances of the prior distribution, number of hidden layers, properties of the activation functions). Our first main contribution stated in Theorem 2 is a tractable risk bound formulated as an oracle inequality. To our knowledge, this inequality is sharper and easier to deal with than its counterparts for the training error minimizing shallow networks. To show potential implications of this oracle inequality, we combine it with known approximation bounds when the Bayes predictor lies in a Sobolev ball. Interestingly, we show that a proper choice of the width of the hidden layer and the variances of the prior leads to minimax

optimal rates of convergence, up to logarithmic factors. More specifically, for the Sobolev ball of smoothness  $r$  and input dimension  $D_0$ , we obtain in Theorem 6 the rate  $n^{-2r/(2r+D_0)} \log^2 n$  for a specific class of sigmoid activation functions. A similar result is obtained for the ReLU activation as well, but with a slightly slower rate  $n^{-2\bar{r}/(2\bar{r}+D_0+1)}$  (up to a polylogarithmic factor) for any  $\bar{r} < r$ .

The rest of the paper is organized as follows. In Section 2, we define the generic PAC-Bayesian framework and instantiate it in the setting of shallow neural networks. In Section 3, we state the main oracle bound for shallow neural networks with a Gaussian prior. Section 4 provides examples of statistical problems where PAC-Bayesian bounds of type (1) are available. Section 5 is devoted to a selective review of the literature on approximation properties of neural networks with bounded (sigmoid) and unbounded (ReLU) activation functions. Finally, Section 6 contains the upper bounds on the worst-case risk which are nearly minimax rate-optimal in the case of sigmoid activation. Some concluding remarks are provided in Section 7. Technical proofs are deferred to the appendices.

## 2. Preliminaries and notation

In this section, we set the general framework of the PAC-Bayesian bound that will be the starting point of our work. We then instantiate it in the specific case of neural networks.

### 2.1. General framework and PAC-Bayesian type bounds

Let  $(\mathcal{Z}, \mathcal{A})$  be a measurable space. We observe one realisation of the random vector  $\mathbf{Z}^n = (Z_1, \dots, Z_n) \in \mathcal{Z}^n$  drawn from an unknown distribution  $\mathbf{P}$  on  $(\mathcal{Z}^n, \mathcal{A}^{\otimes n})$ . We denote by  $\|\mathbf{x}\|_2$  the Euclidean norm of the vector  $\mathbf{x}$  of an Euclidean space. Let  $\mathcal{X} \subset \mathbb{R}^{D_0}$ ,  $D_0 \geq 1$ , be a Borel set and let  $\mu$  be a  $\sigma$ -finite measure on  $(\mathcal{X}, \mathcal{B}(\mathcal{X}))$  such that  $M_2^2 = D_0^{-1} \int_{\mathcal{X}} \|\mathbf{x}\|_2^2 \mu(d\mathbf{x}) < +\infty$ . In the sequel, we denote by  $\mathbb{L}_q(\mu)$ ,  $q \in [1, \infty)$ , the set of all the functions  $f : \mathcal{X} \rightarrow \mathbb{R}^{D_2}$  such that  $\int_{\mathcal{X}} \|f(\mathbf{x})\|_2^q \mu(d\mathbf{x}) < \infty$ . Let  $\mathcal{P}_W$  be the space of all probability measures on  $W$  and let

$$\mathcal{P}_1(\mathcal{F}_W) = \left\{ p \in \mathcal{P}_W : \int_W \|f_w(\mathbf{x})\|_2 p(dw) < \infty, \quad \text{for all } \mathbf{x} \in \mathcal{X} \right\}.$$

We consider the problem of estimating a function  $f_{\mathbf{P}} \in \mathbb{L}_2(\mu)$ . At this stage, one may think of  $f_{\mathbf{P}}$  as the multidimensional regression function when  $\mathcal{Z} = \mathcal{X} \times \mathbb{R}^{D_2}$ , the Bayes classifier when  $\mathcal{Z} = \mathcal{X} \times \{-1, 1\}$  or the density of observations when  $\mathcal{Z} = \mathcal{X}$  (in the last two cases  $D_2 = 1$ ). A common approach in statistics and statistical learning is to use a parametric set  $\mathcal{F}_W := \{f_w, w \in W\} \subset \mathbb{L}_2(\mu)$ , indexed by a measurable set  $W \subset \mathbb{R}^d$ , for some  $d \in \mathbb{N}$ , for constructing an estimator of  $f_{\mathbf{P}}$ . Instances of this approach are the empirical risk minimizer, the Bayesian posterior mean, the exponentially weighted aggregate, etc. The quality of an estimator  $\hat{f}_n$  of  $f_{\mathbf{P}}$  is measured by means of a loss function  $\ell : \mathbb{L}_2(\mu) \times \mathbb{L}_2(\mu) \mapsto \mathbb{R}_+$ ; an estimator  $\hat{f}_n$  is good if its risk

$$\mathbf{E}_{\mathbf{P}} [\ell(\hat{f}_n(\mathbf{Z}^n), f_{\mathbf{P}})] = \int_{\mathcal{Z}^n} \ell(\hat{f}_n(\mathbf{z}), f_{\mathbf{P}}) \mathbf{P}(d\mathbf{z})$$

is small. A widespread choice of the loss function, used throughout this paper except in Section 4, is the squared  $\ell_2$ -norm  $\ell(g, h) = \|g - h\|_{\mathbb{L}_2(\mu)}^2 = \int_{\mathcal{X}} \|g(\mathbf{x}) - h(\mathbf{x})\|_2^2 \mu(d\mathbf{x})$ ,  $\forall g, h \in \mathbb{L}_2(\mu)$ .

We say that the estimator  $\hat{f}_n$  satisfies the PAC-Bayesian bound with prior  $\pi \in \mathcal{P}_1(\mathcal{F}_W)$  and temperature parameter  $\beta > 0$ , if (1) is satisfied (where the infimum in the right hand side is over all

$p$  in  $\mathcal{P}_1(\mathcal{F}_W)$ ). If  $C_{\text{PB}} = 1$ , the bound is called exact or sharp. When the loss function is the squared  $\mathbb{L}_2$ -norm, the PAC-Bayesian bound reads as

$$\mathbf{E}_{\mathbf{P}}[\|\hat{f}_n - f_{\mathbf{P}}\|_{\mathbb{L}_2}^2] \leq C_{\text{PB}} \inf_{p \in \mathcal{P}_1(\mathcal{F}_W)} \left\{ \int_W \|f_{\mathbf{w}} - f_{\mathbf{P}}\|_{\mathbb{L}_2(\mu)}^2 p(d\mathbf{w}) + \frac{\beta}{n} D_{\text{KL}}(p||\pi) \right\}. \quad (2)$$

## 2.2. Shallow neural networks

In the rest of this section, we provide more details on the notations and assumptions that will stand when we estimate  $f_{\mathbf{P}}$  by aggregation of neural networks. We consider the class of networks with a single hidden layer and denote by  $D_1$  the number of units in this layer.

In order to merge weights and biases of a neural network, we note  $\mathbf{x} = (1, x_1, \dots, x_{D_0-1})^\top \in \mathcal{X}$ . The set  $W$  of the weights of a neural network can be divided into the weights of the hidden layer,  $\mathbf{w}_1$ , and the weights of the output layer,  $\mathbf{w}_2$  so that  $\mathbf{w}_1 \in \mathbb{R}^{D_0 \times D_1}$  and  $\mathbf{w}_2 \in \mathbb{R}^{D_1 \times D_2}$ . Therefore,  $\mathbf{w} = (\mathbf{w}_1, \mathbf{w}_2)^\top$  can be seen as an element of  $\mathbb{R}^d$  with the overall dimension  $d = D_0 D_1 + D_1 D_2$ . The neural network parametrized by  $\mathbf{w}$  has the form:

$$f_{\mathbf{w}}(\mathbf{x}) = \mathbf{w}_2^\top \bar{\sigma}(\mathbf{w}_1^\top \mathbf{x}) \in \mathbb{R}^{D_2}, \quad \forall \mathbf{x} \in \mathbb{R}^{D_0} \quad \text{with} \quad \bar{\sigma} : \mathbf{x} \in \mathbb{R}^{D_1} \mapsto \begin{bmatrix} \sigma(x_1) \\ \vdots \\ \sigma(x_{D_1}) \end{bmatrix} \in \mathbb{R}^{D_1}, \quad (3)$$

where  $\sigma : \mathbb{R} \rightarrow \mathbb{R}$  is a scalar activation function. In the next sections, we will consider both the case of bounded and unbounded activation functions in order to cover most of the usual ones. We refer to the bounded case by means of the following assumption.

**Assumption ( $\sigma$ -B)** *The function  $\sigma$  is bounded by  $M_\sigma$ , i.e.,  $|\sigma(u)| \leq M_\sigma$  for all  $u \in \mathbb{R}$ .*

Let us stress that only some of our results require Assumption ( $\sigma$ -B). However, all our results will require the Lipschitz assumption stated below, which is satisfied by sigmoid functions as well as piecewise continuous functions (including ReLU). Without loss of generality, we will assume that the Lipschitz constant is equal to one.

**Assumption ( $\sigma$ -L)** *For every pair of real numbers  $(u, u')$ , we have  $|\sigma(u) - \sigma(u')| \leq |u - u'|$ .*

## 2.3. Spherical Gaussian prior distribution

The prior distribution  $\pi$  defined in the PAC-Bayesian framework can be interpreted as the initial distribution of the weights, or as a regulariser. Most of the time, in practice, the weights are initialized randomly from a Gaussian distribution. Building upon this remark, in the remaining part of the paper, we consider the prior to be a product of spherical Gaussians. Recall that the weights of a neural network are split into two groups: the weights  $\mathbf{w}_1$  of the hidden layer and the weights  $\mathbf{w}_2$  of the output layer. To take into account their different roles we assume the distribution over  $\mathbf{w}$  is a product of two spherical Gaussians with different variances.

**Assumption ( $\mathcal{N}$ )** *The prior  $\pi$  satisfies  $\pi = \pi_1 \otimes \pi_2 = \mathcal{N}(0, \rho_1^2 \mathbf{I}_{D_0 D_1}) \otimes \mathcal{N}(0, \rho_2^2 \mathbf{I}_{D_1 D_2})$ , where  $\mathbf{I}_{D_0 D_1}$  and  $\mathbf{I}_{D_1 D_2}$  are respectively the identity matrices of size  $D_0 D_1 \times D_0 D_1$  and  $D_1 D_2 \times D_1 D_2$ .*

We refer to  $\pi_1$  and  $\pi_2$  as the distribution of the hidden layer and the output layer respectively.

### 3. Oracle inequalities for networks with one hidden layer and Gaussian prior

In this section, we first derive a bound for the risk of the estimator  $\hat{f}_n$  when the prior has an arbitrary centered Gaussian distribution, and subsequently provide an oracle inequality for a carefully chosen Gaussian prior. Let  $\bar{\mathbf{w}} \in \mathcal{W}$  be any value of the parameter. Using the triangle inequality, in conjunction with the fact that  $\sqrt{(a+b)^2 + c^2} \leq a + \sqrt{b^2 + c^2}$  for  $a, b, c \geq 0$ , one can infer from (2) that

$$\left( C_{\text{PB}}^{-1} \mathbf{E}_{\mathbf{P}} [\|\hat{f}_n - f_{\mathbf{P}}\|_{\mathbb{L}_2(\mu)}^2] \right)^{1/2} \leq \|f_{\bar{\mathbf{w}}} - f_{\mathbf{P}}\|_{\mathbb{L}_2(\mu)} + \text{Rem}_n(\bar{\mathbf{w}})^{1/2}, \quad (4)$$

with the remainder term given by

$$\text{Rem}_n(\bar{\mathbf{w}}) \triangleq \inf_{p \in \mathcal{P}_1(\mathcal{F}_{\mathcal{W}})} \left\{ \int_{\mathcal{W}} \|f_{\mathbf{w}} - f_{\bar{\mathbf{w}}}\|_{\mathbb{L}_2(\mu)}^2 p(d\mathbf{w}) + \frac{\beta}{n} D_{\text{KL}}(p\|\pi) \right\}. \quad (5)$$

Considering  $f_{\bar{\mathbf{w}}}$  as an approximator of  $f_{\mathbf{P}}$ , the right hand side of (4) can be seen as the sum of the *approximation error*  $\|f_{\bar{\mathbf{w}}} - f_{\mathbf{P}}\|_{\mathbb{L}_2(\mu)}$  and the *estimation error*  $\text{Rem}_n(\bar{\mathbf{w}})$ . The main goal of this paper is to analyze this estimation error and then to combine it with available bounds on the approximation error. Our approach will consist in replacing the infimum over all measures  $p$  by the infimum over Gaussian distributions, for which mathematical derivations are considerably simpler. It is well-known (see, for example, (McAllester, 2003; Alquier, 2009; Guedj, 2019)) that for a fixed  $\bar{\mathbf{w}}$ , the infimum in (5) is attained by the Gibbs distribution

$$p^*(d\mathbf{w}) \propto \exp \left\{ -\frac{n}{\beta} \|f_{\mathbf{w}} - f_{\bar{\mathbf{w}}}\|_{\mathbb{L}_2(\mu)}^2 \right\} \pi(d\mathbf{w}).$$

Furthermore in this case,

$$\text{Rem}_n(\bar{\mathbf{w}}) = -\frac{\beta}{n} \log \int_{\mathcal{W}} \exp \left\{ -\frac{n}{\beta} \|f_{\mathbf{w}} - f_{\bar{\mathbf{w}}}\|_{\mathbb{L}_2(\mu)}^2 \right\} \pi(d\mathbf{w}).$$

This expression is often referred to as the free energy. The content of the rest of this section can be seen as leveraging the variational formulation (5) for obtaining user-friendly upper bounds.

**Proposition 1** *Let Assumption  $(\sigma\text{-L})$  and Assumption  $(\mathcal{N})$  be satisfied. Recall that  $d = D_0 D_1 + D_1 D_2$  is the number of weights of the neural network and  $n$  is the sample size.*

i) *If Assumption  $(\sigma\text{-B})$  holds true, then*

$$\text{Rem}_n(\bar{\mathbf{w}}) \leq \frac{\beta}{2n} \left\{ \frac{\|\bar{\mathbf{w}}_1\|_{\mathbb{F}}^2}{\rho_1^2} + \frac{\|\bar{\mathbf{w}}_2\|_{\mathbb{F}}^2}{\rho_2^2} + d \log \left( 1 + \frac{2n(A_1 \rho_1^2 + A_2 \rho_2^2)}{d\beta} \right) \right\}$$

where  $A_1 = D_0 D_1 M_2^2 \|\bar{\mathbf{w}}_2\|_{\mathbb{F}}^2$  and  $A_2 = D_1 D_2 \mu(\mathcal{X}) M_{\sigma}^2$ .

ii) *If the activation function vanishes at the origin but is not necessarily bounded, then*

$$\text{Rem}_n(\bar{\mathbf{w}}) \leq \frac{\beta}{2n} \left\{ \frac{\|\bar{\mathbf{w}}_1\|_{\mathbb{F}}^2}{\rho_1^2} + \frac{\|\bar{\mathbf{w}}_2\|_{\mathbb{F}}^2}{\rho_2^2} + 2d \log \left( 1 + \frac{n(A_1 \rho_1^2 + A'_2 \rho_2^2 + A'_3 \rho_1^2 \rho_2^2)}{d\beta} \right) \right\}$$

where  $A'_2 = \bar{M}_2^2 \|\bar{\mathbf{w}}_1\|_{\mathbb{F}}^2 D_2$  and  $A'_3 = M_2^2 D_0 D_2$ .

Quantities  $A_1$ ,  $A_2$ ,  $A'_2$  and  $A'_3$  defined in the proposition, are independent of the sample size  $n$ , the temperature parameter  $\beta$  and the variances  $\rho_1$  and  $\rho_2$  of the prior distribution, but they are dimension-dependent.

There are two dual ways of drawing statistical insights from the above bounds on the estimation error. The first way is to consider  $\rho_1, \rho_2$  and  $D_1$  as “tuning parameters” of the algorithm, and to prove that for a suitable choice of these parameters the predictor  $\hat{f}_n$  is optimal. This line of thought is further developed in Section 6 below. The second way of interpreting the obtained bound is to see which functions are well estimated by  $\hat{f}_n$  based on  $\rho_1, \rho_2$  and  $D_1$ . This leads to the following result.

**Theorem 2** *Let  $\hat{f}_n$  be a method of aggregation of shallow neural networks  $\mathcal{F}_W = \{f_w(\mathbf{x}) = \mathbf{w}_2^\top \bar{\sigma}(\mathbf{w}_1^\top \mathbf{x}) : \mathbf{w}_1 \in \mathbb{R}^{D_0 \times D_1}; \mathbf{w}_2 \in \mathbb{R}^{D_1 \times D_2}\}$ , based on a prior distribution  $\pi$ , satisfying PAC-Bayes bound (2). Let Assumptions **(σ-L)** and **(N)** be satisfied. Then, for  $B_\ell = \rho_\ell \sqrt{2D_{\ell-1}D_\ell}$ ,  $\ell = 1, 2$ , we have*

$$\left( C_{\mathbf{PB}}^{-1} \mathbf{E}_{\mathbf{P}} [\|\hat{f}_n - f_{\mathbf{P}}\|_{\mathbb{L}_2(\mu)}^2] \right)^{1/2} \leq \inf_{\substack{\|\mathbf{w}_1\|_{\mathbf{F}} \leq B_1 \\ \|\mathbf{w}_2\|_{\mathbf{F}} \leq B_2}} \|f_{\bar{\mathbf{w}}} - f_{\mathbf{P}}\|_{\mathbb{L}_2(\mu)} + \left\{ \frac{\beta d}{n} \log \left( 3 + \frac{nE}{d\beta} \right) \right\}^{1/2}, \quad (6)$$

where the constant  $E$  is defined by

$$E = \begin{cases} 3B_2^2(B_1^2 M_2^2 + \mu(\mathcal{X}) M_\sigma^2), & \text{if } \sigma \text{ satisfies Assumption } \mathbf{(σ-B)}, \\ 3B_1^2 B_2^2 (M_2^2 + \bar{M}_2^2 / D_1), & \text{if } \sigma(0) = 0 \text{ but } \sigma \text{ is not necessarily bounded.} \end{cases}$$

An important consequence of this result is that the estimation error, upper bounded by the second term in (6), is of order  $\sqrt{D_1/n}$  (we assume that the input and the output dimensions are fixed and neglect logarithmic factors). This is similar to many non-parametric estimation methods. For instance, if the regression function is estimated by a histogram with  $K$  bins, the estimation error is generally of order  $\sqrt{K/n}$ . Thus, the number of units in the hidden layer of a neural network plays the same role as the number of bins in a histogram. This parameter  $D_1$  has to be chosen carefully, in order to control both the approximation error and the estimation error.

## 4. Examples of application

PAC-Bayes inequality is stated in (1) in a rather general form. In this section, we provide examples of learning problems and learning algorithms for which a version of (1) is satisfied.

### 4.1. Fixed design regression

Regression with deterministic design and additive errors is often used in nonparametric modeling. In the case of Gaussian errors, it corresponds to the observations

$$\mathbf{Y}_i = f_{\mathbf{P}}(\mathbf{x}_i) + \sigma \xi_i, \quad \xi_i \stackrel{\text{iid}}{\sim} \mathcal{N}(0, \mathbf{I}_{D_2}), \quad i = 1, \dots, n,$$

where  $\mathbf{x}_1, \dots, \mathbf{x}_n$  are given deterministic points and  $\mathcal{Y} = \mathbb{R}^{D_2}$ . In this case, the measure  $\mu$  is the empirical uniform distribution:  $\mu = \frac{1}{n} \sum_{i=1}^n \delta_{\mathbf{x}_i}$ .

There are many results of type (1) in the literature for regression with fixed design. In particular, [Leung and Barron \(2006\)](#); [Dalalyan and Tsybakov \(2007, 2008\)](#); [Dalalyan and Salmon](#)

(2012); [Dalalyan and Tsybakov \(2012b\)](#); [Dalalyan \(2020\)](#); [Rigollet and Tsybakov \(2012\)](#) established a PAC-Bayesian bound for the exponentially weighted aggregate defined by  $\hat{f}_n(\mathbf{Y}, \mathbf{x}) = \int_{\mathcal{W}} f_{\mathbf{w}}(\mathbf{x}) \hat{\theta}_{n,\mathbf{w}}(\mathbf{Y}) \pi(d\mathbf{w})$  with

$$\hat{\theta}_{n,\mathbf{w}}(\mathbf{Z}) = \frac{\exp\{-\frac{1}{\beta} \sum_{i=1}^n \|\mathbf{Y}_i - f_{\mathbf{w}}(\mathbf{x}_i)\|_2^2\}}{\int_{\mathcal{W}} \exp\{-\frac{1}{\beta} \sum_{i=1}^n \|\mathbf{Y}_i - f_{\mathbf{u}}(\mathbf{x}_i)\|_2^2\} \pi(d\mathbf{u})}.$$

Note that  $\mathbf{w} \mapsto \hat{\theta}_{n,\mathbf{w}}$  is a probability density on  $(\mathcal{W}, \pi)$ , often referred to as posterior density. As proved in [Leung and Barron \(2006\)](#); [Juditsky et al. \(2008\)](#); [Dalalyan and Tsybakov \(2007\)](#), the exponentially weighted aggregate  $\hat{f}_n(\mathbf{Y}, \mathbf{x})$  satisfies sharp PAC-Bayes bound (2) as soon as  $\beta \geq 4\sigma^2$ .

## 4.2. Random design regression

In the setting of iid observations, sharp PAC-Bayes inequality is valid for the mirror averaging (MA) estimator ([Juditsky et al., 2008](#); [Dalalyan and Tsybakov, 2012a](#); [Gerchinovitz, 2013](#)). We define the estimator in the case of regression with random design, and briefly mention below that similar results hold for density estimation and classification. Interested reader is referred to ([Juditsky et al., 2008](#); [Dalalyan and Tsybakov, 2012a](#)) for more detailed and comprehensive account on the topic. Note that similar inequalities are obtained for the  $Q$ -aggregation procedure ([Dai et al., 2012](#); [Lecué and Rigollet, 2014](#)).

The regression problem writes as in the previous example

$$\mathbf{Y}_i = f_{\mathbf{P}}(\mathbf{X}_i) + \sigma \xi_i, \quad \xi_i \perp\!\!\!\perp \mathbf{X}_i, \quad i = 1, \dots, n,$$

with  $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$ ,  $\mathcal{X} \subset \mathbb{R}^{D_0}$ ,  $\mathcal{Y} \subset \mathbb{R}^{D_2}$  and  $(\mathbf{X}_i, \mathbf{Y}_i)$  being iid. The natural choice of the measure  $\mu$  here is the marginal distribution of  $\mathbf{X}_i$  over  $\mathcal{X}$ .

The mirror averaging procedure satisfying (1) takes the form

$$\hat{f}_n(\mathbf{Z}, x) = \int_{\mathcal{W}} f_{\mathbf{w}}(x) \hat{\theta}_{n,\mathbf{w}}^{\text{MA}}(\mathbf{Z}) \pi(d\mathbf{w}) = \frac{1}{n+1} \sum_{m=0}^n \int_{\mathcal{W}} f_{\mathbf{w}}(x) \hat{\theta}_{m,\mathbf{w}}(\mathbf{Z}) \pi(d\mathbf{w}) \quad (7)$$

with  $\hat{\theta}_{0,\mathbf{w}} = 1$  and

$$\hat{\theta}_{n,\mathbf{w}}^{\text{MA}}(\mathbf{Z}) = \frac{1}{n+1} \sum_{m=0}^n \frac{\exp\{-\frac{1}{\beta} \sum_{i=1}^m Q(\mathbf{Z}_i, f_{\mathbf{w}})\}}{\int_{\mathcal{W}} \exp\{-\frac{1}{\beta} \sum_{i=1}^m Q(\mathbf{Z}_i, f_{\tilde{\mathbf{w}}})\} \pi(d\tilde{\mathbf{w}})}, \quad (8)$$

where  $Q : \mathcal{Z} \times \mathbb{L}_2(\mu) \mapsto \mathbb{R}$  is a mapping satisfying some assumptions under which the minimizer of the loss  $\ell : g \mapsto \ell(g, f)$  coincides with the minimizer of  $g \mapsto \mathbf{E}_{\mathbf{P}}[Q(\mathbf{Z}, g)]$ . In the case of regression, the mirror averaging estimator can be evaluated with the  $\ell_2$ -norm such that in (8), the function  $Q$  is given by  $Q(\mathbf{Z}_i, f_{\mathbf{w}}) = \|\mathbf{Y}_i - f_{\mathbf{w}}(\mathbf{X}_i)\|_2^2$ .

## 4.3. Density estimation

Consider the case where the elements of  $\mathbf{Z}^n = (\mathbf{Z}_1, \dots, \mathbf{Z}_n) \in \mathcal{Z}^n$  are iid random variables drawn from a distribution having  $f_{\mathbf{P}}$  as density with respect to a measure  $\mu$ . We aim to estimate  $f_{\mathbf{P}}$  and

measure the risk using the squared integrated error

$$\ell(\widehat{f}_n, f) = \|\widehat{f}_n - f\|_{\mathbb{L}_2(\mu)}^2 = \int_{\mathcal{X}} (\widehat{f}_n(\mathbf{x}) - f(\mathbf{x}))^2 \mu(d\mathbf{x})$$

such that the mapping  $Q$  in (8) can be defined by  $Q(\mathbf{x}, g) = \|g\|_{\mathbb{L}_2(\mu)}^2 - 2g(\mathbf{x})$ .

#### 4.4. Classification for $\Phi$ -risk

Consider the binary classification problem with  $\mathcal{Z} = \mathbb{R}^{D_0} \times \{-1, +1\}$  and assume that  $\mathbf{Z}^n = ((X_1, Y_1), \dots, (X_n, Y_n))$  are iid observations drawn from a distribution  $\mathbf{P}$  on  $\mathcal{Z}$ . For a twice differentiable convex function  $\Phi$ , the  $\Phi$ -risk of a classifier  $g : \mathbb{R}^{D_0} \rightarrow \{-1, +1\}$  is given by  $R_{\mathbf{P}}^{\Phi}[g] = \mathbf{E}_{\mathbf{P}}[\Phi(-Yg(X))]$ . In this setting, the loss function can be defined as  $\ell(g, f) = R_{\mathbf{P}}^{\Phi}[g] - R_{\mathbf{P}}^{\Phi}[f]$ , and the MA estimator given by (7)–(8) can be used with the function  $Q(z, g) = \Phi(-yg(x))$ .

### 5. Approximation bounds

The goal of this section is to review existing bounds on the approximation error of neural networks for different classes of functions. We are particularly interested in shallow networks and in bounds having explicit dependence on the width of the hidden layer. The main question of interest is the assessment of the distance between a given function and its best approximation by a one-hidden-layer network with  $D_1$  units in the hidden layer. Our focus is on Lipchtiz activation functions such as logistic, tanh, ReLU or quadratic (for bounded inputs). Because of major differences between the sigmoidal and ReLU activation functions, these two cases will be presented separately.

#### 5.1. Bounds for sigmoidal activation functions

For sigmoid activation functions we distinguish the probabilistic approach (Barron, 1993; Delyon et al., 1995; Maiorov and Meir, 2000; Maiorov, 2006) from the deterministic and constructive approaches (Mhaskar and Micchelli, 1994; Petrushev, 1998; Burger and Neubauer, 2001; Cao et al., 2008; Costarelli and Spigler, 2013a,b). For the set of univariate locally  $\alpha$ -Hölder continuous functions with  $\alpha \in (0, 1]$ , the constructive approach of (Cao et al., 2008) leads to an approximation error of order of  $D_1^{-\alpha}$  in  $\ell_{\infty}$ -norm. For  $\alpha > 1$ , Costarelli and Spigler (2013a) shows that the approximation error is  $O(D_1^{-1})$  both for univariate and multivariate functions.

For other classes of functions, a common feature of the results is the requirement of the existence of some type of integral transform (e.g., Fourier, Radon, wavelet) of the function  $f_{\mathbf{P}}$ . Each transform is tailored to a different “smoothness” class. An early example is the constructive approach from (Mhaskar and Micchelli, 1994) that focused on  $2\pi$ -periodic functions from  $\mathbb{L}_2([-\pi, \pi]^{D_0})$  with absolutely convergent Fourier series. For such functions, the approximation error is shown to be  $O(D_1^{-1/2})$ . In the case of random design, the seminal paper (Barron, 1993) established the upper bound  $O(D_1^{-1/2})$  for functions  $f$  satisfying  $\int_{\mathbb{R}^{D_0}} \|z\|_2 |\mathcal{F}[f](z)| dz < \infty$ , with  $\mathcal{F}[f]$  being the Fourier transform of  $f$ .

Note that in the papers mentioned in the previous paragraphs, the smoothness of the function and the dimension of the input variable do not appear in the error bound. In contrast with this, for Sobolev spaces, Petrushev (1998) showed how the dimension of the input space and the smoothness of the Sobolev space impact the approximation. Further, building on (Delyon et al., 1995), Maiorov and Meir (2000); Maiorov (2006) proved that the approximation error is  $O(D_1^{-r/D_0})$

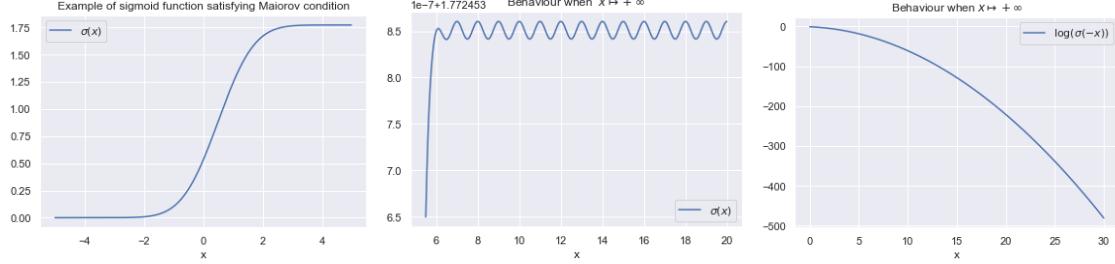


Figure 1: Activation function satisfying the Maiorov condition with  $\varphi(x) = (1/\sqrt{2})e^{-x^2/2}$

up to a  $\log(D_1)$ -factor. We use the results of [Maiorov and Meir \(2000\)](#) and [Maiorov \(2006\)](#) to upper bound the approximation error in (6). For  $f \in (\mathbb{L}_2 \cap \mathbb{L}_1)(\mathbb{R}^{D_0})$  with Fourier transform  $\mathcal{F}[f](z) = (2\pi)^{-D_0/2} \int_{\mathbb{R}^{D_0}} f(x) e^{iz^\top x} dx$ , we define  $D^\alpha f = \mathcal{F}^{-1}[|z|^\alpha \mathcal{F}[f](z)]$ . The unit Sobolev ball of smoothness  $r$  is then

$$W_2^r([0, 1]^{D_0}, \mu) = \left\{ f : \max_{0 \leq |\alpha| \leq r} \|D^\alpha f\|_{\mathbb{L}^2(\mu)} \leq 1 \right\}.$$

To present the precise statement of the result, let  $\varphi, \psi \in \mathbb{L}_2(\mathbb{R}) \cap \mathbb{L}_1(\mathbb{R})$  be functions satisfying

$$\int_0^\infty \frac{1}{a} \mathcal{F}[\varphi](az) \overline{\mathcal{F}[\psi]}(az) da = 1, \quad \forall z. \quad (9)$$

We define  $\Phi^r$  as the set of all functions  $\varphi \in \mathbb{L}_2(\mathbb{R}) \cap \mathbb{L}_1(\mathbb{R})$  such that there exists  $\psi$  satisfying (9) and  $\forall \rho \in [0, r]$ ,  $D^\rho \varphi \in \mathbb{L}_2(\mathbb{R})$ ,  $D^{-\rho} \psi \in \mathbb{L}_1(\mathbb{R})$ .

**Theorem 3 (Maiorov (2006), Theorem 2.3)** *Let  $\mu$  be a measure with a bounded density w.r.t. the Lebesgue measure and let  $\sigma$  be any sigmoid function such that the function  $\varphi(t) = \sigma(t+1) - \sigma(t) \in \Phi^r$ . Then, for any  $f \in W_2^r([0, 1]^{D_0}, \mu)$ , there exists a neural network  $f_{w^*}$  defined as in (3) such that*

$$\|f - f_{w^*}\|_{\mathbb{L}_2(\mu)} \leq c_1 B_\varphi \log(D_1) D_1^{-r/D_0} \quad \text{and} \quad |f_{w^*}(x)| \leq c_2 B_\varphi D_1^{(\frac{1}{2} - \frac{r}{D_0})_+}, \quad \forall x \in [0, 1]^{D_0},$$

where  $c_1$  and  $c_2$  are constants depending only on the problem dimension  $D_0$  and on the regularity parameter  $r$ , whereas  $B_\varphi = \max_{\rho \in [0, r]} \{\|D^\rho \varphi\|_{\mathbb{L}_2(\mathbb{R})}, \|D^{-\rho} \psi\|_{\mathbb{L}_1(\mathbb{R})}\}$ .

Examples of functions  $\varphi$  satisfying (9) are given in [Maiorov and Meir \(2000\)](#); [Maiorov \(2006\)](#) without a detailed analysis of the properties of the resulting function  $\sigma$ . The next result, proved in Appendix E, fills this gap for the example  $\varphi(x) = \frac{1}{\sqrt{2}}e^{-x^2/2}$ . This function satisfies (9) with  $\psi(x) = \frac{1}{\sqrt{2}}(1 - x^2)e^{-x^2/2}$ .

**Lemma 4** *Let  $\varphi(x) = (1/\sqrt{2})e^{-x^2/2}$  and define  $\sigma : \mathbb{R} \mapsto \mathbb{R}$  by  $\sigma(x) = \sum_{j=1}^\infty \varphi(x-j)$ . This function  $\sigma$  is 1-Lipschitz continuous, nonnegative, bounded from above by 2.5 and  $\lim_{x \rightarrow -\infty} \sigma(x) = 0$ .*

In Figure 1 we display the function  $\sigma$  defined in Theorem 4, as well as its limit behaviour when  $|x| \rightarrow +\infty$ . The left plot shows that  $\sigma$  looks very much like a standard sigmoid function. The middle and the right plot zoom on the limit behavior at  $+\infty$  and  $-\infty$ , respectively. We see, in particular,

that  $\sigma$  is not monotone when its values get close to its upper limit, but that it is bounded everywhere and tends exponentially fast to 0 at  $-\infty$ . We can also consider the case where  $\varphi(x) = \frac{(1-|x|)_+}{3}$ , for which we displayed, in Figure 2, the corresponding activation function  $\sigma$ . The function  $\sigma$  is derived using the same methodology as for the case of Theorem 4 (see also Appendix E).

## 5.2. Bounds for the ReLU activation function

The literature on neural networks with ReLU activation has significantly grown these last years thanks to the computational benefits of considering piecewise linear activation functions Yarotsky (2017, 2018); Yarotsky and Zhevnerchuk (2020); Gühring et al. (2020); Lu et al. (2020); Shen et al. (2019). We review below the results concerning shallow networks only, leaving aside the rich literature on approximation properties of deep networks.

For a Lipschitz function  $f$ , approximation error of order  $O(\eta D_1^{-1/D_0})$  is obtained in Bach (2017). Following the seminal work (Makovoz, 1996), results for Barron spectral spaces were developed in (Klusowski and Barron, 2016a; Xu, 2020; Siegel and Xu, 2020). Let  $\Omega \subset \mathbb{R}^{D_0}$  be a bounded domain and  $s > 0$ . The Barron spectral space of order  $s$  on  $\Omega$  is

$$\mathcal{B}^s(\Omega) := \left\{ f : \Omega \mapsto \mathbb{C} : \|f\|_{\mathcal{B}^s(\Omega)} := \inf_{f_e \mid \Omega = f} \int_{\mathbb{R}^d} (1 + \|z\|_2)^s |\mathcal{F}[f_e](z)| dz < \infty \right\},$$

where  $f_e$  is an  $L^1(\mathbb{R}^d)$  extension of  $f$ . It was shown in Klusowski and Barron (2016a) that the approximation error over  $\mathcal{B}^2([0, 1]^{D_0})$  is  $O(D_1^{-(D_0+2)/(2D_0)})$ . The same was proved to hold Xu (2020) for  $\text{ReLU}^k$  activation defined as  $\sigma^{(k)}(x) = \max(0, x)^k$ , when the target function is in  $\mathcal{B}^{k+1}([0, 1]^{D_0})$ . Very recently, Siegel and Xu (2020) made another step forward to assess the approximation error of shallow neural networks. This result being, to the best of our knowledge, the tightest one for shallow networks with  $\text{ReLU}^k$  activations, we provide its statement in the particular case of  $k = 1$ .

**Theorem 5 (Siegel and Xu (2020), Theorem 3)** *Let  $\Omega = [0, 1]^{D_0}$  and  $s \geq 1/2$ . If  $f \in \mathcal{B}^s(\Omega)$  and  $D_1 \geq 2$ , then*

$$\|f - f_{\mathbf{w}^*}\|_{\mathbb{L}_2(\Omega)} \leq C \|f\|_{\mathcal{B}^s(\Omega)} D_1^{-K} \log^m(D_1),$$

where  $C$  is a constant depending on  $s$  and  $D_0$  (but not on  $D_1$ ), whereas  $K$  and  $m$  are given by

$$K = \begin{cases} 2 & \text{if } 2s \geq 3D_0 + 4 \\ \frac{1}{2} + \frac{2s-1}{2(D_0+1)} & \text{if } 2s < 3D_0 + 4 \end{cases} \quad \text{and} \quad m = \begin{cases} 0 & \text{if } 2s < 3D_0 + 4 \\ 1 & \text{if } 2s > 3D_0 + 4 \\ \frac{5}{2} & \text{if } 2s = 3D_0 + 4. \end{cases} \quad (10)$$

Note that in the papers summarized in this section, the values of the constants—which may depend on the input dimension and on the smoothness—are not specified. An unfortunate consequence of this is that we can not keep track of the information on the role of the input dimension in the risk bounds stated in the next section.

## 6. Worst-case risk bounds over smoothness classes

This section is devoted to upper bounds on the minimax risk. We present risk bounds for networks with sigmoid activation functions prior to treating the case of ReLU activation.

## 6.1. Sigmoid activation functions

In this section we focus on real valued functions ( $D_2 = 1$ ) belonging to the unit ball of the Sobolev space,  $f_{\mathbf{P}} \in W_2^r([0, 1]^{D_0}, \mu)$ . Using Theorem 2 and Theorem 3, we can express both the estimation and the approximation error as functions of the size  $D_1$  of the hidden layer for an activation function that satisfies the conditions of Theorem 3. This leads to the risk bound

$$C_{\mathbf{PB}}^{-1} \mathbf{E}_{\mathbf{P}} [\|\widehat{f}_n - f_{\mathbf{P}}\|_{\mathbb{L}_2(\mu)}^2] \leq 2c_1^2 B_{\varphi}^2 \frac{\log^2(D_1)}{D_1^{2r/D_0}} + \frac{4\beta D_1 D_0}{n} \log \left( 3 + \frac{nE}{d\beta} \right), \quad (11)$$

where  $c_1 B_{\varphi}$  is as defined in Theorem 3, and  $E$  is defined in Theorem 2. We clearly see that  $D_1$ , the width of the hidden layer, controls the extent to which finer structure can be modeled. Reducing  $D_1$  decreases the estimation error since we have fewer parameters to estimate. But it increases the approximation error since we use a narrower class of approximators. Our goal below is to determine the value of  $D_1$  guaranteeing the best trade-off between approximation and estimation errors.

**Theorem 6 (Sigmoidal activation and Sobolev balls)** *Let  $\mathcal{X} = [0, 1]^{D_0}$  and  $r > 0$ . Let  $\widehat{f}_n$  be an aggregate of neural networks satisfying PAC-Bayes risk bound (2). If the measure  $\mu$  and the activation function  $\sigma$  satisfy conditions of Theorem 3, then the choice*

$$D_1 = \left( \frac{\beta D_0}{n} \right)^{-\frac{D_0}{2r+D_0}} \quad (12)$$

leads to the worst-case risk bound

$$\sup_{\mathbf{P}: f_{\mathbf{P}} \in W_2^r(\mathcal{X}, \mu)} \mathbf{E}_{\mathbf{P}} [\|\widehat{f}_n - f_{\mathbf{P}}\|_{\mathbb{L}_2(\mu)}^2] \leq g(n) \left( \frac{\beta D_0}{n} \right)^{2r/(2r+D_0)}, \quad (13)$$

where  $g(n)$  is the slowly varying function

$$g(n) = 2C_{\mathbf{PB}} \left( c_1^2 B_{\varphi}^2 \log^2(n/\beta) + 2 \log \left( 3 + \frac{3nB_2^2(B_1^2 M_2^2 + \mu(\mathcal{X}) M_{\sigma}^2)}{d\beta} \right) \right).$$

The proof of this theorem consists in substituting  $D_1$  in (11) by its expression (12). The obtained rate,  $n^{-2r/(2r+D_0)}$ , is the classical minimax rate of estimation over  $D_0$ -variate and  $r$ -smooth functions. We further discuss this result and compare it to prior work in Section 6.3. Note also that without any additional difficulty, one can extend (13) to the mis-specified case, *i.e.*, when  $f_{\mathbf{P}}$  does not necessarily belong to the Sobolev ball. The corresponding result then reads as follows:

$$(\mathbf{E}_{\mathbf{P}} [\|\widehat{f}_n - f_{\mathbf{P}}\|_{\mathbb{L}_2(\mu)}^2])^{1/2} \leq \inf_{f \in W_2^r(\mathcal{X}, \mu)} C_{\mathbf{PB}}^{1/2} \|f - f_{\mathbf{P}}\|_{\mathbb{L}_2(\mu)} + g(n)^{1/2} \left( \frac{\beta D_0}{n} \right)^{r/(2r+D_0)}.$$

## 6.2. ReLU activation function

In the case of ReLU activation, we will state risk bounds for two classes: the Barron spectral space and a specific Sobolev ball. Let us first assume that  $f_{\mathbf{P}} \in \mathcal{B}^s([0, 1]^{D_0})$  and that the conditions of Theorem 5 are satisfied. In view of Theorem 2 and Theorem 5, we have the risk bound

$$\mathbf{E}_{\mathbf{P}} [\|\widehat{f}_n - f_{\mathbf{P}}\|_{\mathbb{L}_2(\mu)}^2] \leq 2C_{\mathbf{PB}} C^2 D_1^{-2K(s, D_0)} \log^{2m}(D_1) + C_{\mathbf{PB}} \frac{4\beta D_1 D_0}{n} \log \left( 3 + \frac{nE}{d\beta} \right)$$

where  $C, K = K(s, D_0)$ ,  $m$  are as in Theorem 5 and  $E$  is as in Theorem 2. The bias-variance balance equation takes the form  $D_1^{-2K} \asymp \beta D_1/n$  and leads to the following proposition.

**Proposition 7 (ReLU activation and Barron spectral spaces)** *Let  $K = K(s, D_0)$  be as in (10) and let  $\hat{f}_n$  be an aggregate of neural networks satisfying PAC-Bayes risk bound (2). If the conditions of Theorem 5 hold then the choice  $D_1 = (\beta D_0/n)^{-1/(2K+1)}$  leads to the risk bound*

$$\sup_{\|f_{\mathbf{P}}\|_{\mathcal{B}^s([0,1]^{D_0})} \leq 1} \mathbf{E}_{\mathbf{P}} [\|\hat{f}_n - f_{\mathbf{P}}\|_{\mathbb{L}_2(\mu)}^2] \leq \bar{g}(n) \left( \frac{\beta D_0}{n} \right)^{2K/(2K+1)},$$

with the slowly varying function

$$\bar{g}(n) = 2C_{\text{PB}} C^2 \log^{2m}(n/\beta) + 4C_{\text{PB}} \log \left( 3 + \frac{3nB_1^2 B_2^2 (M_2^2 + \bar{M}_2^2)}{d\beta} \right)$$

and  $C$  is a constant that depends on  $s$  and  $D_0$  but not on  $D_1$ .

To get a risk over Sobolev spaces, we can rely on the inclusion  $W^{s+D_0/2+\varepsilon,2}(\mathcal{X}) \subset \mathcal{B}^s(\mathcal{X})$ , true for arbitrarily small  $\varepsilon > 0$  (Xu, 2020, Lemma 2.5). This is equivalent to  $W^{r,2}(\mathcal{X}) \subset \mathcal{B}^{\bar{r}-D_0/2}(\mathcal{X})$  for every  $r, \bar{r}$  such that  $D_0/2 \leq \bar{r} < r$ . Depending on the order of the Barron spectral space and the dimension of the problem, this might require a significant level of smoothness for the function  $f$  we want to approximate. Keeping this constraint in mind, we proceed with the next proposition which is more easily comparable to Theorem 6.

**Proposition 8 (ReLU activation and Sobolev space)** *Let  $r \in (D_0/2, 2D_0 + 2)$  and let  $\hat{f}_n$  be an aggregate satisfying (2). For every  $\bar{r} < r$  there is a slowly varying function  $g_{\bar{r}} : \mathbb{N} \rightarrow \mathbb{R}_+$  such that*

$$\sup_{f_{\mathbf{P}} \in W_2^r([0,1]^{D_0})} \mathbf{E}_{\mathbf{P}} [\|\hat{f}_n - f_{\mathbf{P}}\|_{\mathbb{L}_2(\mu)}^2] \leq g_{\bar{r}}(n) n^{-\frac{2\bar{r}}{2\bar{r}+D_0+1}}.$$

This result is weaker than the one of Theorem 6 in three aspects. First, it has the constraint  $r \in (D_0/2, 2D_0 + 2)$  limiting the order of smoothness of Sobolev classes. The constraint  $r < 2D_0 + 2$  stems from the fact that we want  $K(s, D_0)$  to take the value  $(2s + D_0)/(2D_0 + 2)$ . If  $r \geq 2D_0 + 2$ , the claim of the last proposition holds true if we replace  $2\bar{r}/(2\bar{r} + D_0 + 1)$  by  $4/5$ . The second weakness is that  $\bar{r}$ , present in the rate of convergence, is strictly smaller than the true smoothness  $r$ . Finally, the denominator in the exponent has an additional term increasing the dimension  $D_0$  by 1, leading thus to a slightly slower rate of convergence than the minimax rate over Sobolev balls. This is a direct consequence of approximation properties of ReLU neural nets in Sobolev spaces.

### 6.3. Related work on risk bounds for (penalized) ERM neural networks

For shallow neural networks with sigmoid activation, (Barron, 1994) showed that the risk of the suitably penalized empirical risk minimizer (ERM) is  $O(n^{-1/2} \log n)$ , provided that the function  $f$  is very smooth ( $\int \|z\|_1 |\mathcal{F}[f](z)| dz < \infty$ ). This was improved to  $O(n^{-2\bar{r}/(2\bar{r}+D_0+5)})$ ,  $\forall \bar{r} < r$ , for specific cosine activation McCaffrey and Gallant (1994). To our knowledge, this is the best known result for a one-hidden-layer network provided by ERM. In the case of two-hidden-layer networks with sigmoid activation, the rate  $O(n^{-2r/(2r+D_0)} \log^3 n)$  was obtained in Bauer and Kohler (2019) for functions satisfying a generalized hierarchical interaction model. Our risk bound (13), of order  $O(n^{-2r/(2r+D_0)} \log n)$ , matches the nonparametric minimax rate (Stone, 1982; Tsybakov, 2008), and is better than known rates for the ERM networks with one hidden layer. Roughly speaking, this

shows that aggregation acts as an additional layer, so that the aggregated one-hidden-layer networks achieve the same rate as the ERM two-hidden-layer networks.

We switch now to neural networks with ReLU activation functions. For one-hidden-layer networks, [Bach \(2017\)](#) established a risk bound of order  $n^{-2/(D_0+3)}$ . On a related note, [Klusowski and Barron \(2016b\)](#) considered bounded ramp activation functions and the low dimensional setting  $D_0 \ll n$ . For functions belonging to  $\mathcal{B}^2([0, 1]^{D_0})$ , they proved that the risk of the penalized ERM is  $O(n^{-(D_0+4)/(2D_0+6)})$ . This result can be directly compared to ours, in the particular case  $s = 2$ ; Theorem 7 and the fact that  $2K/(2K+1) = (2s+D_0)/(2s+2D_0+1) = (D_0+4)/(2D_0+5)$ , yield a leading term of order  $O(n^{-(D_0+4)/(2D_0+5)})$ . This improves the result of [Klusowski and Barron \(2016b\)](#) by a factor  $O(n^{-(D_0+4)/2(3+D_0)(2D_0+5)})$ . For instance, if  $D_0 = 3$  or  $D_0 = 4$ , we get the improvement factors  $n^{-7/132}$  and  $n^{-4/91}$ , respectively. This improvement vanishes when  $D_0$  increases to infinity. For multilayer ReLU networks, [Schmidt-Hieber \(2020\)](#) established the counterpart of the risk bound of [Bauer and Kohler \(2019\)](#) for  $\beta$ -Hölder functions. In particular, the worst-case risk was shown to be  $O(n^{-2\beta/(2\beta+D_0)})$ , see also ([Suzuki, 2019](#)) for an analogous result over Besov spaces. Hence, the minimax rate is achieved by the ERM over multilayer ReLU networks. In view of Theorem 8, this provides a bound for the ERM over multilayer networks smaller by a factor  $O(n^{-2\bar{r}/(2\bar{r}+D_0+1)(2\bar{r}+D_0)})$  than the bound for the aggregate of one-hidden-layer networks.

## 7. Conclusion and outlook

We have analyzed the estimation error of an aggregate of neural networks having one hidden layer and Lipschitz continuous activation function, under the condition that the aggregate satisfies the PAC-Bayes inequality. We focused our attention on Gaussian priors and obtained risk bounds in which the dependence on all the involved parameters is explicit. All these bounds on the estimation error come with explicit constants. We then combined our bounds on the estimation error with bounds on approximation error available in the literature. This allowed us to prove that aggregation of one-layer neural networks achieves the minimax risk over conventional smoothness classes. On the down side, since the constants in the bounds on the approximation error available in the literature are not explicit, the same is true for risk bounds of the present work. Therefore, it would be highly relevant to refine the existing approximation bounds to make appear all the constants.

The results of the present work can be extended in different directions. First, it would be interesting to consider the problem of aggregation of deep neural networks in order to understand possible benefits of increasing the depth. Second, it might be relevant to analyze the case of a prior with heavier tails, such as the Laplace prior or the Student prior, with a hope to cover the case of high dimension  $D_0 > n$  under some kind of sparsity assumption. Finally, another avenue of future research is to explore the computational benefits of considering aggregated neural networks in conjunction with the Langevin-type algorithms.

## Acknowledgments

Arnak Dalalyan's research was supported by the grant Investissements d'Avenir (ANR-11-IDEX-0003/Labex Ecodec/ANR-11-LABX-0047) and by the center Hi!PARIS.

## References

Pierre Alquier. PAC-Bayesian bounds for randomized empirical risk minimizers. *Mathematical methods of statistics*, 17(4):279–304, 2009.

Pierre Alquier. User-friendly introduction to PAC-Bayes bounds. Technical report, 2021. URL <https://arxiv.org/abs/2110.11216>.

Pierre Alquier and Gérard Biau. Sparse single-index model. *J. Mach. Learn. Res.*, 14(1):243–280, 2013.

Martin Anthony and Peter L. Bartlett. *Neural Network Learning: Theoretical Foundations*. Cambridge University Press, 1999.

Jimmy Ba, Murat A. Erdogdu, Taiji Suzuki, Denny Wu, and Tianzong Zhang. Generalization of two-layer neural networks: An asymptotic viewpoint. In *ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*, 2020.

Francis Bach. Breaking the curse of dimensionality with convex neural networks. *The Journal of Machine Learning Research*, 18(1):629–681, 2017.

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

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

Peter L. Bartlett, Vitaly Maiorov, and Ron Meir. Almost linear VC-dimension bounds for piecewise polynomial networks. *Neural Comput.*, 10(8):2159–2173, 1998.

Peter L. Bartlett, Nick Harvey, Christopher Liaw, and Abbas Mehrabian. Nearly-tight VC-dimension and pseudodimension bounds for piecewise linear neural networks. *Journal of Machine Learning Research*, 20(63):1–17, 2019.

Peter L. Bartlett, Andrea Montanari, and Alexander Rakhlin. Deep learning: a statistical viewpoint. *Acta Numerica*, 30:87–201, 2021.

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

Felix Biggs and Benjamin Guedj. Differentiable PAC-Bayes objectives with partially aggregated neural networks. *Entropy*, 23(10):1280, 2021.

Felix Biggs and Benjamin Guedj. On margins and derandomisation in pac-bayes. *International Conference on Artificial Intelligence and Statistics*, pages 3709–3731, 2022a.

Felix Biggs and Benjamin Guedj. Non-vacuous generalisation bounds for shallow neural networks. *arXiv preprint arXiv:2202.01627*, 2022b.

Martin Burger and Andreas Neubauer. Error bounds for approximation with neural networks. *Journal of Approximation Theory*, 112(2):235–250, 2001.

Feilong Cao, Tingfan Xie, and Zongben Xu. The estimate for approximation error of neural networks: a constructive approach. *Neurocomputing*, 71(4-6):626–630, 2008.

Yuan Cao and Quanquan Gu. Tight sample complexity of learning one-hidden-layer convolutional neural networks. In *NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 10611–10621, 2019.

Olivier Catoni. PAC-Bayesian supervised classification: The thermodynamics of statistical learning. *Lecture Notes-Monograph Series*, 56, 2007.

Danilo Costarelli and Renato Spigler. Approximation results for neural network operators activated by sigmoidal functions. *Neural Networks*, 44:101–106, 2013a.

Danilo Costarelli and Renato Spigler. Multivariate neural network operators with sigmoidal activation functions. *Neural Networks*, 48:72–77, 2013b.

Dong Dai, Philippe Rigollet, and Tong Zhang. Deviation optimal learning using greedy  $Q$ -aggregation. *The Annals of Statistics*, 40(3):1878 – 1905, 2012.

Arnak S. Dalalyan. Exponential weights in multivariate regression and a low-rankness favoring prior. *Ann. Inst. Henri Poincaré Probab. Stat.*, 56(2):1465–1483, 2020.

Arnak S. Dalalyan and Joseph Salmon. Sharp oracle inequalities for aggregation of affine estimators. *The Annals of Statistics*, 40(4):2327 – 2355, 2012.

Arnak S. Dalalyan and Alexandre B. Tsybakov. Aggregation by exponential weighting and sharp oracle inequalities. In *International Conference on Computational Learning Theory*, pages 97–111. Springer, 2007.

Arnak S. Dalalyan and Alexandre B. Tsybakov. Aggregation by exponential weighting, sharp PAC-Bayesian bounds and sparsity. *Mach. Learn.*, 72(1-2):39–61, 2008.

Arnak S. Dalalyan and Alexandre B. Tsybakov. Mirror averaging with sparsity priors. *Bernoulli*, 18(3):914–944, 2012a.

Arnak S. Dalalyan and Alexandre B. Tsybakov. Sparse regression learning by aggregation and Langevin Monte-Carlo. *Journal of Computer and System Sciences*, 78(5):1423–1443, 2012b.

B. Delyon, A. Juditsky, and A. Benveniste. Accuracy analysis for wavelet approximations. *IEEE Transactions on Neural Networks*, 6(2):332–348, 1995. doi: 10.1109/72.363469.

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

Jianqing Fan, Cong Ma, and Yiqiao Zhong. A Selective Overview of Deep Learning. *Statistical Science*, 36(2):264 – 290, 2021.

Louis Fortier-Dubois, Gaël Letarte, Benjamin Leblanc, François Laviolette, and Pascal Germain. Learning aggregations of binary activated neural networks with probabilities over representations. *CoRR*, abs/2110.15137, 2021. URL <https://arxiv.org/abs/2110.15137>.

Sébastien Gerchinovitz. Sparsity regret bounds for individual sequences in online linear regression. *J. Mach. Learn. Res.*, 14:729–769, 2013. ISSN 1532-4435.

Pascal Germain, Alexandre Lacasse, François Laviolette, and Mario Marchand. Pac-bayesian learning of linear classifiers. In *Proceedings of the 26th Annual International Conference on Machine Learning*, ICML '09, page 353–360. Association for Computing Machinery, 2009. ISBN 9781605585161. doi: 10.1145/1553374.1553419.

Benjamin Guedj. A primer on PAC-Bayesian learning. *arXiv preprint arXiv:1901.05353*, 2019.

Ingo Gühring, Gitta Kutyniok, and Philipp Petersen. Error bounds for approximations with deep relu neural networks in  $W^s, p$  norms. *Analysis and Applications*, 18(05):803–859, 2020.

A. Juditsky, P. Rigollet, and A. B. Tsybakov. Learning by mirror averaging. *The Annals of Statistics*, 36(5):2183 – 2206, 2008.

Jason M. Klusowski and A. Barron. Uniform approximation by neural networks activated by first and second order ridge splines. *arXiv preprint arXiv:1607.07819v1*, 2016a.

Jason M Klusowski and Andrew R Barron. Risk bounds for high-dimensional ridge function combinations including neural networks. *arXiv preprint arXiv:1607.01434*, 2016b.

Guillaume Lecué and Philippe Rigollet. Optimal learning with Q-aggregation. *The Annals of Statistics*, 42(1):211 – 224, 2014.

Gaël Letarte, Pascal Germain, Benjamin Guedj, and François Laviolette. Dichotomize and generalize: PAC-Bayesian binary activated deep neural networks. In *NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 6869–6879, 2019.

G. Leung and A. R. Barron. Information theory and mixing least-squares regressions. *IEEE Transactions on Information Theory*, 52(8):3396–3410, 2006. doi: 10.1109/TIT.2006.878172.

Gilbert Leung and Andrew R. Barron. Information theory and mixing least-squares regressions. *IEEE Trans. Inform. Theory*, 52(8):3396–3410, 2006.

Guy Lever, François Laviolette, and John Shawe-Taylor. Tighter PAC-Bayes bounds through distribution-dependent priors. *Theor. Comput. Sci.*, 473:4–28, 2013.

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

VE Maiorov and Ron Meir. On the near optimality of the stochastic approximation of smooth functions by neural networks. *Advances in Computational Mathematics*, 13(1):79–103, 2000.

Vitaly Maiorov. Approximation by neural networks and learning theory. *Journal of Complexity*, 22(1):102–117, 2006.

Y. Makovoz. Random approximants and neural networks. *Journal of Approximation Theory*, 85(1):98–109, 1996.

David A McAllester. Some PAC-Bayesian theorems. *Machine Learning*, 37(3):355–363, 1999.

David A McAllester. PAC-Bayesian stochastic model selection. *Machine Learning*, 51(1):5–21, 2003.

Daniel F McCaffrey and A Ronald Gallant. Convergence rates for single hidden layer feedforward networks. *Neural Networks*, 7(1):147–158, 1994.

Hrushikesh Narhar Mhaskar and Charles A Micchelli. Dimension-independent bounds on the degree of approximation by neural networks. *IBM Journal of Research and Development*, 38(3):277–284, 1994.

Behnam Neyshabur, Srinadh Bhojanapalli, David Mcallester, and Nati Srebro. Exploring generalization in deep learning. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

Maria Perez-Ortiz, Omar Rivasplata, Benjamin Guedj, Matthew Gleeson, Jingyu Zhang, John Shawe-Taylor, Miroslaw Bober, and Josef Kittler. Learning PAC-Bayes priors for probabilistic neural networks. Submitted., 2021a.

Maria Perez-Ortiz, Omar Rivasplata, Emilio Parrado-Hernandez, Benjamin Guedj, and John Shawe-Taylor. Progress in self-certified neural networks. In *NeurIPS 2021 workshop Bayesian Deep Learning [BDL]*, 2021b.

Maria Pérez-Ortiz, Omar Rivasplata, John Shawe-Taylor, and Csaba Szepesvári. Tighter risk certificates for neural networks. *Journal of Machine Learning Research*, 22, 2021.

Pencho P Petrushev. Approximation by ridge functions and neural networks. *SIAM Journal on Mathematical Analysis*, 30(1):155–189, 1998.

Philippe Rigollet and Alexandre B. Tsybakov. Sparse Estimation by Exponential Weighting. *Statistical Science*, 27(4):558 – 575, 2012.

Omar Rivasplata, Emilio Parrado-Hernandez, John S Shawe-Taylor, Shiliang Sun, and Csaba Szepesvari. PAC-Bayes bounds for stable algorithms with instance-dependent priors. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.

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

Zuowei Shen, Haizhao Yang, and Shijun Zhang. Deep network approximation characterized by number of neurons. *arXiv preprint arXiv:1906.05497*, 2019.

Jonathan W Siegel and Jinchao Xu. High-order approximation rates for neural networks with  $\text{relu}^k$  activation functions. *arXiv preprint arXiv:2012.07205*, 2020.

Charles J Stone. Optimal global rates of convergence for nonparametric regression. *The Annals of Statistics*, pages 1040–1053, 1982.

Taiji Suzuki. Adaptivity of deep ReLU network for learning in Besov and mixed smooth Besov spaces: optimal rate and curse of dimensionality. In *ICLR 2019, New Orleans, USA*, 2019.

Alexandre B Tsybakov. *Introduction to nonparametric estimation*. Springer Science & Business Media, 2008.

Bo Xie, Yingyu Liang, and Le Song. Diverse neural network learns true target functions. In *AISTATS 2017*, volume 54 of *Proceedings of Machine Learning Research*, pages 1216–1224. PMLR, 2017.

Jinchao Xu. The finite neuron method and convergence analysis. *arXiv preprint arXiv:2010.01458*, 2020.

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

Dmitry Yarotsky. Optimal approximation of continuous functions by very deep relu networks. In *Conference on Learning Theory*, pages 639–649. PMLR, 2018.

Dmitry Yarotsky and Anton Zhevnerchuk. The phase diagram of approximation rates for deep neural networks. In *NeurIPS 2020, December 6-12*, 2020.

Anatolii Borisovich Yuditskii, Aleksandr Viktorovich Nazin, Aleksandr Borisovich Tsybakov, and Nikolas Vayatis. Recursive aggregation of estimators by mirror descent algorithm with averaging. *Problemy Perekhodni Informatsii*, 41(4):78–96, 2005.

Kai Zhong, Zhao Song, Prateek Jain, Peter L. Bartlett, and Inderjit S. Dhillon. Recovery guarantees for one-hidden-layer neural networks. In *ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, volume 70 of *Proceedings of Machine Learning Research*, pages 4140–4149. PMLR, 2017.

Wenda Zhou, Victor Veitch, Morgane Austern, Ryan P. Adams, and Peter Orbanz. Non-vacuous generalization bounds at the imagenet scale: a PAC-Bayesian compression approach. In *ICLR 2019, New Orleans, USA, May 6-9*, 2019.

## Appendix A. Preliminary remark

As a preliminary remark let us note that, as a mixing measure, we expect the distribution  $p$  to aggregate the predictors  $f_w$  so that the resulting estimator is almost as good as the best predictors in  $\mathcal{F}_W$ . A direct consequence of it is that “a good choice” of  $p$  should be centered in  $\bar{w}$ . This is an heuristic way to choose the mean, and all along the appendix we will fix the distribution of  $p$  as

$$p = p_1 \otimes p_2 \sim \mathcal{N}(\bar{w}_1, \tau_1^2 \mathbf{I}_{D_1 D_0}) \otimes \mathcal{N}(\bar{w}_2, \tau_2^2 \mathbf{I}_{D_1 D_2}), \quad \tau_1, \tau_2 > 0 \quad (14)$$

where  $\bar{w} \in \operatorname{argmin}_{w \in W} \|f_w - f_P\|_{\mathbb{L}_2(\mu)}$ . The additional condition (14) is the starting point of our choice for  $p$ , it is now left to set values for the variance  $(\tau_1^2, \tau_2^2)$ .

## Appendix B. Some useful lemmas

In what follows, when appropriate, we will write  $f_{w_1, w_2}$  instead of  $f_w$ .

**Lemma 9** *If the probability distribution  $p$  is such that  $p(d\mathbf{w}) = p_1(d\mathbf{w}_1)p_2(d\mathbf{w}_2)$  with*

$$\int_{W_2} \mathbf{w}_2 p_2(d\mathbf{w}_2) = \bar{\mathbf{w}}_2$$

*then*

$$\int_W \|f_{\mathbf{w}} - f_{\bar{\mathbf{w}}}\|_{L_2(\mu)}^2 p(d\mathbf{w}) = \int_W \|f_{\mathbf{w}} - f_{\mathbf{w}_1, \bar{\mathbf{w}}_2}\|_{L_2(\mu)}^2 p(d\mathbf{w}) + \int_{W_1} \|f_{\mathbf{w}_1, \bar{\mathbf{w}}_2} - f_{\bar{\mathbf{w}}}\|_{L_2}^2 p_1(d\mathbf{w}_1).$$

**Proof** Simple algebra yields

$$\begin{aligned} \int_W (f_{\mathbf{w}} - f_{\bar{\mathbf{w}}})^2(\mathbf{x}) p(d\mathbf{w}) &= \int_W (f_{\mathbf{w}} - f_{\mathbf{w}_1, \bar{\mathbf{w}}_2} + f_{\mathbf{w}_1, \bar{\mathbf{w}}_2} - f_{\bar{\mathbf{w}}})^2(\mathbf{x}) p(d\mathbf{w}) \\ &= \int_W (f_{\mathbf{w}} - f_{\mathbf{w}_1, \bar{\mathbf{w}}_2})^2(\mathbf{x}) p(d\mathbf{w}) + \int_W (f_{\mathbf{w}_1, \bar{\mathbf{w}}_2} - f_{\bar{\mathbf{w}}})^2(\mathbf{x}) p(d\mathbf{w}) \\ &\quad + 2 \underbrace{\int_W (f_{\mathbf{w}_1, \bar{\mathbf{w}}_2} - f_{\mathbf{w}_1, \bar{\mathbf{w}}_2})(\mathbf{x}) (f_{\mathbf{w}_1, \bar{\mathbf{w}}_2} - f_{\bar{\mathbf{w}}_1, \bar{\mathbf{w}}_2})(\mathbf{x}) p(d\mathbf{w})}_{\triangleq A}. \end{aligned}$$

To complete the proof it suffices to integrate the previous equality with respect to  $\mu(d\mathbf{x})$  in virtue of Fubini-Tonelli theorem and to check that  $A = 0$ . The latter property follows from the fact that  $p$  is a product measure and, for all  $\mathbf{w}_1 \in W_1$ ,

$$\int_{W_2} (f_{\mathbf{w}_1, \bar{\mathbf{w}}_2} - f_{\mathbf{w}_1, \bar{\mathbf{w}}_2})(\mathbf{x}) p_2(d\mathbf{w}_2) = \int_{W_2} (\bar{\mathbf{w}}_2 - \bar{\mathbf{w}}_2)^\top \bar{\sigma}(\mathbf{w}_1^\top \mathbf{x}) p_2(d\mathbf{w}_2) = 0.$$

This yields the claim of the lemma. ■

In this section and the next one, let us define the two quantities:

$$G_1(\mathbf{w}) = \|f_{\mathbf{w}} - f_{\mathbf{w}_1, \bar{\mathbf{w}}_2}\|_{L_2(\mu)}^2, \quad \text{and} \quad G_2(\mathbf{w}_1) = \|f_{\mathbf{w}_1, \bar{\mathbf{w}}_2} - f_{\bar{\mathbf{w}}}\|_{L_2}^2.$$

**Lemma 10** *If Assumptions [\(σ-L\)](#) and  $M_2 < \infty$  are satisfied, and  $p$  is chosen as in [\(14\)](#), then*

$$\int_{W_1} G_2(\mathbf{w}_1) p_1(d\mathbf{w}_1) \leq D_0 (M_2 \tau_1 \|\bar{\mathbf{w}}_2\|_{1,2})^2 \leq C_1 D_0 D_1 \tau_1^2$$

with  $C_1 = (M_2 \|\bar{\mathbf{w}}_2\|_{\mathbb{F}})^2$ .

**Proof** [Proof of Theorem 10] We first use the fact that  $\sigma$  is 1-Lipschitz. On the one hand, in conjunction with the Fubini-Tonelli theorem, this yields

$$\begin{aligned} \int_{W_1} G_2(\mathbf{w}_1) p_1(d\mathbf{w}_1) &= \int_{\mathcal{X}} \int_W \left\| \bar{\mathbf{w}}_2^\top \left\{ \bar{\sigma}(\mathbf{w}_1^\top \mathbf{x}) - \bar{\sigma}(\bar{\mathbf{w}}_1^\top \mathbf{x}) \right\} \right\|_2^2 p(d\mathbf{w}) \mu(d\mathbf{x}) \\ &\leq \int_{\mathcal{X}} \int_W \sum_{j=1}^{D_2} \left( \sum_{i=1}^{D_1} |\bar{w}_{2,ij}| |(\mathbf{w}_1 - \bar{\mathbf{w}}_1)_i^\top \mathbf{x}| \right)^2 p(d\mathbf{w}) \mu(d\mathbf{x}) \\ &\leq \int_{\mathcal{X}} \sum_{j=1}^{D_2} \left( \sum_{i=1}^{D_1} |\bar{w}_{2,ij}| \left\{ \int_W |(\mathbf{w}_1 - \bar{\mathbf{w}}_1)_i^\top \mathbf{x}|^2 p(d\mathbf{w}) \right\}^{1/2} \right)^2 \mu(d\mathbf{x}) \\ &= D_0 M_2^2 \tau_1^2 \sum_{j=1}^{D_2} \left( \sum_{i=1}^{D_1} |\bar{w}_{2,ij}| \right)^2 \leq D_0 M_2^2 \tau_1^2 D_1 \|\bar{\mathbf{w}}_2\|_{\mathbb{F}}^2 \end{aligned}$$

and the claim of the lemma follows.  $\blacksquare$

In view of Theorem 9 and Theorem 10, we have

$$\int_{\mathcal{W}} \|f_{\mathbf{w}} - f_{\bar{\mathbf{w}}}\|_{\mathbb{L}_2(\mu)}^2 p(d\mathbf{w}) = \int_{\mathcal{W}} G_1(\mathbf{w}) p(d\mathbf{w}) + \int_{\mathcal{W}} G_2(\mathbf{w}_1) p_1(d\mathbf{w}_1).$$

and

$$\int_{\mathcal{W}} G_2(\mathbf{w}_1) p(d\mathbf{w}) \leq D_1 (M_2 \|\bar{\mathbf{w}}_2\|_{\mathbb{F}})^2 \tau_1^2.$$

We now state two distinct lemmas to bound the quantity  $\int_{\mathcal{W}} G_1(\mathbf{w}) p(d\mathbf{w})$ . Theorem 11 account for bounded activation functions whereas Theorem 12 focuses on unbounded ones.

**Lemma 11** *Under Assumption  $(\sigma\text{-B})$  and  $M_2 < \infty$ , if  $p$  is given by (14), we have*

$$\int_{\mathcal{W}} G_1(\mathbf{w}) p(d\mathbf{w}) \leq (M_{\sigma} \tau_2)^2 \mu(\mathcal{X}) D_1 D_2.$$

**Proof** [Proof of Theorem 11] Using Fubini-Tonelli theorem, we get

$$\int_{\mathcal{W}} G_1(\mathbf{w}) p(d\mathbf{w}) = \int_{\mathcal{X}} \int_{\mathcal{W}_1} \underbrace{\int_{\mathcal{W}_2} \left\| (\mathbf{w}_2 - \bar{\mathbf{w}}_2)^{\top} \bar{\sigma}(\mathbf{w}_1^{\top} \mathbf{x}) \right\|_{\mathbb{L}_2(\mu)}^2 p_2(d\mathbf{w}_2) p_1(d\mathbf{w}_1) \mu(d\mathbf{x})}_{:=I(\mathbf{x}, \mathbf{w}_1)}.$$

For the inner integral, simple algebra yields

$$\begin{aligned} I(\mathbf{x}, \mathbf{w}_1) &= \int_{\mathcal{W}} \bar{\sigma}(\mathbf{w}_1^{\top} \mathbf{x})^{\top} (\mathbf{w}_2 - \bar{\mathbf{w}}_2) (\mathbf{w}_2 - \bar{\mathbf{w}}_2)^{\top} \bar{\sigma}(\mathbf{w}_1^{\top} \mathbf{x}) p(d\mathbf{w}) \\ &= \bar{\sigma}(\mathbf{w}_1^{\top} \mathbf{x})^{\top} \left\{ \int_{\mathcal{W}_2} (\mathbf{w}_2 - \bar{\mathbf{w}}_2) (\mathbf{w}_2 - \bar{\mathbf{w}}_2)^{\top} p_2(d\mathbf{w}_2) \right\} \bar{\sigma}(\mathbf{w}_1^{\top} \mathbf{x}) \\ &= \tau_2^2 D_2 \|\bar{\sigma}(\mathbf{w}_1^{\top} \mathbf{x})\|_2^2. \end{aligned} \tag{15}$$

Therefore,

$$\int_{\mathcal{X}} \int_{\mathcal{W}_1} I(\mathbf{x}, \mathbf{w}_1) p_1(d\mathbf{w}_1) \mu(d\mathbf{x}) \leq M_{\sigma}^2 \mu(\mathcal{X}) \tau_2^2 D_1 D_2.$$

This completes the proof of the lemma.  $\blacksquare$

**Lemma 12** *Let  $\bar{M}_2 = \|\int_{\mathcal{X}} \mathbf{x} \mathbf{x}^{\top} \mu(d\mathbf{x})\|_{\text{sp}}$  be the spectral norm of the ‘‘covariance’’ matrix of the design. Under Assumption  $(\sigma\text{-L})$ , if  $p$  is given by (14) and  $\sigma(0) = 0$ , we have*

$$\int_{\mathcal{W}} G_1(\mathbf{w}) p(d\mathbf{w}) \leq M_2^2 D_0 D_1 D_2 \tau_1^2 \tau_2^2 + \bar{M}_2^2 D_2 \|\bar{\mathbf{w}}_1\|_{\mathbb{F}}^2 \tau_2^2.$$

**Proof** [Proof of Theorem 12] Using (15), we get

$$\begin{aligned}
 \int_W G_1(\mathbf{w}) p(d\mathbf{w}) &= \tau_2^2 D_2 \int_{\mathcal{X}} \int_{W_1} \|\bar{\sigma}(\mathbf{w}_1^\top \mathbf{x})\|_2^2 p_1(d\mathbf{w}_1) \mu(d\mathbf{x}) \\
 &\leq \tau_2^2 D_2 \int_{\mathcal{X}} \int_{W_1} \|\mathbf{w}_1^\top \mathbf{x}\|_2^2 p_1(d\mathbf{w}_1) \mu(d\mathbf{x}) \\
 &= \tau_2^2 D_2 \int_{\mathcal{X}} \int_{W_1} \|(\mathbf{w}_1 - \bar{\mathbf{w}}_1)^\top \mathbf{x}\|_2^2 p_1(d\mathbf{w}_1) \mu(d\mathbf{x}) + \tau_2^2 D_2 \int_{\mathcal{X}} \|(\bar{\mathbf{w}}_1)^\top \mathbf{x}\|_2^2 \mu(d\mathbf{x}) \\
 &= M_2^2 D_0 D_1 D_2 \tau_1^2 \tau_2^2 + \bar{M}_2^2 D_2 \|\bar{\mathbf{w}}_1\|_F^2 \tau_2^2.
 \end{aligned}$$

This completes the proof of the lemma.  $\blacksquare$

**Lemma 13** *Under Assumption  $(\sigma\text{-L})$  and  $M_2 < \infty$ , if  $p$  is given by (14), then*

$$\int_W \|f_{\mathbf{w}} - f_{\bar{\mathbf{w}}}\|_{\mathbb{L}_2(\mu)}^2 p(d\mathbf{w}) \leq M_2^2 \|\bar{\mathbf{w}}_2\|_F^2 D_0 D_1 \tau_1^2 + \bar{M}_2^2 D_2 \|\bar{\mathbf{w}}_1\|_F^2 \tau_2^2 + M_2^2 D_0 D_1 D_2 \tau_1^2 \tau_2^2. \quad (16)$$

If, in addition, Assumption  $(\sigma\text{-B})$  is satisfied, then

$$\int_W \|f_{\mathbf{w}} - f_{\bar{\mathbf{w}}}\|_{\mathbb{L}_2(\mu)}^2 p(d\mathbf{w}) \leq M_2^2 \|\bar{\mathbf{w}}_2\|_F^2 D_0 D_1 \tau_1^2 + \mu(\mathcal{X}) M_{\sigma}^2 D_1 D_2 \tau_2^2. \quad (17)$$

**Proof** In Theorem 9 we have checked that

$$\int_W \|f_{\mathbf{w}} - f_{\bar{\mathbf{w}}}\|_{\mathbb{L}_2(\mu)}^2 p(d\mathbf{w}) = \int_W G_1(\mathbf{w}) p(d\mathbf{w}) + \int_W G_2(\mathbf{w}_1) p_1(d\mathbf{w}_1).$$

Theorem 11 and Theorem 10 take care of both integrals in the right hand side of the equality for bounded activation functions and we directly get (17). Similarly, Theorem 12 and Theorem 10 can be applied for unbounded activation functions, leading to (16).  $\blacksquare$

## Appendix C. Proof of Proposition 1

Recall that the goal is to find an upper bound for the remainder term

$$\text{Rem}_n(\bar{\mathbf{w}}) \triangleq \inf_{p \in \mathcal{P}_1(\mathcal{F}_W)} \left\{ \int_W \|f_{\mathbf{w}} - f_{\bar{\mathbf{w}}}\|_{\mathbb{L}_2(\mu)}^2 p(d\mathbf{w}) + \frac{\beta}{n} D_{\text{KL}}(p||\pi) \right\}.$$

We start this proof by considering the case where Assumptions  $(\sigma\text{-L})$ ,  $(\sigma\text{-B})$  and  $(\mathcal{N})$  are satisfied. We choose as  $p$  the product of two spherical Gaussian distributions with variances  $\tau_1^2$  and  $\tau_2^2$ , as specified in (14). In this case, the Kullback-Leibler divergence  $D_{\text{KL}}(p||\pi)$  is given by

$$D_{\text{KL}}(p||\pi) = \frac{1}{2} \sum_{\ell=1}^2 \left\{ \frac{\|\bar{\mathbf{w}}_\ell\|_F^2}{\rho_\ell^2} + D_{\ell-1} D_\ell \left[ \left( \frac{\tau_\ell}{\rho_\ell} \right)^2 - 1 - \log \left( \frac{\tau_\ell^2}{\rho_\ell^2} \right) \right] \right\}.$$

It is now left to find good values for  $\tau_1^2$  and  $\tau_2^2$ . Combining with the result (17) of Theorem 13, we get the inequality

$$\text{Rem}_n(\bar{\mathbf{w}}) \leq \frac{\beta \|\bar{\mathbf{w}}_1\|_{\mathbb{F}}^2}{2n\rho_1^2} + \frac{\beta \|\bar{\mathbf{w}}_2\|_{\mathbb{F}}^2}{2n\rho_2^2} + \frac{\beta}{2n} \sum_{\ell=1}^2 D_{\ell-1} D_{\ell} \left\{ C_{\ell} \left( \frac{\tau_{\ell}}{\rho_{\ell}} \right)^2 - 1 - \log \left( \frac{\tau_{\ell}^2}{\rho_{\ell}^2} \right) \right\}$$

where

$$C_1 = \frac{2nM_2^2 \|\bar{\mathbf{w}}_2\|_{\mathbb{F}}^2 \rho_1^2}{\beta} + 1, \quad C_2 = \frac{2n\mu(\mathcal{X}) M_{\sigma}^2 \rho_2^2}{\beta} + 1.$$

One can easily check that the minimum of the function  $u \mapsto Cu - 1 - \log u$  is attained at  $u_{\min} = 1/C$  and the value at this point is  $\log C$ . This implies that

$$\begin{aligned} \text{Rem}_n(\bar{\mathbf{w}}) &\leq \frac{\beta \|\bar{\mathbf{w}}_1\|_{\mathbb{F}}^2}{2n\rho_1^2} + \frac{\beta \|\bar{\mathbf{w}}_2\|_{\mathbb{F}}^2}{2n\rho_2^2} + \frac{\beta}{2n} \sum_{\ell=1}^2 D_{\ell-1} D_{\ell} \log C_{\ell} \\ &\stackrel{(1)}{\leq} \frac{\beta \|\bar{\mathbf{w}}_1\|_{\mathbb{F}}^2}{2n\rho_1^2} + \frac{\beta \|\bar{\mathbf{w}}_2\|_{\mathbb{F}}^2}{2n\rho_2^2} + \frac{\beta d}{2n} \log \left( \frac{D_0 D_1 C_1 + D_1 D_2 C_2}{d} \right) \\ &\leq \frac{\beta \|\bar{\mathbf{w}}_1\|_{\mathbb{F}}^2}{2n\rho_1^2} + \frac{\beta \|\bar{\mathbf{w}}_2\|_{\mathbb{F}}^2}{2n\rho_2^2} + \frac{\beta d}{2n} \log \left( 1 + \frac{2n(D_0 D_1 M_2^2 \|\bar{\mathbf{w}}_2\|_{\mathbb{F}}^2 \rho_1^2 + D_1 D_2 \mu(\mathcal{X}) M_{\sigma}^2 \rho_2^2)}{\beta d} \right), \end{aligned} \quad (18)$$

where in (1) we have used the concavity of the function  $u \mapsto \log u$ . This completes the proof of the first claim of Theorem 1.

In the case where Assumption **(σ-B)** is not fulfilled, but instead  $\sigma(0) = 0$ , we repeat the same scheme of proof as above by using (16) instead of (17). This leads to

$$\begin{aligned} \text{Rem}_n(\bar{\mathbf{w}}) &\leq \frac{\beta d}{2n} \left\{ C'_1 \left( \frac{\tau_1}{\rho_1} \right)^2 + C'_2 \left( \frac{\tau_2}{\rho_2} \right)^2 + C'_3 \left( \frac{\tau_1}{\rho_1} \right)^2 \left( \frac{\tau_2}{\rho_2} \right)^2 - 2 - \log \left( \frac{\tau_1^2 \tau_2^2}{\rho_1^2 \rho_2^2} \right) \right\} \\ &\quad + \frac{\beta \|\bar{\mathbf{w}}_1\|_{\mathbb{F}}^2}{2n\rho_1^2} + \frac{\beta \|\bar{\mathbf{w}}_2\|_{\mathbb{F}}^2}{2n\rho_2^2}. \end{aligned} \quad (19)$$

where

$$C'_1 = \frac{2nD_0 M_2^2 \|\bar{\mathbf{w}}_2\|_{\mathbb{F}}^2 \rho_1^2}{\beta(D_0 + D_2)} + 1, \quad C'_2 = \frac{2n\bar{M}_2^2 \|\bar{\mathbf{w}}_1\|_{\mathbb{F}}^2 D_2 \rho_2^2}{\beta(D_0 + D_2) D_1} + 1, \quad C'_3 = \frac{2nM_2^2 \rho_1^2 \rho_2^2 D_0 D_2}{\beta(D_0 + D_2)}.$$

We choose  $\tau_1$  and  $\tau_2$  so that

$$\left( \frac{\tau_1}{\rho_1} \right)^2 = \frac{1}{C'_1 + C'_3 (\tau_2/\rho_2)^2}, \quad \left( \frac{\tau_2}{\rho_2} \right)^2 = 1/C'_2.$$

With this choice of  $\tau_1$  and  $\tau_2$  in (19) and simple algebra, we get

$$\text{Rem}_n(\bar{\mathbf{w}}) \leq \frac{\beta d}{2n} \log (C'_1 C'_2 + C'_3) + \frac{\beta \|\bar{\mathbf{w}}_1\|_{\mathbb{F}}^2}{2n\rho_1^2} + \frac{\beta \|\bar{\mathbf{w}}_2\|_{\mathbb{F}}^2}{2n\rho_2^2}.$$

To complete the proof, we use the following inequalities

$$\begin{aligned}\ln(C'_1 C'_2 + C'_3) &\leq \log(C'_1(C'_2 + C'_3)) \\ &= \log C'_1 + \log(C'_2 + C'_3) \\ &\leq 2 \ln((C'_1 + C'_2 + C'_3)/2),\end{aligned}$$

where the first inequality follows from the fact that  $C'_2 \geq 1$  whereas the last inequality is a consequence of the concavity of the logarithm.

**Remark 14** *The distribution  $p$  is centered on the oracle choice  $\bar{w}$  for the weights of the neural network and we observe that the optimized variances  $(\tau_1^2, \tau_2^2)$  in the proof of Theorem 1 are of the form  $\tau_\ell^2 = \rho_\ell^2/(1 + c_\ell n \rho_\ell^2)$ ,  $\ell = 1, 2$ , for some positive constants  $c_1, c_2$ . These values of  $\tau_\ell$  arbitrate between the prior beliefs and the information brought by data. Indeed, (1) when no training data is available the uncertainty around  $\bar{w}$  corresponds to the prior uncertainty  $(\rho_1^2, \rho_2^2)$ , (2) when the amount of observations  $n$  is unlimited and goes to infinity the uncertainty around the oracle value converges to 0 and  $p$  becomes close to the Dirac mass in  $\bar{w}$ .*

## Appendix D. Proof of Theorem 2

The main idea is to choose  $\rho_1$  and  $\rho_2$  minimizing the upper bound of the worst-case value of the remainder term

$$\sup_{\bar{w}: \|\bar{w}_\ell\|_F \leq B_\ell} \text{Rem}_n(\bar{w})$$

furnished by Theorem 1. Instead of using the exact minimizer, we use a surrogate obtained by simplifying expressions of  $\rho_1$  and  $\rho_2$ . This is done by the following result.

**Corollary 15** *Let Assumptions  $(\sigma\text{-L})$  and  $(\mathcal{N})$  be satisfied, set  $B_\ell = \rho_\ell \sqrt{2D_{\ell-1}D_\ell}$  for  $\ell = 1, 2$ .*

i) *If Assumption  $(\sigma\text{-B})$  holds true, then*

$$\sup_{\bar{w}: \|\bar{w}_\ell\|_F \leq B_\ell} \text{Rem}_n(\bar{w}) \leq \frac{\beta d}{n} \log \left( 3 + \frac{3nB_2^2(B_1^2M_2^2 + \mu(\mathcal{X})M_\sigma^2)}{d\beta} \right).$$

ii) *If the activation function is unbounded but vanishes at the origin, then*

$$\sup_{\bar{w}: \|\bar{w}_\ell\|_F \leq B_\ell} \text{Rem}_n(\bar{w}) \leq \frac{\beta d}{n} \log \left( 3 + \frac{3nB_1^2B_2^2(M_2^2 + \bar{M}_2^2/D_1)}{d\beta} \right).$$

The rest of this section is devoted to the proof of this claim, which implies the claim of Theorem 2. In view of (18), we have

$$\text{Rem}_n(\bar{w}) \leq \frac{\beta \|\bar{w}_1\|_F^2}{2n\rho_1^2} + \frac{\beta \|\bar{w}_2\|_F^2}{2n\rho_2^2} + \frac{\beta}{2n} \sum_{\ell=1}^2 D_{\ell-1}D_\ell \log(1 + F_\ell \rho_\ell^2)$$

with

$$F_1 = \frac{2nM_2^2\|\bar{\mathbf{w}}_2\|_{\mathbb{F}}^2}{\beta} \quad \text{and} \quad F_2 = \frac{2n\mu(\mathcal{X})M_{\sigma}^2}{\beta}.$$

Taking the maximum over all  $\bar{\mathbf{w}}$  such that the Frobenius norms of  $\bar{\mathbf{w}}_1$  and  $\bar{\mathbf{w}}_2$  are bounded by  $B_1$  and  $B_2$ , we get

$$\sup_{\|\bar{\mathbf{w}}_1\|_{\mathbb{F}} \leq B_1} \sup_{\|\bar{\mathbf{w}}_2\|_{\mathbb{F}} \leq B_2} \text{Rem}_n(\bar{\mathbf{w}}) \leq \frac{\beta B_1^2}{2n\rho_1^2} + \frac{\beta B_2^2}{2n\rho_2^2} + \frac{\beta}{2n} \sum_{\ell=1}^2 D_{\ell-1} D_{\ell} \log(1 + \bar{F}_{\ell} \rho_{\ell}^2) \quad (20)$$

with

$$\bar{F}_1 = \frac{2nM_2^2 B_2^2}{\beta} \quad \text{and} \quad \bar{F}_2 = \frac{2n\mu(\mathcal{X})M_{\sigma}^2}{\beta}.$$

The first order necessary condition for optimizing the right hand side with respect to  $\rho_1^2$  and  $\rho_2^2$  reads as

$$-\frac{B_{\ell}^2}{\rho_{\ell}^4} + \frac{D_{\ell-1} D_{\ell} \bar{F}_{\ell}}{1 + \bar{F}_{\ell} \rho_{\ell}^2} = 0 \iff \rho_{\ell}^4 - \frac{B_{\ell}^2}{D_{\ell-1} D_{\ell}} \rho_{\ell}^2 - \frac{B_{\ell}^2}{D_{\ell-1} D_{\ell} \bar{F}_{\ell}} = 0.$$

This second-order equation has only one positive root given by

$$\begin{aligned} \rho_{\ell}^2 &= \frac{B_{\ell}^2}{2D_{\ell-1} D_{\ell}} + \left( \frac{B_{\ell}^4}{4D_{\ell-1}^2 D_{\ell}^2} + \frac{B_{\ell}^2}{D_{\ell-1} D_{\ell} \bar{F}_{\ell}} \right)^{1/2} \\ &= \frac{B_{\ell}^2}{2D_{\ell-1} D_{\ell}} \left\{ 1 + \left( 1 + \frac{4D_{\ell-1} D_{\ell}}{B_{\ell}^2 \bar{F}_{\ell}} \right)^{1/2} \right\}. \end{aligned}$$

We simplify computations by choosing

$$\rho_{\ell}^2 = \frac{B_{\ell}^2}{2D_{\ell-1} D_{\ell}}.$$

Replacing these values of  $\rho_{\ell}^2$  in (20), we get

$$\begin{aligned} \sup_{\|\bar{\mathbf{w}}_{\ell}\|_{\mathbb{F}} \leq B_{\ell}} \text{Rem}_n(\bar{\mathbf{w}}) &\leq \frac{\beta}{n} \sum_{\ell=1}^2 D_{\ell-1} D_{\ell} \left\{ 1 + \frac{1}{2} \log \left( 1 + \frac{B_{\ell}^2 \bar{F}_{\ell}}{2D_{\ell-1} D_{\ell}} \right) \right\} \\ &\leq \frac{\beta d}{n} \sum_{\ell=1}^2 \left\{ 1 + \frac{1}{2} \log \left( 1 + \frac{B_1^2 \bar{F}_1 + B_2^2 \bar{F}_2}{2d} \right) \right\}, \end{aligned}$$

where the last inequality follows from the concavity of the logarithm. Replacing  $\bar{F}_1$  and  $\bar{F}_2$  with their respective expressions, we get the inequality

$$\begin{aligned} \sup_{\|\bar{\mathbf{w}}_{\ell}\|_{\mathbb{F}} \leq B_{\ell}} \text{Rem}_n(\bar{\mathbf{w}}) &\leq \frac{\beta d}{n} \left( 1 + \frac{1}{2} \log \left( 1 + \frac{nB_2^2(B_1^2 M_2^2 + \mu(\mathcal{X}) M_{\sigma}^2)}{d\beta} \right) \right) \\ &\leq \frac{\beta d}{n} \log \left( 3 + \frac{3nB_2^2(B_1^2 M_2^2 + \mu(\mathcal{X}) M_{\sigma}^2)}{d\beta} \right), \end{aligned}$$

which coincides with the first claim of the corollary.

The second claim of the proposition is obtained by replacing  $\rho_{\ell}$ 's by their respective expressions in the second claim of Theorem 1.

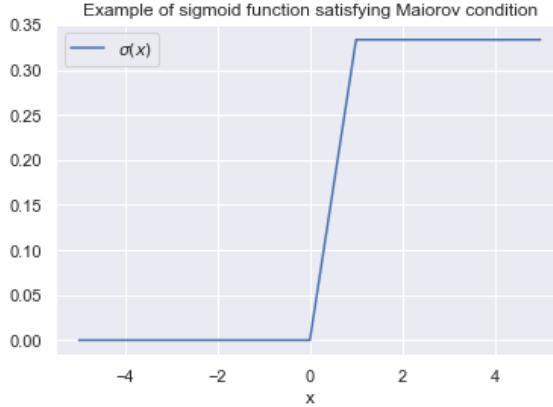


Figure 2: Sigmoid function satisfying the Maiorov condition with  $\varphi(x) = (1 - |x|)_+ / 3$ .

## Appendix E. Proof of Lemma 4

Since

$$\sigma(x) = \sum_{j=1}^{\infty} \varphi(x - j).$$

we have

$$\begin{aligned} \sigma(x+1) - \sigma(x) &= \sum_{j=1}^{\infty} \varphi(x+1 - j) - \sum_{j=1}^{\infty} \varphi(x - j) \\ &= \sum_{j=0}^{\infty} \varphi(x - j) - \sum_{j=1}^{\infty} \varphi(x - j) \\ &= \varphi(x). \end{aligned}$$

Now, recall that we use the function  $\varphi(x) = \frac{1}{\sqrt{2}} e^{-x^2/2}$ . It is clear, that the series

$$\sum_{j=1}^{\infty} |\varphi'(j - x)|$$

converges uniformly on any bounded interval. This implies that  $\sigma$  is differentiable and

$$\sigma'(x) = \sum_{j=1}^{\infty} \frac{(j - x)}{\sqrt{2}} e^{-(x-j)^2/2}.$$

Let us denote by  $[x]$  the integer part of  $x$  and by  $\{x\} = x - [x]$  the fractional part of  $x$ . Recall also that the function  $u \mapsto e^{-u^2/2}$  is increasing on  $(-\infty, 0]$  and decreasing on  $[0, +\infty)$ . Therefore, we

have

$$\begin{aligned}
 \sigma(x) &= \frac{1}{\sqrt{2}} \sum_{j=1}^{[x]} e^{-(x-j)^2/2} + \frac{1}{\sqrt{2}} \sum_{j=[x]+1}^{\infty} e^{-(x-j)^2/2} \\
 &\leq \frac{1}{\sqrt{2}} \sum_{j=1}^{[x]-1} \int_{x-j-1}^{x-j} e^{-u^2/2} du + \frac{e^{-\{x\}^2/2} + e^{-(1-\{x\})^2/2}}{\sqrt{2}} + \frac{1}{\sqrt{2}} \sum_{j=[x]+2}^{\infty} \int_{j-x-1}^{j-x} e^{-u^2/2} du \\
 &\leq \frac{1}{\sqrt{2}} \int_{\{x\}}^{x-1} e^{-u^2/2} du + \frac{e^{-\{x\}^2/2} + e^{-(1-\{x\})^2/2}}{\sqrt{2}} + \frac{1}{\sqrt{2}} \int_{-\infty}^{\{x\}-1} e^{-u^2/2} du \\
 &\leq \frac{1}{\sqrt{2}} \int_{-\infty}^{+\infty} e^{-u^2/2} du + \frac{1}{\sqrt{2}} = \sqrt{\pi} + \frac{1}{\sqrt{2}}.
 \end{aligned}$$

For  $x > 0$ , using similar arguments and the fact that the function  $u \mapsto ue^{-u^2/2}$  is decreasing on  $[1, \infty)$ , we get

$$\begin{aligned}
 \sqrt{2}\sigma'(x) &= - \sum_{j=1}^{[x]} (x-j) e^{-(x-j)^2/2} + \sum_{j=[x]+1}^{\infty} (j-x) e^{-(j-x)^2/2} \\
 &\leq (1-\{x\}) e^{-(1-\{x\})^2/2} + \sum_{j=[x]+2}^{\infty} \int_{j-x-1}^{j-x} ue^{-u^2/2} du \\
 &\leq (1-\{x\}) e^{-(1-\{x\})^2/2} + \int_{1-\{x\}}^{\infty} ue^{-u^2/2} du \\
 &= (2-\{x\}) e^{-(1-\{x\})^2/2} \leq \sqrt{2}.
 \end{aligned}$$

In the same way, one can check that  $\sqrt{2}\sigma'(x) \geq -\sqrt{2}$  for every  $x > 0$ . Therefore,  $|\sigma'(x)| \leq 1$  for every positive  $x$ . On the other hand, for  $x \leq 0$ , we have  $\sigma'(x) \geq 0$  and

$$\begin{aligned}
 \sigma'(x) &\leq \sigma'(0) = \frac{1}{\sqrt{2}} \sum_{j=1}^{\infty} j e^{-j^2/2} \\
 &\leq \frac{1}{\sqrt{2}} \left( e^{-1/2} + \sum_{j=2}^{\infty} \int_{j-1}^j ue^{-u^2/2} du \right) \\
 &= \frac{1}{\sqrt{2}} (e^{-1/2} + e^{-1/2}) \leq 1.
 \end{aligned}$$

This completes the proof of the fact that  $\sigma$  is 1-Lipschitz.

## Appendix F. Proof of Proposition 8

**Proof** Let assume  $r \in (\frac{D_0}{2}, 2D_0 + 2)$  and  $\bar{r} \in [D_0/2, r)$ . Then,  $W^{r,2}(\mathcal{X}) \subset \mathcal{B}^{\bar{r}-D_0/2}(\mathcal{X})$  (Xu, 2020, Lemma 2.5), and since  $\frac{2K}{2K+1} = \frac{2s+D_0}{2s+2D_0+1}$ , substituting  $s$  by  $\bar{r} - D_0/2$ , we obtain:

$$\frac{2K}{2K+1} = \frac{2\bar{r}}{2\bar{r} + D_0 + 1}.$$

■

Substituting the terms in Theorem 7, this yields the result with

$$\bar{g}_r(n) = 2C_{\text{PB}}C^2 + 4C_{\text{PB}} \log \left( 3 + \frac{3nB_1^2B_2^2(M_2^2 + \bar{M}_2^2)}{d\beta} \right).$$