# Misclassification bounds for PAC-Bayesian sparse deep learning

The Tien Mai

Department of Mathematical Sciences, Norwegian University of Science and Technology, Trondheim, 7034, Norway .

Corresponding author(s). E-mail(s): the.t.mai@ntnu.no;

**Abstract**

Recently, there has been a significant focus on exploring the theoretical aspects of deep learning, especially regarding its performance in classification tasks. Bayesian deep learning has emerged as a unified probabilistic framework, seeking to integrate deep learning with Bayesian methodologies seamlessly. However, there exists a gap in the theoretical understanding of Bayesian approaches in deep learning for classification. This study presents an attempt to bridge that gap. By leveraging PAC-Bayes bounds techniques, we present theoretical results on the prediction or misclassification error of a probabilistic approach utilizing Spike-and-Slab priors for sparse deep learning in classification. We establish non-asymptotic results for the prediction error. Additionally, we demonstrate that, by considering different architectures, our results can achieve minimax optimal rates in both low and high-dimensional settings, up to a logarithmic factor. Moreover, our additional logarithmic term yields slight improvements over previous works. Additionally, we propose and analyze an automated model selection approach aimed at optimally choosing a network architecture with guaranteed optimality.

**Keywords:** deep neural network, binary classification, prediction bounds, PAC-Bayes bounds, fast rate, low-rank priors

## 1 Introduction

Deep learning has achieved remarkable success in various applications, including computer vision, natural language processing, and other areas related to pattern recognition and classification (Goodfellow et al., 2016; LeCun et al., 2015). As its applicability continues to expand, there is a growing interest in understanding the

theoretical foundations that underlie its effectiveness. To gain theoretical insights into deep learning, numerous studies have been conducted from various perspectives, including approximation theory and statistical learning theory. The statistical learning community has extensively investigated the generalization properties of Deep Neural Networks (DNNs), (Barron, 1994; Bartlett et al., 2017; Neyshabur et al., 2018; Zhang et al., 2021; Schmidt-Hieber, 2020; Suzuki, 2018; Imaizumi and Fukumizu, 2019; Suzuki, 2019). Specifically, Schmidt-Hieber (2020) and Suzuki (2019) demonstrated that estimators based on sparsely connected DNNs with ReLU activation functions and carefully selected architectures can achieve minimax estimation rates (up to logarithmic factors) under classical smoothness assumptions on the regression function. Additionally, Imaizumi and Fukumizu (2019) and Hayakawa and Suzuki (2020) highlighted the superiority of DNNs over linear operators in certain scenarios, where DNNs achieve the minimax rate of convergence while alternative methods fall short.

On the Bayesian front, theoretical research has been relatively limited. Studies such as Polson and Ročková (2018); Suzuki (2018); Lee and Lee (2022); Kong et al. (2023) have explored the properties of the posterior distribution, while Vladimirova et al. (2019) investigated the regularization effects of prior distributions at the unit level. Furthermore, Chérief-Abdellatif (2020); Bai et al. (2020) examined properties related to variational inference.

In the realm of classification, theoretical investigations of deep learning classifiers have been undertaken quite recently in various setups. Such as Kim et al. (2021); Shen et al. (2022); Hu et al. (2022); Bos and Schmidt-Hieber (2022); Wang and Shang (2022); Kohler et al. (2022); Meyer (2023); Kohler and Langer (2024). While a variational Bayes classification for dense deep neural networks has been studied Bhattacharya et al. (2024), however, the prediction error for classification using Bayesian sparse deep learning remains unexplored. This study aims to address this gap by presenting an effort to shed light on this aspect.

We investigate a probabilistic approach for a sparse deep neural network classifier, where the network's architecture is governed by a Gibbs posterior incorporating a risk concept based on the hinge loss (Zhang, 2004). The hinge loss is commonly utilized for classification tasks in deep learning (Kim et al., 2021; Shen et al., 2022; Hu et al., 2022; Bos and Schmidt-Hieber, 2022; Wang and Shang, 2022; Kohler et al., 2022; Kohler and Langer, 2024). Our analysis of prediction errors relies on the PAC-Bayes bound technique, originally introduced by McAllester (1998); Shawe-Taylor and Williamson (1997). This technique aims to provide numerical generalization certificates for various probabilistic machine learning algorithms, and subsequently, Catoni (2004, 2007) demonstrated its utility in deriving oracle inequalities and rates of convergence for different methods. This methodology shares significant connections with the "information bounds" presented by Zhang (2006); Russo and Zou (2019). For a comprehensive exploration of this topic, we recommend consulting Guedj (2019); Alquier (2024).

Utilizing Spike-and-Slab priors, our main results are presented in terms of non-asymptotic oracle inequalities. These inequalities demonstrate that the prediction error of our proposed method is as good as the best possible errors. Our results are applicable in a general setting. Specifically, we show that by considering different architectures, our results can achieve minimax optimal rates for Hölder smooth functions in both

low and high dimensions, up to a logarithmic factor. This alignment with works in the frequentist literature, such as Kim et al. (2021) and Wang and Shang (2022), underscores the robustness of our findings. Moreover, our inclusion of an additional logarithmic term yields slight improvements over previous works. Furthermore, we also propose and analyze an automated model selection approach aimed at optimally choosing a network architecture with guaranteed optimality.

The rest of the paper is structured as follows: Section 2 introduces the problem, describes the deep neural network, and presents our proposed method. Our main results are outlined in Section 3. All technical proofs are provided in Section 4.

### Notations

Let us introduce the notations and the statistical framework we adopt in this paper. For any vector $x = (x_1, ..., x_d) \in [-1, 1]^d$ and any real-valued function $f$ defined on $[-1, 1]^d$, $d > 0$, we denote: $\|x\|_\infty = \max_{1\leq i\leq d} |x_i|, \|f\|_2 = (\int f^2)^{1/2}$ and $\|f\|_\infty = \sup_{y\in[-1,1]^d} |f(y)|$. Put $(a)_+ := \max(a, 0), \forall a \in \mathbb{R}$. For any $\mathbf{k} \in \{0, 1, 2, ...\}^d$, we define $|\mathbf{k}| = \sum_{i=1}^d k_i$ and the mixed partial derivatives when all partial derivatives up to order $|\mathbf{k}|$ exist: $D^{\mathbf{k}} f(x) = \frac{\partial^{|\mathbf{k}|} f}{\partial^{k_1} x_1 ... \partial^{k_d} x_d}(x)$. For $\beta > 0$, let $\lfloor\beta\rfloor$ denote the largest integer strictly smaller than $\beta$ and . Then $f$ is said to be $\beta$-Hölder continuous (Tsybakov, 2009) if all partial derivatives up to order $\lfloor\beta\rfloor$ exist and are bounded. The norm $\|f\|_{\mathcal{C}_\beta}$ of the Hölder space $\mathcal{C}_\beta = \{f : \|f\|_{\mathcal{C}_\beta} < +\infty\}$ is defined as:
$\|f\|_{\mathcal{C}_\beta} := \max_{|\mathbf{k}|\leq\lfloor\beta\rfloor} \|D^{\mathbf{k}} f\|_\infty + \max_{|\mathbf{k}|=\lfloor\beta\rfloor} \sup_{x,y\in[-1,1]^d, x\neq y} \frac{|D^{\mathbf{k}} f(x) - D^{\mathbf{k}} f(y)|}{\|x-y\|_\infty^{\beta-\lfloor\beta\rfloor}}$.

# 2 Problem and method

## 2.1 Problem statement

This study examines a general binary classification problem. Given a high-dimensional feature vector $x \in \mathbb{R}^d$, the class label outcome $Y$ is a binary variable derived from $\{-1, 1\}$ with an unknown probability $p(x)$. Specifically, $Y$ is equal to 1 with a probability of $p(x)$, and $-1$ with a probability of $1-p(x)$ when conditioned on $x$. The accuracy of a classifier $\eta$ is determined by the prediction or misclassification error, defined as

$$R(\eta) = \mathbb{P}(Y \neq \eta(x)).$$

However, the probability function $p(x)$ is unknown and the resulting classifier $\hat{\eta}(x)$ should be designed from the data $D_n$: a random sample of $n$ independent observations $(x_1, Y_1), \ldots, (x_n, Y_n)$. The design points $x_i$ may be considered as fixed or random. The corresponding (conditional) prediction error of $\hat{\eta}$ is $R(\hat{\eta}) = \mathbb{P}(Y \neq \hat{\eta}(x) \,|D_n)$. In this study, our objective is to construct a classifier $\hat{\eta}$ with a prediction error that closely approximates the ideal Bayes error $R^* := \inf R(\eta)$, (Vapnik, 1998; Devroye et al., 1996).

## 2.2 Deep neural networks

We define a deep neural network as any function $f_\theta : \mathbb{R}^d \to \mathbb{R}$, recursively defined as:

$$\begin{cases} x^{(0)} := x, \\ x^{(\ell)} := \rho(A_\ell x^{(\ell-1)} + b_\ell) \quad \text{for } \ell = 1, ..., L-1, \\ f_\theta(x) := A_L x^{(L-1)} + b_L \end{cases}$$

where $L \geq 3$ and $\rho$ is an activation function acting componentwise.

For example, we can employ the ReLU activation function, $\rho(u) = \max(u, 0)$. Each $A_\ell \in \mathbb{R}^{D_\ell \times D_{\ell-1}}$ serves as a weight matrix. Its entry at the $(i, j)$-th position, called edge weight, connects the $j$-th neuron from the $(\ell-1)$-th layer to the $i$-th neuron of the $\ell$-th layer. And, each node vector $b_\ell \in \mathbb{R}^{D_\ell}$ acts as a shift vector. Its $i$-th entry signifies the weight linked with the $i$-th node of the $\ell$-th layer.

Put $D_0 = d$ the number of units in the input layer, $D_L = 1$ the number of units in the output layer and $D_\ell = D$ the number of units in the hidden layers. The architecture of the network is characterized by its number of edges $S$ (the total number of nonzero entries in matrices $A_\ell$ and vectors $b_\ell$), its number of layers $L \geq 3$ (excluding the input layer), and its width $D \geq 1$. We have $S \leq T := \sum_{\ell=1}^{L} D_\ell(D_{\ell-1} + 1)$, the total number of coefficients in a fully connected network.

At this point, we assume that $S$, $L$, and $D$ are constants. We require $d \leq D$ and $T \leq LD(D+1)$. It is assumed that the absolute values of all coefficients are bounded above by a constant $C_B \geq 2$. The parameter for a Deep Neural Network (DNN) is denoted as $\theta = \{(A_1, b_1), \ldots, (A_L, b_L)\}$, and we define $\Theta_{S,L,D}$ as the set of all possible parameters. Alternatively, we will also consider the stacked coefficients parameter, $\theta = (\theta_1, \ldots, \theta_T)$.

## 2.3 EWA procedure using empirical hinge loss

We construct a classifier $\hat{\eta}(x) := \text{sign}(f_\theta(x))$.

We adopt a PAC-Bayesian approach (Alquier, 2024) and consider an exponentially weighted aggregate (EWA) procedure. Let's consider the following posterior distribution:

$$\hat{\rho}_\lambda(\theta) \propto \exp[-\lambda r_n^h(\theta)]\pi(\theta) \tag{1}$$

where $\lambda > 0$ is a tuning parameter that will be discussed later and $\pi(\theta)$ is a prior distribution, given in (2), that promotes (approximately) sparsity on the parameter vector $\theta$. In this paper, we primarily focus on the hinge loss resulting in the following hinge empirical risk:

$$r_n^h(\theta) = \frac{1}{n} \sum_{i=1}^{n} (1 - Y_i \, f_\theta(x_i))_+ \, .$$

It is noteworthy that several studies in deep learning for classification have employed hinge loss as a surrogate loss to facilitate computation, as in (Kim et al., 2021; Wang and Shang, 2022).

The EWA procedure has found application in various contexts in prior works (Dalalyan et al., 2018; Dalalyan and Tsybakov, 2012; Dalalyan, 2020; Dalalyan and Tsybakov, 2008). The term $\hat{\rho}_\lambda$ is also referred to as the Gibbs posterior (Alquier et al., 2016; Catoni, 2007). The incorporation of $\hat{\rho}_\lambda$ is driven by the minimization problem presented in Lemma 6, rather than strictly adhering to conventional Bayesian principles.

### 2.3.1 Sparsity prior

We adopt a spike-and-slab prior $\pi$ (Castillo et al., 2015; Chérief-Abdellatif, 2020) over the parameter space $\Theta_{S,L,D}$. The spike-and-slab prior is recognized as a relevant alternative to dropout in Bayesian deep learning, as discussed in Polson and Ročková (2018). More specifically, the prior is defined hierarchically as follows. Initially, a binary vector of indicators $\gamma = (\gamma_1, \ldots, \gamma_T)$ is sampled from the set $\mathcal{S}_T^S$ of $T$-dimensional binary vectors with precisely $S$ non-zero entries. Subsequently, given $\gamma_t$ for each $t = 1, \ldots, T$, we apply a spike-and-slab prior to $\theta_t$, which returns 0 if $\gamma_t = 0$ and a random sample from a uniform distribution on $[-C_B, C_B]$ otherwise:

$$\begin{cases} \gamma \sim \mathcal{U}(\mathcal{S}_T^S), \\ \theta_t | \gamma_t \sim \gamma_t \, \mathcal{U}([-C_B, C_B]) + (1 - \gamma_t)\delta_{\{0\}}, \quad t = 1, ..., T \end{cases} \tag{2}$$

where $\delta_{\{0\}}$ is a point mass at 0 and $\mathcal{U}([-C_B, C_B])$ is a uniform distribution on $[-C_B, C_B]$. We recall that the sparsity level $S$ is fixed here and that this assumption will be relaxed in Section 3.4.

**Remark 1.** *We opt for uniform distributions for simplicity, as done in related studies (Polson and Ročková, 2018; Suzuki, 2018). However, Gaussian distributions can also be employed when dealing with an unbounded parameter set $\Theta_{S,L,D}$, as indicated in Chérief-Abdellatif (2020).*

# 3 Main results

## 3.1 Assumptions

The result in this section applies to a wide range of activation functions, including the popular ReLU activation and the identity map. In the following, we make the following assumption regarding the activation function.

**Assumption 1.** *Assume that the activation function $\rho$ is 1-Lispchitz continuous with respect to the absolute value: $|\rho(x)| \leq |x|$ for any $x \in \mathbb{R}$.*

Put $R^* := \min_\theta \mathbb{P}(Y \neq \text{sign}(f_\theta(x)))$ ; $\theta^* := \arg\min_\theta \mathbb{P}(Y \neq \text{sign}(f_\theta(x)))$ ; $r_n(\theta) = n^{-1} \sum_{i=1}^n \mathbf{1}\{Y_i f_\theta(x_i) < 0\}$ , and $r_n^* := r_n(\theta^*)$.

**Assumption 2.** *We assume that there is a constant $C' > 0$ such that $r_n^h(\theta^*) \leq (1 + C')r_n^*$.*

Assumption 2 implies that for the underlying parameter $\theta^*$, the function $f_{\theta^*}(x)$ is bounded above by a constant.

**Assumption 3** (Low-noise assumption). *We assume that there is a constant $C \geq 1$ such that:* $\mathbb{E}\left[\left(\mathbf{1}_{Yf_\theta(x)\leq 0} - \mathbf{1}_{Yf_{\theta^*}(x)\leq 0}\right)^2\right] \leq C[R(\theta) - R^*]$.

Assumption 3 serves as the noise condition, capturing the difference between a classifier's performance and random guessing. This assumption is also commonly referred to as the margin assumption, as discussed in Mammen and Tsybakov (1999), and further explored in Tsybakov (2004); Bartlett et al. (2006).

## 3.2 General results

Here, we present general results that do not require $f_\theta$ to be $\beta$-Hölder and we consider any structure $(S, L, D)$.

**Theorem 1.** *Given Assumption 1 and Assumption 2, for $\lambda = \sqrt{n}$, we find that with a probability of at least $1 - 2\epsilon$, where $\epsilon \in (0, 1)$, the following holds:*

$$\int R d\hat{\rho}_\lambda \leq (1 + 2C')R^* + c\frac{S\log\left(\frac{nTD^L[(d+1)L+1]}{S(D-1)}\right) + \log(1/\epsilon)}{\sqrt{n}},$$

*where $c$ depends only ong $C_B, C'$.*

In addition to the results in Theorem 1 for the integrated error, we can derive a result for, $\hat{\theta} \sim \hat{\rho}_\lambda$, a stochastic classifier sampled from the Gibbs-posterior (1).

**Theorem 2.** *Under the same assumptions for Theorem 1, and the same definition for $\lambda$, for any small $\varepsilon \in (0, 1)$. one has that $\mathbb{E}\left[\mathbb{P}_{\hat{\theta}\sim\hat{\rho}_\lambda}(\hat{\theta} \in \mathcal{B})\right] \geq 1 - \varepsilon$, where*

$$\mathcal{B} = \left\{\theta \in \Theta_{S,L,D} : R \leq (1 + 2C')R^* + c\frac{S\log\left(\frac{TnD^L[(d+1)L+1]}{S(D-1)}\right)+\log(2/\epsilon)}{\sqrt{n}}\right\}.$$

The oracle inequality presented in Theorem 1 links the integrated prediction risk of our method to the minimum attainable risk. By incorporating additional assumptions, the bounds provided in Theorem 1 can be improved.

**Theorem 3.** *Suppose Assumption 1, Assumption 2 and Assumption 3 are satisfied. For $\lambda = 2n/(3C+2)$, we find that with a probability of at least $1-2\epsilon$, where $\epsilon \in (0, 1)$, the following holds:*

$$\int R d\hat{\rho}_\lambda \leq (1 + 3C')R^* + c'\frac{S\log\left(\frac{TnD^L[(d+1)L+1]}{S(D-1)}\right) + \log(1/\epsilon)}{n}.$$

*where $c'$ is a universal constant depending only on $C, C', C_B$.*

In contrast to Theorem 1, the bound in Theorem 3 exhibits a faster rate, scaling as $1/n$ rather than $1/\sqrt{n}$. These bounds enable a comparison between the out-of-sample error of our method and the optimal error $R^*$.

For the noiseless case, i.e. $Y = \text{sign}(f_{\theta^*}(x))$ almost surely, then $R^* = 0$ and consequently that $\mathbb{E}\left[\left(\mathbf{1}_{Yf_\theta(x)\leq 0} - \mathbf{1}_{Yf_{\theta^*}(x)\leq 0}\right)^2\right] = \mathbb{E}\left[\mathbf{1}^2_{Yf_\theta(x)\leq 0}\right] = \mathbb{E}\left[\mathbf{1}_{Yf_\theta(x)\leq 0}\right] = R(\theta) - R^*$. Thus, Assumption 3 is satisfied with $C = 1$.

**Corollary 4.** *In the noiseless case, for $\lambda = 2n/5$, with probability at least $1 - 2\epsilon, \epsilon \in (0,1)$, Theorem 3 leads to*

$$\int R d\hat{\rho}_\lambda \leq c' \frac{S \log\left(\frac{TnD^L[(d+1)L+1]}{S(D-1)}\right) + \log(1/\epsilon)}{n}.$$

*where $c'$ is a universal constant depending only on $C', C_B$.*

In analogy to Theorem 2, we can establish that a stochastic classifier, $\hat{\beta} \sim \hat{\rho}_\lambda$, drawn from our proposed pseudo-posterior in Equation (1) exhibits a fast rate.

**Theorem 5.** *Under the same assumptions for Theorem 3, and the same definition for $\lambda$, along with any small $\varepsilon \in (0,1)$, we find that $\mathbb{E}\left[\mathbb{P}_{\hat{\theta}\sim\hat{\rho}_\lambda}(\hat{\theta} \in \Omega)\right] \geq 1 - \varepsilon$, where*

$$\Omega_n = \left\{\theta \in \Theta_{S,L,D} : R \leq (1+3C')R^* + c' \frac{S \log\left(\frac{TnD^L[(d+1)L+1]}{S(D-1)}\right) + \log(2/\epsilon)}{n}\right\}.$$

In this section, we establish specific values for the tuning parameters $\lambda$ in our proposed method within theoretical results for prediction errors. However, it is noted that these values may not necessarily be optimal for practical applications. Cross-validation can be employed in practical scenarios to fine-tune the parameters appropriately. Nevertheless, the theoretical values identified in our analysis offer insights into the scale of the tuning parameters when utilized in real-world situations.

## 3.3 Rates in specific settings

### Low-dimension setup

Now, let's explore the scenario of low-dimensional settings, where we assume that $d$ is fixed and smaller than $n$. We will derive prediction error rates for the noiseless case, considering the following architecture.

**Example 1.** *Let us assume that $f_{\theta^*}$ is $\beta$-Hölder smooth with $0 < \beta < d$ and that the activation function is ReLU. We consider the architecture of Polson and Ročková (2018), for some positive constant $C_D$ independent of $n$, that: $L = 8 + (\lfloor \log_2 n \rfloor + 5)(1 + \lceil \log_2 d \rceil)$, $D = C_D \lfloor n^{\frac{d}{2\beta+d}} / \log n \rfloor$, $S \leq 94 d^2 (\beta+1)^{2d} D(L + \lceil \log_2 d \rceil)$.*

**Proposition 6.** *For the architecture as in Example 1, then the rate in Corollary 4 is of order $n^{\frac{-2\beta}{2\beta+d}} \log n$.*

**Remark 2.** *The rate provided in Proposition 6, adjusted by a $\log n$ factor, is minimax optimal Tsybakov (2004). Additionally, it is noteworthy that our $\log n$ factor represents a slight enhancement over the $\log^3 n$ factor achieved by the authors in Kim et al. (2021) for obtaining the minimax optimal rate.*

### High-dimensional setup

Next, we delve into the realm of high-dimensional settings, where we assume that $d > n$. We will consider the following architecture.

**Example 2.** *Let us assume that $f_{\theta^*}$ is $\beta$-Hölder smooth with $0 < \beta < d$ and that the activation function is ReLU. We consider the architecture of Wang and Shang (2022): $L \asymp \log n, D \asymp d, S \asymp n^{\frac{d}{2\beta+d}} \log n$.*

**Proposition 7.** *For the architecture as in Example 2, then the rate in Corollary 4 is of order* $n^{\frac{-2\beta}{2\beta+d}} \log n \log d$.

**Remark 3.** *The rate,* $n^{\frac{-2\beta}{2\beta+d}} \log d$*, provided in Proposition 7, up to a* $\log n$ *factor, is minimax optimal in a high dimensional setting as in Wang and Shang (2022). Additionally, it is noteworthy that our* $\log n$ *factor represents a slight enhancement over the* $\log^2 n \log d$ *factor achieved by the authors in Wang and Shang (2022) for obtaining the minimax optimal rate.*

### 3.4 Architecture design via model selection

Let $\mathcal{M}_{S,L,D}$ denote the model associated with the parameter set $\Theta_{S,L,D}$. We consider a countable number of models, and introduce prior beliefs $\mathrm{p}_{S,L,D}$ over the sparsity, the depth and the width of the network, hierarchically.

More specifically, $\mathrm{p}_{S,L,D} = \mathrm{p}_{S|L,D}\mathrm{p}_{D|L}\mathrm{p}_L$ where $\mathrm{p}_L = 2^{-L}$, $\mathrm{p}_{D|L}$ follows a uniform distribution over $\{d, ..., \max(e^L, d)\}$ given $L$, and $\mathrm{p}_{S|L,D}$ is a uniform distribution over $\{1, ..., T\}$ given $L$ and $D$. Here, $T$ represents the number of coefficients in a fully connected network. This particular choice is sensible as it allows to consider any number of hidden layers and (at most) an exponentially large width with respect to the depth of the network. We continue to employ spike-and-slab priors in (2) on $\theta_{S,L,D} \in \Theta_{S,L,D}$ given model $\mathcal{M}_{S,L,D}$, now denoted as $\pi_{(S,L,D)}(\theta)$.

The Gibbs posterior is now as $\hat{\rho}_\lambda^{(S,L,D)}(\theta) \propto \exp[-\lambda r_n^h(\theta)]\pi_{(S,L,D)}(\theta)$. We formulate the following optimization problem to select the architecture of the network.

$$(\hat{S}, \hat{L}, \hat{D}) = \arg\min_{S,L,D} \left[ \int r_n^h \mathrm{d}\hat{\rho}_\lambda^{(S,L,D)} + \frac{\mathrm{KL}(\hat{\rho}_\lambda^{(S,L,D)}, \pi_{(S,L,D)}) + \log(1/\mathrm{p}_{S,L,D})}{\lambda} \right] \quad (3)$$

The following theorem demonstrates that this strategy for selecting the architecture yields identical results to those in our main theorem, Theorem 3.

**Theorem 8.** *Suppose Assumption 1, Assumption 2 and Assumption 3 are satisfied. For* $\lambda = 2n/(3C+2)$*, we find that with a probability of at least* $1-2\epsilon$*, where* $\epsilon \in (0,1)$*, the following holds:*

$$\int R d\hat{\rho}_\lambda^{(\hat{S},\hat{L},\hat{D})} \leq (1+3C')R^* + \inf_{S,L,D} \left[ c' \frac{S \log\left(\frac{TnD^L[(d+1)L+1]}{S(D-1)}\right) + \log(\frac{1}{\mathrm{p}_{S,L,D}}) + \log(\frac{1}{\epsilon})}{n} \right].$$

*where* $c'$ *is a universal constant depending only on* $C, C', C_B$.

Theorem 8 illustrates that once the complexity term $\log(1/\mathrm{p}_{S,L,D})/n$, which reflects the prior beliefs, falls below the effective error $r_n^{S,L,D} := S\max\left(\log(L), L\log(D), \log(nd)\right)/n$, the selection method in (3) adaptively achieves the best possible rate. This scenario leads to (near-)minimax rates for Hölder smooth functions and selects the optimal architecture. Notably, for the prior beliefs choice $\pi_L = 2^{-L}$, $\pi_{D|L} = 1/(\max(e^L, d) - d + 1)$, $\pi_{S|L,D} = 1/T$, we obtain, up to some

absolute constant $K > 0$, that:

$$\frac{\log(\frac{1}{\mathrm{p}_{S,L,D}})}{n} \leq K \frac{\max(\log(D), \log L, L, \log(d))}{n}$$

which is lower than $r_n^{S,L,D}$ (up to a factor).

It is worth noting that various model selection approaches for choosing the architecture of the network have been proposed in the study of deep learning, such as Kim et al. (2021); Chérief-Abdellatif (2020).

# 4 Proof

## Proof for slow rate

**Proof of Theorem 1.**
*Step 1:*

Put $U_i = \mathbf{1}_{Y_i f_\theta(x_i) \leq 0} - \mathbf{1}_{Y_i f_{\theta^*}(x_i) \leq 0}$, then, $-1 \leq U_i \leq 1$. One can utilize Hoeffding's Lemma 5 to derive that

$$\mathbb{E} \exp\left\{\lambda[R(\theta) - R^*] - \lambda[r_n(\theta) - r_n^*]\right\} \leq \exp\left(\lambda^2/2n\right).$$

Integrating with respect to $\pi$ and then applying Fubini's theorem, one gets, for any $\lambda \in (0, n)$, that

$$\mathbb{E} \int \exp\left\{\lambda[R(\theta) - R^*] - \lambda[r_n(\theta) - r_n^*] - \frac{\lambda^2}{2n}\right\} \mathrm{d}\pi(\theta) \leq 1, \tag{4}$$

Consequently, using Lemma 6,

$$\mathbb{E} \exp\left\{\sup_\rho \int \left\{\lambda[R(\theta) - R^*] - \lambda[r_n(\theta) - r_n^*] - \frac{\lambda^2}{2n}\right\} \rho(d\theta) - \mathrm{KL}(\rho, \pi)\right\} \leq 1.$$

Applying Markov's inequality, for $\epsilon \in (0, 1)$, we obtain that:

$$\mathbb{P}\left\{\sup_\rho \int \left[\lambda[R(\theta) - R^*] - \lambda[r_n(\theta) - r_n^*] - \frac{\lambda^2}{2n}\right] \rho(d\theta) - \mathrm{KL}(\rho, \pi) + \log \epsilon > 0\right\} \leq \epsilon.$$

Subsequently, considering the complement, we derive that with a probability of at least $1 - \epsilon$, the following holds:

$$\forall \rho, \quad \lambda \int [R(\theta) - R^*] \rho(d\theta) \leq \lambda \int [r_n(\theta) - r_n^*] \rho(d\theta) + \mathrm{KL}(\rho, \pi) + \frac{\lambda^2}{2n} + \log \frac{1}{\epsilon}.$$

Now, note that as $r_n^h \geq r_n$ and as it stands for all $\rho$ then the right hand side can be minimized and, from Lemma 6, the minimizer over $\mathcal{P}(\Theta_{S,L,D})$ is $\hat{\rho}_\lambda$. Thus we get,

when $\lambda > 0$,

$$\int R d\hat{\rho}_\lambda \leq R^* + \inf_{\rho \in \mathcal{P}(\Theta_{S,L,D})} \left[ \int r_n^h d\rho + \frac{1}{\lambda} \mathrm{KL}(\rho, \pi) \right] - r_n^* + \frac{\lambda}{2n} + \frac{1}{\lambda} \log \frac{1}{\epsilon}. \quad (5)$$

Step 2:
First, we have that,

$$\begin{aligned}
\int r_n^h(\theta) \rho(d\theta) &= \frac{1}{n} \int \sum_{i=1}^n (1 - Y_i f_\theta(x_i))_+ \, \rho(d\theta) \\
&\leq \frac{1}{n} \left[ \sum_{i=1}^n (1 - Y_i f_{\theta^*}(x_i))_+ + \int \sum_{i=1}^n |f_\theta(x_i) - f_{\theta^*}(x_i)| \, \rho(d\theta) \right] \\
&\leq r_n^h(\theta^*) + \frac{1}{n} \sum_{i=1}^n \int |f_\theta(x_i) - f_{\theta^*}(x_i)| \, \rho(d\theta). \quad (6)
\end{aligned}$$

And using Lemma 4,

$$\begin{aligned}
&\int |f_\theta(x_i) - f_{\theta^*}(x_i)| \, \rho(d\theta) \\
&\leq \int r_L(\theta) \rho(d\theta) \\
&\leq \int (C_B D)^{L-1} \frac{C_B D(d+1) - d}{C_B D - 1} \sum_{u=1}^L \tilde{A}_u \rho(d\theta) + \int \sum_{u=1}^L (C_B D)^{L-u} \tilde{b}_u \rho(d\theta) \\
&\leq (C_B D)^{L-1} \frac{C_B D(d+1) - d}{C_B D - 1} \sum_{u=1}^L \int \tilde{A}_u \rho(d\theta) + \sum_{u=1}^L (C_B D)^{L-u} \int \tilde{b}_u \rho(d\theta).
\end{aligned}$$

Here, we take $\rho(d\theta) = q_n^*(\theta)$, as defined in equation (11). We find that

$$\int \tilde{A}_\ell q_n^*(\mathrm{d}\theta) = \int \sup_{i,j} |A_{\ell,i,j} - A^*_{\ell,i,j}| q_n^*(\mathrm{d}A_{\ell,i,j}) \leq s_n,$$

and $\int \tilde{b}_u q_n^*(\mathrm{d}\theta) = \int \sup_j |b_{u,j} - b^*_{u,j}| q_n^*(\mathrm{d}b_{u,j}) \leq s_n$. Therefore,

$$\begin{aligned}
\int |f_\theta(x_i) - f_{\theta^*}(x_i)| \, q_n^*(\mathrm{d}\theta) &\leq (C_B D)^{L-1} \frac{C_B D(d+1) - d}{C_B D - 1} L s_n + s_n \sum_{u=1}^L (C_B D)^{L-u} \\
&\leq (C_B D)^{L-1} \frac{C_B D(d+1) - d}{C_B D - 1} L s_n + s_n \sum_{\ell=0}^{L-1} (C_B D)^\ell \\
&\leq (C_B D)^{L-1} \frac{C_B D(d+1) - d}{C_B D - 1} L s_n + s_n \frac{(C_B D)^L - 1}{C_B D - 1}
\end{aligned}$$

$$\leq s_n \left[ (C_B D)^L \frac{C_B D(d+1) - d}{(C_B D - 1) C_B D} L + \frac{(C_B D)^L}{C_B D - 1} \right]$$
$$\leq s_n \frac{(C_B D)^L}{C_B D - 1} \left[ (d+1)L + 1 \right]. \tag{7}$$

From Assumption 2, we have $r_n^h(\theta^*) \leq (1 + C') r_n^*$. Plug in (7) and (12), into (5), we obtain

$$\int R d\hat{\rho}_\lambda \leq R^* + C' r_n^* + s_n \frac{(C_B D)^L}{C_B D - 1} \left[ (d+1)L + 1 \right] + \frac{S \log(TB) + \frac{S}{2} \log\left(\frac{1}{s_n^2}\right)}{\lambda} + \frac{\lambda}{2n} + \frac{1}{\lambda} \log\left(\frac{1}{\epsilon}\right).$$

Then, we use Lemma 1, with probability at least $1 - 2\epsilon$, to obtain that

$$\int R d\hat{\rho}_\lambda \leq (1 + 2C') R^* + C' \frac{1}{n\varsigma} \log \frac{1}{\epsilon} + s_n \frac{(C_B D)^L}{C_B D - 1} \left[ (d+1)L + 1 \right] + \frac{S \log(TC_B) + S \log\left(\frac{1}{s_n}\right)}{\lambda} + \frac{\lambda}{2n} + \frac{1}{\lambda} \log\left(\frac{1}{\epsilon}\right). \tag{8}$$

By taking $s_n = \frac{S}{n} \left[ \frac{(C_B D)^L}{C_B D - 1} \left[ (d+1)L + 1 \right] \right]^{-1}$ and $\lambda = \sqrt{n}$, we can obtain that

$$\int R d\hat{\rho}_\lambda \leq (1 + 2C') R^* + C' \frac{1}{n\varsigma} \log \frac{1}{\epsilon} + \frac{S}{n} + \frac{S \log\left( T C_B \frac{n}{S} \frac{(C_B D)^L}{C_B D - 1} \left[ (d+1)L + 1 \right] \right)}{\sqrt{n}} + \frac{\log(1/\epsilon)}{\sqrt{n}}.$$

The proof is completed. □

***Proof of Theorem 2.*** From (4), we have that

$$\mathbb{E}\left[ \int \exp\left\{ \lambda[R(\theta) - R^*] - \lambda[r_n(\theta) - r_n^*] - \log\left[ \frac{d\hat{\rho}_\lambda}{d\pi}(\theta) \right] - \frac{\lambda^2}{2n} - \log\frac{1}{\varepsilon} \right\} \hat{\rho}_\lambda(d\theta) \right] \leq \varepsilon.$$

We employ the Chernoff's trick, which is based on $\exp(x) \geq 1_{\mathbb{R}_+}(x)$, leading to $\mathbb{E}\left[ \mathbb{P}_{\theta \sim \hat{\rho}_\lambda}(\theta \in \mathcal{B}) \right] \geq 1 - \varepsilon$, where

$$\mathcal{B} = \left\{ \theta : \lambda[R(\theta) - R^*] - \lambda[r_n(\theta) - r_n^*] \leq \log\left[ \frac{d\hat{\rho}_\lambda}{d\pi}(\theta) \right] + \frac{\lambda^2}{2n} + \log\frac{2}{\varepsilon} \right\}.$$

Utilizing the definition of $\hat{\rho}_\lambda$ and Lemma 6, and recognizing that $r_n \le r_n^h$, for $\theta \in \mathcal{B}$, we find that:

$$\begin{aligned}
\lambda[R(\theta) - R^*] &\le \lambda\Big(r(\theta) - r_n^*\Big) + \log\left[\frac{d\hat{\rho}_\lambda}{d\pi}(\theta)\right] + \frac{\lambda^2}{2n} + \log\frac{2}{\varepsilon} \\
&\le \lambda\Big(r_n^h(\theta) - r_n^*\Big) + \log\left[\frac{d\hat{\rho}_\lambda}{d\pi}(\theta)\right] + \frac{\lambda^2}{2n} + \log\frac{2}{\varepsilon} \\
&\le -\log\int \exp\left[-\lambda r_n^h(\theta)\right]\pi(\mathrm{d}\theta) - \lambda r_n^* + \frac{\lambda^2}{2n} + \log\frac{2}{\varepsilon} \\
&= \lambda\Big(\int r_n^h(\theta)\hat{\rho}_\lambda(\mathrm{d}\theta) - r_n^*\Big) + \mathcal{K}(\hat{\rho}_\lambda, \pi) + \frac{\lambda^2}{2n} + \log\frac{2}{\varepsilon} \\
&= \inf_\rho\left\{\lambda\Big(\int r_n^h(\theta)\rho(\mathrm{d}\theta) - r_n^*\Big) + \mathcal{K}(\rho, \pi) + \frac{\lambda^2}{2n} + \log\frac{2}{\varepsilon}\right\}.
\end{aligned}$$

We upper-bound the right-hand side exactly as Step 2 in the proof of Theorem 1. The result of the theorem is followed. □

## Proof for fast rate

**Proof of Theorem 3.**
*Step 1:*

Fix any $\theta$ and put $U_i = \mathbf{1}_{Y_i f_\theta(x_i)\le 0} - \mathbf{1}_{Y_i f_{\theta^*}(x_i)\le 0}$. Under Assumption 3, we have that $\sum_i \mathbb{E}[U_i^2] \le nC[R(\theta) - R^*]$. Now, for any integer $k \ge 3$, as the 0-1 loss is bounded, it results in

$$\sum_i \mathbb{E}\left[(U_i)_+^k\right] \le \sum_i \mathbb{E}\left[|U_i|^{k-2}|U_i|^2\right] \le \sum_i \mathbb{E}\left[|U_i|^2\right].$$

Thus, we can apply Lemma 2 with $v := nC[R(\theta) - R^*]$, $w := 1$ and $\zeta := \lambda/n$. We obtain, for any $\lambda \in (0, n)$,

$$\mathbb{E}\exp\{\lambda([R(\theta) - R^*] - [r_n(\theta) - r_n^*])\} \le \exp\left\{\frac{C\lambda^2[R(\theta) - R^*]}{2n(1 - \lambda/n)}\right\},$$

By integrating with respect to $\pi$ and subsequently applying Fubini's theorem, we find that

$$\mathbb{E}\left[\int \exp\left\{(\lambda - \frac{C\lambda^2}{2n(1-\lambda/n)})[R(\theta) - R^*] - \lambda[r_n(\theta) - r_n^*]\right\}\pi(d\theta)\right] \le 1. \tag{9}$$

Consequently, using Lemma 6,

$$\mathbb{E}\left[e^{\sup_\rho \int\left\{(\lambda - \frac{C\lambda^2}{2n(1-\lambda/n)})[R(\theta)-R^*] - \lambda[r_n(\theta)-r_n^*]\right\}\rho(dM) - \mathrm{KL}(\rho,\pi)}\right] \le 1.$$

By applying Markov's inequality and subsequently considering the complement, we establish that with a probability of at least $1-\epsilon$, the following holds:

$$\forall \rho,\ \left(\lambda - \frac{C\lambda^2}{2n(1-\lambda/n)}\right)\int [R(\theta)-R^*]\rho(d\theta) \leq \lambda \int [r_n(\theta)-r_n^*]\rho(d\theta) + \mathrm{KL}(\rho,\pi) + \log\frac{1}{\epsilon}.$$

Now, observe that $r_n^h \geq r_n$,

$$\lambda\left[\int r_n d\rho - r_n^*\right] + \mathrm{KL}(\rho,\pi) \leq \lambda\left[\int r_n^h d\rho + \frac{1}{\lambda}\mathrm{KL}(\rho,\pi)\right] - \lambda r_n^*.$$

As it stands for all $\rho$ then the right hand side can be minimized and, from Lemma 6, the minimizer over $\mathcal{P}(\Theta_{S,L,D})$ is $\hat{\rho}_\lambda$. Thus we get, when $\lambda < 2n/(C+2)$,

$$\int R d\hat{\rho}_\lambda \leq R^* + \frac{1}{1-\frac{C\lambda}{2n(1-\lambda/n)}}\left\{\inf_{\rho\in\mathcal{P}(\Theta_{S,L,D})}\left[\int r_n^h d\rho + \frac{\mathrm{KL}(\rho,\pi)+\log\frac{1}{\epsilon}}{\lambda}\right] - r_n^*\right\}. \tag{10}$$

*Step 2:*
From (6), we have that,

$$\int r_n^h(\theta)\rho(d\theta) \leq r_n^h(\theta^*) + \frac{1}{n}\sum_{i=1}^n \int |f_\theta(x_i) - f_{\theta^*}(x_i)|\,\rho(d\theta).$$

Here, we take $\rho(\theta) = q_n^*(\theta)$, as defined in equation (11). Then from (7), and (12), we deduce (10) into that, from Assumption 2, as $r_n^h(\theta^*) \leq (1+C')r_n^*$, we have that

$$\int R d\hat{\rho}_\lambda \leq R^* + \frac{1}{1-\frac{C\lambda}{2n(1-\lambda/n)}}\left\{C' r_n^* + s_n\frac{(C_B D)^L}{C_B D - 1}\left[(d+1)L+1\right] + \frac{S\log(TC_B) + \frac{S}{2}\log\left(\frac{1}{s_n^2}\right) + \log\left(\frac{1}{\epsilon}\right)}{\lambda}\right\}.$$

Taking $\lambda = 2n/(3C+2)$, we obtain:

$$\int R d\hat{\rho}_\lambda \leq R^* + \frac{3}{2}\left\{C' r_n^* + s_n\frac{(C_B D)^L}{C_B D - 1}\left[(d+1)L+1\right] + (3C+2)\frac{S\log(TC_B) + \frac{S}{2}\log\left(\frac{1}{s_n^2}\right) + \log\left(\frac{1}{\epsilon}\right)}{2n}\right\}.$$

Then, we use Lemma 1, with probability at least $1-2\epsilon$, to obtain that

$$\int R d\hat{\rho}_\lambda \leq (1+3C')R^* + \frac{3}{2}\left\{C'\frac{1}{n\varsigma}\log\frac{1}{\epsilon} + s_n\frac{(C_BD)^L}{C_BD-1}\left[(d+1)L+1\right] \right.$$
$$\left. +(3C+2)\frac{S\log(TC_B) + S\log\left(\frac{1}{s_n}\right) + \log(1/\epsilon)}{2n}\right\}.$$

By setting $s_n = \frac{S}{n}\left[\frac{(C_BD)^L}{C_BD-1}\left[(d+1)L+1\right]\right]^{-1}$, one finds that

$$\int R d\hat{\rho}_\lambda \leq (1+3C')R^* + \frac{3}{2}\left\{C'\frac{1}{n\varsigma}\log\frac{1}{\epsilon} + \frac{S}{n}\right.$$
$$\left. +(3C+2)\frac{S\log\left(TC_B\frac{n}{S}\frac{(C_BD)^L}{C_BD-1}\left[(d+1)L+1\right]\right) + \log(1/\epsilon)}{2n}\right\}.$$

The proof is completed. □

***Proof of Theorem 5.*** From (9), we have that

$$\mathbb{E}\left[\int \exp\left\{(\lambda - \frac{C\lambda^2}{2n(1-\lambda/n)})[R(\theta)-R^*] - \lambda[r_n(\theta)-r_n^*] - \log\left[\frac{d\hat{\rho}_\lambda}{d\pi}(\theta)\right] - \log\frac{1}{\varepsilon}\right\}\hat{\rho}_\lambda(d\theta)\right] \leq \varepsilon.$$

Using Chernoff's trick, i.e. $\exp(x) \geq 1_{\mathbb{R}_+}(x)$, this gives: $\mathbb{E}\Big[\mathbb{P}_{\theta\sim\hat{\rho}_\lambda}(\theta\in\Omega)\Big] \geq 1-\varepsilon$ where

$$\Omega = \left\{\theta : (\lambda - \frac{C\lambda^2}{2n(1-\lambda/n)})[R(\theta)-R^*] - \lambda[r_n(\theta)-r_n^*] \leq \log\left[\frac{d\hat{\rho}_\lambda}{d\pi}(\theta)\right] + \log\frac{2}{\varepsilon}\right\}.$$

Using the definition of $\hat{\rho}_\lambda$ and noting that $r_n \leq r_n^h$, for $\theta\in\Omega$ we have

$$\begin{aligned}
\Big(\lambda - \frac{C\lambda^2}{2n(1-\lambda/n)}\Big)\Big[R(\theta)-R^*\Big] &\leq \lambda\Big(r_n(\theta)-r_n^*\Big) + \log\left[\frac{d\hat{\rho}_\lambda}{d\pi}(\theta)\right] + \log\frac{2}{\varepsilon} \\
&\leq \lambda\Big(r_n^h(\theta)-r_n^*\Big) + \log\left[\frac{d\hat{\rho}_\lambda}{d\pi}(\theta)\right] + \log\frac{2}{\varepsilon} \\
&\leq -\log\int \exp\left[-\lambda r_n^h(\theta)\right]\pi(\mathrm{d}\theta) - \lambda r_n^* + \log\frac{2}{\varepsilon} \\
&= \lambda\Big(\int r_n^h(\theta)\hat{\rho}_\lambda(\mathrm{d}\theta) - r_n^*\Big) + \mathcal{K}(\hat{\rho}_\lambda,\pi) + \log\frac{2}{\varepsilon} \\
&= \inf_\rho\left\{\lambda\Big(\int r_n^h(\theta)\rho(\mathrm{d}\theta) - r_n^*\Big) + \mathcal{K}(\rho,\pi) + \log\frac{2}{\varepsilon}\right\}.
\end{aligned}$$

We upper-bound the right-hand side exactly as Step 2 in the proof of Theorem 3. The result of the theorem is followed. □

*Proof of Proposition 6.* For certain constants $K > 0$, independent of $n$, we have that

$$\begin{aligned}\frac{S\log\left(\frac{TnD^L[(d+1)L+1]}{S(D-1)}\right)}{n} &\leq K\frac{S\log\left(\frac{LD(D+1)nD^L[(d+1)L+1]}{S(D-1)}\right)}{n}\\ &\leq K\frac{S\max\left(\log(L), L\log(D), \log n\right)}{n}\\ &\leq K\frac{n^{\frac{d}{2\beta+d}}}{n}\log n = Kn^{\frac{-2\beta}{2\beta+d}}\log n.\end{aligned}$$

□

*Proof of Proposition 7.* For certain constants $K > 0$, independent of $n, d$, we have that

$$\begin{aligned}\frac{S\log\left(\frac{TnD^L[(d+1)L+1]}{S(D-1)}\right)}{n} &\leq K\frac{S\log\left(\frac{LD(D+1)nD^L[(d+1)L+1]}{S(D-1)}\right)}{n}\\ &\leq K\frac{S\max\left(\log(L), L\log(D), \log(nd)\right)}{n}\\ &\leq K\frac{n^{\frac{d}{2\beta+d}}\log n\log d}{n} = Kn^{\frac{-2\beta}{2\beta+d}}\log n\log d.\end{aligned}$$

□

## Proof of Section 3.4

*Proof of Theorem 8.* From (10), we have that

$$\int R d\hat{\rho}_\lambda^{(S,L,D)} \leq R^* + \frac{1}{1-\frac{C\lambda}{2n(1-\lambda/n)}}\left\{\inf_\rho\left[\int r_n^h d\rho + \frac{\mathrm{KL}(\rho,\pi_{(S,L,D)}) + \log\frac{1}{\epsilon}}{\lambda}\right] - r_n^*\right\}.$$

From Lemma 6, it holds that $\hat{\rho}_\lambda^{(S,L,D)}$ is the minimizer of $\int r_n^h d\rho + \frac{\mathrm{KL}(\rho,\pi_{(S,L,D)})}{\lambda}$, thus

$$\begin{aligned}\int R d\hat{\rho}_\lambda^{(S,L,D)} \leq R^* &+ \frac{1}{1-\frac{C\lambda}{2n(1-\lambda/n)}}\times\\ &\left\{\int r_n^h d\hat{\rho}_\lambda^{(S,L,D)} + \frac{\mathrm{KL}(\hat{\rho}_\lambda^{(S,L,D)},\pi_{(S,L,D)}) + \log(\frac{1}{\mathrm{p}_{S,L,D}}) + \log\frac{1}{\epsilon}}{\lambda} - r_n^*\right\}.\end{aligned}$$

Now using the definition of $(\hat{S},\hat{L},\hat{D})$ in (3), we arrive at

$$\int R d\hat{\rho}_\lambda^{(\hat{S},\hat{L},\hat{D})} \leq R^* + \frac{1}{1-\frac{C\lambda}{2n(1-\lambda/n)}}\times$$

$$\left\{ \inf_{S,L,D} \left[ \int r_n^h d\hat{\rho}_\lambda^{(S,L,D)} + \frac{\mathrm{KL}(\hat{\rho}_\lambda^{(S,L,D)}, \pi_{(S,L,D)}) + \log(\frac{1}{\mathrm{p}_{S,L,D}}) + \log\frac{1}{\epsilon}}{\lambda} \right] - r_n^* \right\},$$

using Lemma 6, we deduce that

$$\int R d\hat{\rho}_\lambda^{(\hat{S},\hat{L},\hat{D})} \leq R^* + \frac{1}{1 - \frac{C\lambda}{2n(1-\lambda/n)}} \times \left\{ \inf_{S,L,D} \inf_{\rho} \left[ \int r_n^h d\rho + \frac{\mathrm{KL}(\rho, \pi_{(S,L,D)}) + \log(\frac{1}{\mathrm{p}_{S,L,D}}) + \log\frac{1}{\epsilon}}{\lambda} \right] - r_n^* \right\}.$$

To obtain the result, proceed as in Step 2 in the proof of Theorem 3, page 13. This completes the proof. □

## Auxiliary lemmas

**Lemma 1.** *[Lemma 6 in Cottet and Alquier (2018)] For $\epsilon \in (0,1)$, with probability at least $1-\epsilon$, we have, for every $\varsigma \in (0,1)$, that $r_n^* \leq (1+\varsigma)R^* + \frac{1}{n\varsigma}\log\frac{1}{\epsilon}$, or we can have that $r_n^* \leq 2R^* + \frac{1}{n\varsigma}\log\frac{1}{\epsilon}$.*

We will use the following version of the Bernstein's lemma, from (Massart, 2007, page 24).

**Lemma 2.** *Let $U_1, \ldots, U_n$ be independent real valued random variables. Assuming that there exist two constants $v$ and $w$ such that $\sum_{i=1}^n \mathbb{E}[U_i^2] \leq v$ and that for all integers $k \geq 3$, $\sum_{i=1}^n \mathbb{E}\left[(U_i)_+^k\right] \leq vk!w^{k-2}/2$. Then, for any $\zeta \in (0, 1/w)$,*

$$\mathbb{E}\exp\left[\zeta \sum_{i=1}^n [U_i - \mathbb{E}U_i]\right] \leq \exp\left(\frac{v\zeta^2}{2(1-w\zeta)}\right).$$

**Definition 1.** *Put $q_n^*$ as*

$$\begin{cases} \gamma_t^* = \mathbb{I}(\theta_t^* \neq 0), \\ \theta_t \sim \gamma_t^* \, \mathcal{U}([\theta_t^* - s_n, \theta_t^* + s_n]) + (1-\gamma_t^*)\delta_{\{0\}}, \quad \text{for each } t = 1, ..., T. \end{cases} \tag{11}$$

**Lemma 3.** *We have, for $q_n^*$ in (11), that*

$$\mathrm{KL}(q_n^* \| \pi) \leq S\log(TC_B) + \frac{S}{2}\log\left(\frac{1}{s_n^2}\right). \tag{12}$$

*Proof.* The proof of this Lemma can be found in the proof of Theorem 2 in Chérief-Abdellatif (2020), in particular in the third step of the proof. □

**Lemma 4.** *Define the loss of the output layer as $r_\ell(\theta) := \sup_{x\in[-1,1]^d} |f_\theta^L(x) - f_{\theta^*}^L(x)| = \sup_{x\in[-1,1]^d} |f_\theta(x) - f_{\theta^*}(x)|$. We have, under Assumption 1, that for any $\ell = 1, ..., L$:*

$r_\ell(\theta) \leq (C_B D)^{\ell-1} \frac{C_B D(d+1)-d}{C_B D-1} \sum_{u=1}^{\ell} \tilde{A}_u + \sum_{u=1}^{\ell} (C_B D)^{\ell-u} \tilde{b}_u$,
*where* $\tilde{A}_u = \sup_{i,j} |A_{u,i,j} - A^*_{u,i,j}|$ *and* $\tilde{b}_u = \sup_j |b_{u,j} - b^*_{u,j}|$.

*Proof.* The proof of this Lemma can be found in the proof of Theorem 2 in Chérief-Abdellatif (2020), in particular in the second step of the proof. □

We remind here a version of Hoeffding's inequality for bounded random variables.

**Lemma 5.** *Let* $U_i, i = 1, \ldots, n$ *be* $n$ *independent random variables with* $a \leq U_i \leq b$ *a.s., and* $\mathbb{E}(U_i) = 0$. *Then, for any* $\lambda > 0$, $\mathbb{E} \exp\left(\frac{\lambda}{n} \sum_{i=1}^{n} U_i\right) \leq \exp\left(\frac{\lambda^2 (b-a)^2}{8n}\right)$.

**Lemma 6** (Catoni (2007)Lemma 1.1.3)**.** *Let* $\mu \in \mathcal{P}(\Theta)$. *For any measurable, bounded function* $h : \Theta \to \mathbb{R}$ *we have:* $\log \int \mathrm{e}^{h(\theta)} \mu(\mathrm{d}\theta) = \sup_{\rho \in \mathcal{P}(\Theta)} \left[ \int h(\theta) \rho(\mathrm{d}\theta) - \mathrm{KL}(\rho, \mu) \right]$. *Moreover, the supremum w.r.t* $\rho$ *in the right-hand side is reached for the Gibbs distribution,* $\hat{\rho}(d\theta) \propto \exp(h(\theta))\pi(d\theta)$.

**Acknowledgements.** The author was supported by the Norwegian Research Council, grant number 309960, through the Centre for Geophysical Forecasting at NTNU.

**Conflict of interest/Competing interests.** The author declares no potential conflict of interests.

# References

Alquier, P.: User-friendly introduction to PAC-Bayes bounds. Foundations and Trends® in Machine Learning **17**(2), 174–303 (2024)

Alquier, P., Ridgway, J., Chopin, N.: On the properties of variational approximations of Gibbs posteriors. The Journal of Machine Learning Research **17**(1), 8374–8414 (2016)

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

Bartlett, P.L., Foster, D.J., Telgarsky, M.J.: Spectrally-normalized margin bounds for neural networks. Advances in neural information processing systems **30** (2017)

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

Bhattacharya, S., Liu, Z., Maiti, T.: Comprehensive study of variational bayes classification for dense deep neural networks. Statistics and Computing **34**(1), 17 (2024)

Bai, J., Song, Q., Cheng, G.: Efficient variational inference for sparse deep learning with theoretical guarantee. Advances in Neural Information Processing Systems **33**, 466–476 (2020)

Bos, T., Schmidt-Hieber, J.: Convergence rates of deep relu networks for multiclass classification. Electronic Journal of Statistics **16**(1), 2724–2773 (2022)

Cottet, V., Alquier, P.: 1-Bit matrix completion: PAC-Bayesian analysis of a variational approximation. Machine Learning **107**(3), 579–603 (2018)

Chérief-Abdellatif, B.-E.: Convergence rates of variational inference in sparse deep learning. In: III, H.D., Singh, A. (eds.) Proceedings of the 37th International Conference on Machine Learning, vol. 119, pp. 1831–1842 (2020). PMLR

Catoni, O.: Statistical Learning Theory and Stochastic Optimization. Saint-Flour Summer School on Probability Theory 2001 (Jean Picard ed.), Lecture Notes in Mathematics, vol. 1851, p. 272. Springer, Berlin (2004). https://doi.org/10.1007/b99352

Catoni, O.: PAC-Bayesian Supervised Classification: the Thermodynamics of Statistical Learning. IMS Lecture Notes—Monograph Series, 56, p. 163. Institute of Mathematical Statistics, Beachwood, OH (2007)

Castillo, I., Schmidt-Hieber, J., Der Vaart, A.: Bayesian linear regression with sparse priors. Annals of Statistics **43**(5), 1986–2018 (2015)

Dalalyan, A.S.: Exponential weights in multivariate regression and a low-rankness favoring prior. Annales de l'Institut Henri Poincaré, Probabilités et Statistiques **56**(2), 1465–1483 (2020)

Devroye, L., Györfi, L., Lugosi, G.: A Probabilistic Theory of Pattern Recognition vol. 31. Springer, ??? (1996)

Dalalyan, A.S., Grappin, E., Paris, Q.: On the exponentially weighted aggregate with the laplace prior. The Annals of Statistics **46**(5), 2452–2478 (2018)

Dalalyan, A., Tsybakov, A.B.: Aggregation by exponential weighting, sharp PAC-Bayesian bounds and sparsity. Machine Learning **72**(1-2), 39–61 (2008)

Dalalyan, A.S., Tsybakov, A.B.: Sparse regression learning by aggregation and langevin monte-carlo. Journal of Computer and System Sciences **78**(5), 1423–1443 (2012)

Goodfellow, I., Bengio, Y., Courville, A.: Deep Learning. MIT press, ??? (2016)

Guedj, B.: A primer on PAC-Bayesian learning. In: SMF 2018: Congrès de la Société Mathématique de France. Sémin. Congr., vol. 33, pp. 391–413. Soc. Math. France, ??? (2019)

Hu, T., Liu, R., Shang, Z., Cheng, G.: Minimax optimal deep neural network classifiers under smooth decision boundary. arXiv preprint arXiv:2207.01602 (2022)

Hayakawa, S., Suzuki, T.: On the minimax optimality and superiority of deep neural network learning over sparse parameter spaces. Neural Networks **123**, 343–361 (2020)

Imaizumi, M., Fukumizu, K.: Deep neural networks learn non-smooth functions effectively. In: The 22nd International Conference on Artificial Intelligence and Statistics, pp. 869–878 (2019). PMLR

Kohler, M., Krzyżak, A., Walter, B.: On the rate of convergence of image classifiers based on convolutional neural networks. Annals of the Institute of Statistical Mathematics **74**(6), 1085–1108 (2022)

Kohler, M., Langer, S.: Statistical theory for image classification using deep convolutional neural networks with cross-entropy loss. Journal of Statistical Planning and Inference ( to appear), 2011–13602 (2024)

Kim, Y., Ohn, I., Kim, D.: Fast convergence rates of deep neural networks for classification. Neural Networks **138**, 179–197 (2021)

Kong, I., Yang, D., Lee, J., Ohn, I., Baek, G., Kim, Y.: Masked bayesian neural networks: Theoretical guarantee and its posterior inference. In: International Conference on Machine Learning, pp. 17462–17491 (2023). PMLR

LeCun, Y., Bengio, Y., Hinton, G.: Deep learning. nature **521**(7553), 436–444 (2015)

Lee, K., Lee, J.: Asymptotic properties for bayesian neural network in besov space. Advances in Neural Information Processing Systems **35**, 5641–5653 (2022)

Massart, P.: Concentration Inequalities and Model Selection. Lecture Notes in Mathematics, vol. 1896, p. 337. Springer, Saint-Flour (2007). Lectures from the 33rd Summer School on Probability Theory, July 6–23, 2003, Edited by Jean Picard

McAllester, D.A.: Some PAC-Bayesian theorems. In: Proceedings of the Eleventh Annual Conference on Computational Learning Theory, pp. 230–234. ACM, New York (1998)

Meyer, J.T.: Optimal convergence rates of deep neural networks in a classification setting. Electronic Journal of Statistics **17**(2), 3613–3659 (2023)

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

Neyshabur, B., Bhojanapalli, S., Srebro, N.: A pac-bayesian approach to spectrally-normalized margin bounds for neural networks. In: International Conference on Learning Representations (2018)

Polson, N.G., Ročková, V.: Posterior concentration for sparse deep learning. In: Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., Garnett, R. (eds.) Advances in Neural Information Processing Systems, vol. 31. Curran Associates, Inc., ??? (2018)

Russo, D., Zou, J.: How much does your data exploration overfit? controlling bias

via information usage. IEEE Transactions on Information Theory **66**(1), 302–323 (2019)

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

Shen, G., Jiao, Y., Lin, Y., Huang, J.: Approximation with cnns in sobolev space: with applications to classification. Advances in Neural Information Processing Systems **35**, 2876–2888 (2022)

Shawe-Taylor, J., Williamson, R.: A PAC analysis of a Bayes estimator. In: Proceedings of the Tenth Annual Conference on Computational Learning Theory, pp. 2–9. ACM, New York (1997)

Suzuki, T.: Fast generalization error bound of deep learning from a kernel perspective. In: International Conference on Artificial Intelligence and Statistics, pp. 1397–1406 (2018). PMLR

Suzuki, T.: Adaptivity of deep relu network for learning in besov and mixed smooth besov spaces: optimal rate and curse of dimensionality. In: The 7tth International Conference on Learning Representations (2019)

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

Tsybakov, A.B.: Introduction to Nonparametric Estimation. Springer Series in Statistics, p. 214. Springer, ??? (2009). https://doi.org/10.1007/b13794 . Revised and extended from the 2004 French original, Translated by Vladimir Zaiats

Vapnik, V.N.: Statistical Learning Theory. Wiley, ??? (1998)

Vladimirova, M., Verbeek, J., Mesejo, P., Arbel, J.: Understanding priors in bayesian neural networks at the unit level. In: International Conference on Machine Learning, pp. 6458–6467 (2019). PMLR

Wang, S., Shang, Z.: Minimax optimal high-dimensional classification using deep neural networks. Stat **11**(1), 482 (2022)

Zhang, C., Bengio, S., Hardt, M., Recht, B., Vinyals, O.: Understanding deep learning (still) requires rethinking generalization. Communications of the ACM **64**(3), 107–115 (2021)

Zhang, T.: Statistical behavior and consistency of classification methods based on convex risk minimization. The Annals of Statistics **32**(1), 56–85 (2004)

Zhang, T.: Information-theoretic upper and lower bounds for statistical estimation. IEEE Transactions on Information Theory **52**(4), 1307–1321 (2006)