Convergence Rates of Variational Inference in Sparse Deep Learning

Badr-Eddine Chérief-Abdellatif [1]

Abstract

Variational inference is becoming more and more popular for approximating intractable posterior distributions in Bayesian statistics and machine learning. Meanwhile, a few recent works have provided theoretical justification and new insights on deep neural networks for estimating smooth functions in usual settings such as nonparametric regression. In this paper, we show that variational inference for sparse deep learning retains precisely the same generalization properties than exact Bayesian inference. In particular, we show that a wise choice of the neural network architecture leads to near-minimax rates of convergence for Hölder smooth functions. Additionally, we show that the model selection framework over the architecture of the network via ELBO maximization does not overfit and adaptively achieves the optimal rate of convergence.

1. Introduction

The Bayesian approach to learning in neural networks has a long history. Bayesian Neural Networks have been first proposed in the 90s and widely studied since then (MacKay, 1992a; Neal, 1995). They offer a probabilistic interpretation and a measure of uncertainty for deep learning models. They are more robust to overfitting than classical neural networks and still achieve great performance even on small datasets. A prior distribution is put on the parameters of the network, namely the weight matrices and the bias vectors, for instance a Gaussian or a uniform distribution, and Bayesian inference is done through the likelihood specification. Nevertheless, state-of-the-art neural networks may contain millions of parameters and the form of a neural network is not adapted to exact integration, which makes the posterior distribution be intractable in practice.

[1]CREST - ENSAE - Institut Polytechnique de Paris. Correspondence to: Badr-Eddine Chérief-Abdellatif <badr.eddine.cherief.abdellatif@ensae.fr>.

Proceedings of the 37^{th} International Conference on Machine Learning, Vienna, Austria, PMLR 119, 2020.

Modern approximate inference mainly relies on variational inference (VI), with sometimes a flavor of sampling techniques. A lot of recent papers have investigated variational inference for Deep Neural Networks (DNNs) (Hinton & van Camp, 1993; Graves, 2011; Blundell et al., 2015) to fit an approximate posterior that maximizes the Evidence Lower Bound (ELBO). For instance, (Blundell et al., 2015) introduced Bayes by Backprop, one of the most famous techniques of VI applied to neural networks, which derives a fully factorized Gaussian approximation to the posterior: using the reparameterization trick (Opper & Archambeau, 2008), the gradients of ELBO towards parameters of the Gaussian approximation can be computed by backpropagation, and then be used for updates. Another point of interest in DNNs is the choice of the prior. (Blundell et al., 2015) introduced a mixture of Gaussians prior on the weights, with one component tightly concentrated around zero, imitating the sparsity-inducing spike-and-slab prior. This offers a Bayesian alternative to the dropout regularization procedure (Srivastava et al., 2014) which injects sparsity in the network by switching off randomly some of the weights of the network. This idea goes back to David MacKay who discussed in his thesis the possibility of choosing a spike-and-slab prior over the weights of the neural network (MacKay, 1992b). More recently, (Rockova & Polson, 2018) introduced Spike-and-Slab Deep Learning (SS-DL), a fully Bayesian alternative to dropout for improving generalizability of deep ReLU networks.

1.1. Related work

Although deep learning is extremely popular, the study of generalization properties of DNNs is still an open problem. Some works have been conducted in order to investigate the theoretical properties of neural networks from different points of view. The literature developed in the past decades can be mainly shared in three parts. First, the approximation theory wonders how well a function can be approximated by neural networks. The first studies were mostly conducted to obtain approximation guarantees for shallow neural nets with a single hidden layer (Cybenko, 1989; Barron, 1993). Since then, modern research has focused on the expressive power of depth and extended the previous results to deep neural networks with a larger number of layers (Bengio & Delalleau, 2011; Yarotsky, 2016;

Petersen & Voigtländer, 2017; Grohs et al., 2019). Indeed, even though the universal approximation theorem (Cybenko, 1989) states that a shallow neural network containing a finite number of neurons can approximate any continuous function on compact sets under mild assumptions on the activation function, recent advances showed that a shallow network requires exponentially many neurons in terms of the dimension to represent a monomial function, whereas linearly many neurons are sufficient for a deep network (Rolnick & Tegmark, 2018). Second, as the objective function in deep learning is known to be nonconvex, the optimization community has discussed the landscape of the objective as well as the dynamics of some learning algorithms such as Stochastic Gradient Descent (SGD) (Baldi & Hornik, 1989; Stanford et al., 2000; Soudry & Carmon, 2016; Kawaguchi, 2016; Kawaguchi et al., 2019; Oymak & Soltanolkotabi, 2019; Nguyen et al., 2019; Allen-Zhu et al., 2019; Du et al., 2019). Finally, the statistical learning community has investigated generalization properties of DNNs, see (Barron, 1994; Zhang et al., 2017; Schmidt-Hieber, 2017; Suzuki, 2018; Imaizumi & Fukumizu, 2019; Suzuki, 2019). In particular, (Schmidt-Hieber, 2017) and (Suzuki, 2019) showed that estimators in nonparametric regression based on sparsely connected DNNs with ReLU activation function and wisely chosen architecture achieve the minimax estimation rates (up to logarithmic factors) under classical smoothness assumptions on the regression function. In the same time, (Bartlett et al., 2017) and (Neyshabur et al., 2018) respectively used Rademacher complexity and PAC-Bayes theory to get spectrally-normalized margin bounds for deep ReLU networks, as well as (Dziugaite & Roy, 2017). More recently, (Imaizumi & Fukumizu, 2019) and (Hayakawa & Suzuki, 2019) showed the superiority of DNNs over linear operators in some situations when DNNs achieve the minimax rate of convergence while alternative methods fail. From a Bayesian point of view, (Rockova & Polson, 2018) and (Suzuki, 2018) studied the concentration of the posterior distribution while (Vladimirova et al., 2019) investigated the regularization effect of prior distributions at the level of the units.

Such as for generalization properties of DNNs, only little attention has been put in the literature towards the theoretical properties of VI until recently. (Alquier et al., 2016) studied generalization properties of variational approximations of Gibbs distributions in machine learning for bounded loss functions. (Alquier & Ridgway, 2017; Zhang & Gao, 2017; Sheth & Khardon, 2017; Bhattacharya et al., 2018; Chérief-Abdellatif & Alquier, 2018; Cherief-Abdellatif, 2019; Jaiswal et al., 2019a) extended the previous guarantees to more general statistical models and studied the concentration of variational approximations of the posterior distribution, while (Wang & Blei, 2018) provided Bernstein-von-Mises' theorems for variational approximations in parametric models. (Huggins et al., 2018; Campbell & Li, 2019; Jaiswal et al., 2019b) discussed theoretical properties of variational inference algorithms based on various divergences (respectively Wasserstein and Hellinger distances, and Rényi divergence). More recently, (Chérief-Abdellatif et al., 2019) presented generalization bounds for online variational inference. All these works show that under mild conditions, the variational approximation is consistent and achieves the same rate of convergence than the Bayesian posterior distribution it approximates. Note that (Alquier & Ridgway, 2017; Bhattacharya et al., 2018; Chérief-Abdellatif & Alquier, 2018; Cherief-Abdellatif, 2019) restricted their studies to tempered versions of the posterior distribution where the likelihood is raised to an α-power ($\alpha < 1$) as it is known to require less stringent assumptions to obtain consistency and to be robust to misspecification, see respectively (Bhattacharya et al., 2016) and (Grünwald & Van Ommen, 2017). Nevertheless, some questions remain unanswered, as the theoretical study of generalization of variational inference for deep neural networks.

2. Outline

This paper aims at filling the gap between theory and practice when using variational approximations for tempered Bayesian Deep Neural Networks. To the best of our knowledge, this is the first paper to present theoretical generalization error bounds of variational inference for Bayesian deep learning. Inspired by the related literature, our work is motivated by the following questions:

- Do consistency of Bayesian DNNs still hold when an approximation is used instead of the exact posterior distribution, and can we obtain the same rates of convergence than those obtained for the regular posterior distribution and frequentist estimators ?
- Is it possible to obtain a nonasymptotic generalization error bound that holds for (almost) any generating distribution function and that gives a general formula ?
- What about the consistency of numerical algorithms used to compute these variational approximations ?
- Can we obtain new insights on the structure of the networks ?

It also raises the question of finding a relevant general definition of consistency that can be used to provide theoretical properties for the exact Bayesian DNNs distribution and their variational approximations. Indeed, a classical criterion used to assess frequentist guarantees for Bayesian estimators is the concentration of the posterior (to the true distribution) (Ghosal et al., 2000). Nevertheless, posterior

concentration to the true distribution only applies when the model is well specified, or at least when the model contains distributions in the neighborhood of the true distribution, which is problematic for misspecified models e.g. when the neural network does not sufficiently approximate the generating distribution. And although the posterior distribution may concentrate to the best approximation of the true distribution in KL divergence in such misspecified models, there exists pathological cases where the regular Bayesian posterior is not consistent at all, see (Grünwald & Van Ommen, 2017). This is the reason why we focus here on tempered posteriors which are robust to such misspecification. Therefore, we introduce in Section 3 a notion of consistency of a Bayesian estimator which is closely related to the notion of concentration - even stronger - and which enables a more robust formulation of generalization error bounds for variational approximations. See Appendix A in the supplementary material for more details on the connection between the notions of consistency and concentration.

The main contribution of this paper, a nonasymptotic generalization error bound for variational inference in sparse DL in the nonparametric regression framework, answers the first two motivating questions. This generalization result is similar to theoretical inequalities in the seminal works of (Suzuki, 2018; Imaizumi & Fukumizu, 2019; Rockova & Polson, 2018) on generalization properties of deep neural networks, and is inspired by the general literature on the consistency of variational approximations (Alquier & Ridgway, 2017; Bhattacharya et al., 2018). In particular, it states that under the same conditions, sparse variational approximations of posterior distributions of deep neural networks are consistent at the same rate of convergence than the exact posterior.

Then we focus on optimization aspects. We no longer assume an ideal optimization, as done for instance in (Schmidt-Hieber, 2017; Imaizumi & Fukumizu, 2019). We address in this paper the question of the consistency of numerical algorithms used to compute our ideal approximations. We consider an optimization error given by any algorithm and independent to the statistical error, and we show how it affects our generalization result. Our upper bound highlights the connection between the consistency of the variational approximation and the convergence of the ELBO.

We also provide insights on the structure of the network which leads to optimal rates of convergence, i.e. its depth, its width and its sparsity. Indeed, in our first generalization error bound, the structure of the network is ideally tuned for some choice of the generating function. Nevertheless, the characteristics of the regression function may be unknown, e.g. we may know that the regression function is Hölder continuous but we ignore its level of smoothness. We propose here an automated method for choosing the architecture of the network. We introduce a classical model selection framework based on the ELBO criterion (Cherief-Abdellatif, 2019), and we show that the variational approximation associated with the selected structure does not overfit and adaptively achieves the optimal rate of convergence even without any oracle information.

The rest of this paper is organized as follows. Section 3 introduces the notations and the framework that will be considered in the paper, and presents sparse spike-and-slab variational inference for deep neural networks. Section 4 provides theoretical generalization error bounds for variational approximations of DNNs and shows the optimality of the method for estimating Hölder smooth functions. Finally, insights on the choice of the architecture of the network are given in Section 5 via the ELBO maximization framework. All the technical proofs are deferred to the appendices in the supplementary material.

3. Sparse deep variational inference

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 $\|x\|_\infty = \max_{1\leq i\leq d} |x_i|$, $\|f\|_2^2 = \int f^2$ and $\|f\|_\infty = \sup_{y\in[-1,1]^d} |f(y)|$. We also introduce the notion of β-Hölder continuity for $\beta > 0$ (Tsybakov, 2008) which is rigorously defined in Appendix C in the supplementary material.

3.1. Nonparametric regression

We consider the nonparametric regression framework. We have a collection of random variables $(X_i, Y_i) \in [-1, 1]^d \times \mathbb{R}$ for $i = 1, ..., n$ which are independent and identically distributed (i.i.d.) with the generating process:

$$\begin{cases} X_i \sim \mathcal{U}([-1, 1]^d), \\ Y_i = f_0(X_i) + \zeta_i \end{cases}$$

where $\mathcal{U}([-1, 1]^d)$ is the uniform distribution on the interval $[-1, 1]^d$, $\zeta_1, ..., \zeta_n$ are i.i.d. Gaussian random variables with mean 0 and known variance σ^2, and $f_0 : [-1, 1]^d \to \mathbb{R}$ is the true unknown function. For instance, the true regression function f_0 may belong to the set of Hölder functions with level of smoothness β.

3.2. Deep neural networks

We call deep neural network any map $f_\theta : \mathbb{R}^d \to \mathbb{R}$ defined recursively as follows:

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

where a value $L = 2$ corresponds to a shallow network and $L \geq 3$ to a deep neural network. ρ is an activation function acting componentwise. For instance, we can choose the ReLU activation function $\rho(u) = \max(u, 0)$. Each $A_\ell \in \mathbb{R}^{D_\ell \times D_{\ell-1}}$ is a weight matrix such that its (i, j) coefficient, called edge weight, connects the j-th neuron of the $(\ell-1)$-th layer to the i-th neuron of the ℓ-th layer, and each $b_\ell \in \mathbb{R}^{D_\ell}$ is a shift vector such that its i-th coefficient, called node vector, represents the weight associated with the i-th node of layer ℓ. We set $D_0 = d$ 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, i.e. the total number of nonzero entries in matrices A_ℓ and vectors b_ℓ, its number of layers $L \geq 2$ (excluding the input layer), and its width $D \geq 1$. We have $S \leq T$ where $T = \sum_{\ell=1}^{L} D_\ell(D_{\ell-1} + 1)$ is the total number of coefficients in a fully connected network. By now, we consider that S, L and D are considered deterministic, and $d = \mathcal{O}(1)$ as $n \to +\infty$. In particular, we assume that $d \leq D$, which implies that $T \leq LD(D + 1)$. We also suppose that the absolute values of all coefficients are upper bounded by some positive constant $B \geq 2$. This boundedness assumption will be relaxed in the appendix, see Appendix G. Then, the parameter of a DNN is $\theta = \{(A_1, b_1), ..., (A_L, b_L)\}$, and we denote $\Theta_{S,L,D}$ the set of all possible parameters. We will also alternatively consider the stacked coefficients parameter $\theta = (\theta_1, ..., \theta_T)$.

3.3. Bayesian modeling

We adopt a Bayesian approach, and we place a spike-and-slab prior π (Castillo et al., 2015) over the parameter space $\Theta_{S,L,D}$ (equipped with some suited sigma-algebra) that is defined hierarchically. The spike-and-slab prior is known to be a relevant alternative to dropout for Bayesian deep learning, see (Rockova & Polson, 2018). First, we sample a vector of binary indicators $\gamma = (\gamma_1, ..., \gamma_T) \in \{0, 1\}^T$ uniformly among the set $\mathcal{S}_T^S$ of T-dimensional binary vectors with exactly S nonzero entries, and then given γ_t for each $t = 1, ..., T$, we put a spike-and-slab prior on θ_t that returns 0 if $\gamma_t = 0$ and a random sample from a uniform distribution on $[-B, B]$ otherwise:

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

where $\delta_{\{0\}}$ is a point mass at 0 and $\mathcal{U}([-B, B])$ is a uniform distribution on $[-B, B]$. We recall that the sparsity level S is fixed here and that this assumption will be relaxed in Section 5.

Remark 3.1. *We consider uniform distributions for simplicity as in similar works (Rockova & Polson, 2018; Suzuki, 2018), but Gaussian distributions can be used as well when working on an unbounded parameter set $\Theta_{S,L,D}$, see Appendix G in the supplementary material.*

Then we define the tempered posterior distribution $\pi_{n,\alpha}$ on parameter $\theta \in \Theta_{S,L,D}$ using prior π for any $\alpha \in (0, 1)$:

$$\pi_{n,\alpha}(d\theta) \propto \exp\left(- \frac{\alpha}{2\sigma^2} \sum_{i=1}^{n} (Y_i - f_\theta(X_i))^2 \right) \pi(d\theta),$$

which is a slight variant of the definition of the regular Bayesian posterior (for which $\alpha = 1$). This distribution is known to be easier to sample from, to require less stringent assumptions to obtain concentration, and to be robust to misspecification, see respectively (Behrens et al., 2012), (Bhattacharya et al., 2016) and (Grünwald & Van Ommen, 2017).

3.4. Sparse variational inference

The variational Bayes approximation $\tilde{\pi}_{n,\alpha}$ of the tempered posterior is defined as the projection (with respect to the Kullback-Leibler divergence) of the tempered posterior onto some set $\mathcal{F}_{S,L,D}$:

$$\tilde{\pi}_{n,\alpha} = \underset{q \in \mathcal{F}_{S,L,D}}{\arg\min} \, \mathrm{KL}(q \| \pi_{n,\alpha}).$$

which is equivalent to:

$$\underset{q \in \mathcal{F}_{S,L,D}}{\arg\min} \left\{ \frac{\alpha}{2\sigma^2} \sum_{i=1}^{n} \int (Y_i - f_\theta(X_i))^2 q(d\theta) + \mathrm{KL}(q \| \pi) \right\} \tag{1}$$

where the function inside the argmin operator in (1) is the opposite of the evidence lower bound $\mathcal{L}_n(q)$.

We choose a sparse spike-and-slab variational set $\mathcal{F}_{S,L,D}$ - see for instance (Tonolini et al., 2019) - which can be seen as an extension of the popular mean-field variational set with a dependence assumption specifying the number of active neurons. The mean-field approximation is based on a decomposition of the space of parameters $\Theta_{S,L,D}$ as a product $\theta = (\theta_1, ..., \theta_T)$ and consists in compatible product distributions on each parameter θ_t, $t = 1, ..., T$. Here, we fit a distribution in the family that matches the prior: we first choose a distribution q_γ on the set $\mathcal{S}_T^S$ that selects a T-dimensional binary vector γ with S nonzero entries, and then we place a spike-and-slab variational approximation on each θ_t given γ_t:

$$\begin{cases} \gamma \sim q_\gamma, \\ \theta_t | \gamma_t \sim \gamma_t \, \mathcal{U}([l_t, u_t]) + (1 - \gamma_t)\delta_{\{0\}}, \quad t = 1, ..., T \end{cases}$$

where $-1 \leq l_t \leq u_t \leq 1$, with the distribution q_γ and the intervals $[l_t, u_t]$, $t = 1, ..., T$ as the hyperparameters of the variational set $\mathcal{F}_{S,L,D}$. In particular, if we choose a deterministic $q_\gamma = \delta_{\{\gamma'\}}$ with $\gamma' \in \mathcal{S}_T^S$, then we will obtain

a parametric mean-field approximation. See Section 6.6 of the PhD thesis of (Gal, 2016) for a more detailed discussion on the connection between Gaussian mean-field and sparse spike-and-slab posterior approximations.

The generalization error of a Bayesian estimator ρ (either the tempered posterior $\pi_{n,\alpha}$ or its variational approximation $\tilde{\pi}_{n,\alpha}$) is the expected average of the squared L_2-distance to the true generating function over ρ:

$$\mathbb{E}\left[\int \|f_\theta - f_0\|_2^2 \rho(d\theta)\right].$$

We say that a Bayesian estimator is consistent at rate $r_n \to 0$ if its generalization error is upper bounded by r_n. Notice that consistency of the Bayesian estimator implies concentration to f_0. Again, see Appendix A for the connection between these two notions.

4. Generalization of variational inference for neural networks

The first result of this section is a variant of the result of (Rockova & Polson, 2018) on the Bayesian distribution for Hölder regression functions. Indeed, we provide a concentration result on the posterior distribution for the expected L_2-distance instead of the empirical L_2-distance, which enables generalization instead of reconstruction on the training datapoints. This result is then extended again to the variational approximation for our definition of consistency: we show that we can still achieve near-optimality using an approximation of the posterior without any additional assumption. Finally, we explain how we can incorporate optimization error in our generalization results.

4.1. Concentration of the posterior

(Rockova & Polson, 2018) gives the first posterior concentration result for deep ReLU networks when estimating Hölder smooth functions in nonparametric regression with empirical L_2-distance. The authors highlight the flexibility of DNNs over other methods for estimating β-Hölder smooth functions as there is a large range of values of the level of smoothness β for which one can obtain concentration, e.g. $0 < \beta < d$ for a DNN against $0 < \beta < 1$ for a Bayesian tree.

The following theorem provides the concentration of the tempered posterior distribution $\pi_{n,\alpha}$ for deep ReLU neural networks when using the expected L_2-distance for some suitable architecture of the network:

Theorem 1. *Let us assume that $\alpha \in (0,1)$, that f_0 is β-Hölder smooth with $0 < \beta < d$ and that the activation function is ReLU. We consider the architecture of (Rockova & Polson, 2018) for some positive constant C_D independent of n:*

$$L = 8 + (\lfloor \log_2 n \rfloor + 5)(1 + \lceil \log_2 d \rceil),$$

$$D = C_D \lfloor n^{\frac{d}{2\beta+d}} / \log n \rfloor,$$

$$S \leq 94 d^2 (\beta+1)^{2d} D (L + \lceil \log_2 d \rceil).$$

Then the tempered posterior distribution $\pi_{n,\alpha}$ concentrates at the minimax rate $r_n = n^{\frac{-2\beta}{2\beta+d}}$ up to a (squared) logarithmic factor for the expected L_2-distance in the sense that:

$$\pi_{n,\alpha}\left(\theta \in \Theta_{S,L,D} \;/\; \|f_\theta - f_0\|_2^2 > M_n \cdot n^{\frac{-2\beta}{2\beta+d}} \cdot \log^2 n\right) \to 0$$

in probability as $n \to +\infty$ for any $M_n \to +\infty$.

In order to prove Theorem 1, we actually have to check that the so-called *prior mass* condition is satisfied:

$$\pi\left(\theta \in \Theta_{S,L,D} \;/\; \|f_\theta - f_0\|_2^2 \leq r_n\right) \geq e^{-n r_n}. \qquad (2)$$

This assumption, introduced in (Ghosal et al., 2000) in order to obtain the concentration of the regular posterior distribution states that the prior must give enough mass to some neighborhood of the true parameter. As shown in (Bhattacharya et al., 2016), this condition is even sufficient for tempered posteriors. Actually, this inequality was first stated using the KL divergence instead of the expected L_2-distance (see Condition 2.4 in Theorem 2.1 in (Ghosal et al., 2000)), but the KL metric is equivalent to the squared L_2-metric in regression problems with Gaussian noise. This prior mass condition gives us the rate of convergence of the tempered posterior $r_n = n^{\frac{-2\beta}{2\beta+d}}$ (up to a squared logarithmic factor) which is known to be optimal when estimating β-Hölder smooth functions (Tsybakov, 2008). Note that the $\log^2 n$ term is common in the theoretical deep learning literature (Imaizumi & Fukumizu, 2019; Suzuki, 2019; Schmidt-Hieber, 2017).

Remark 4.1. *The number of parameters of order $n^{\frac{2d}{2\beta+d}} / \log n \in [n^{2/3}/\log(n), n^2/\log(n)]$ is high compared to standard machine learning methods, which may lead to overfitting and hence prevent the procedure from achieving the minimax rate of convergence. The sparsity parameter S which gives a network with a small number of nonzero parameters along with the spike-and-slab prior help us tackle this issue and obtain optimal rates of convergence (up to logarithmic factors).*

4.2. A generalization error bound

The result we state in this subsection applies to a wide range of activation functions, including the popular ReLU activation and the identity map:

Assumption 4.1. *In the following, we assume that the activation function ρ is 1-Lispchitz continuous (with respect to the aboluste value) and is such that for any $x \in \mathbb{R}$, $|\rho(x)| \leq |x|$.*

We do not assume any longer that the regression function is β-Hölder and we consider any structure (S, L, D). The following theorem gives a generalization error bound when using variational approximations instead of exact tempered posteriors for DNNs. The proof is given in Appendix B and is based on PAC-Bayes theory (Catoni, 2007; Guedj, 2019):

Theorem 2. *For any $\alpha \in (0, 1)$,*

$$\mathbb{E}\left[\int \|f_\theta - f_0\|_2^2 \tilde{\pi}_{n,\alpha}(d\theta)\right] \tag{3}$$
$$\leq \frac{2}{1-\alpha} \inf_{\theta^* \in \Theta_{S,L,D}} \|f_{\theta^*} - f_0\|_2^2 + \frac{2}{1-\alpha}\left(1 + \frac{\sigma^2}{\alpha}\right) r_n^{S,L,D}$$

with

$$r_n^{S,L,D} = \frac{LS}{n}\log(BD) + \frac{2S}{n}\log(BLD) + \frac{S}{n}\log\left(7dL\max\left(\frac{n}{S}, 1\right)\right).$$

The oracle inequality (3) ensures consistency of variational Bayes for estimating neural networks and provides the associated rate of convergence given the structure (S, L, D). Indeed, if f_0 is a neural network with structure (S, L, D), then the infimum term on the right hand side of the inequality vanishes and we obtain a rate of convergence of order

$$r_n^{S,L,D} \sim \max\left(\frac{S\log(nL/S)}{n}, \frac{LS\log D}{n}\right),$$

which underlines a linear dependence on the number of layers and the sparsity. In fact, this rate of convergence is determined by the *extended prior mass condition* (Alquier & Ridgway, 2017; Chérief-Abdellatif & Alquier, 2018; Cherief-Abdellatif, 2019), which requires that in addition to the previous prior mass condition of (Ghosal et al., 2000) and (Bhattacharya et al., 2016), the variational set $\mathcal{F}_{S,L,D}$ must contain probability distributions q that are concentrated enough around the true generating function f_0. One of the main findings of Theorem 2 is that our choice of the sparse spike-and-slab variational set $\mathcal{F}_{S,L,D}$ is rich enough and that both conditions are actually similar and lead to the same rate of convergence. Hence, the rate of convergence is the one that satisfies the prior mass condition (2). In particular, as the prior distribution is uniform over the parameter space, the negative logarithm of the prior mass of the neighborhood of the true regression function in Equation (2) is a local covering entropy, that is the logarithm of the number of $r_n^{S,L,D}$-balls needed to cover a neighborhood of the true regression function. Especially, it has been shown in previous studies that this local covering entropy fully characterizes the rate of convergence of the empirical risk minimizer for DNNs (Schmidt-Hieber, 2017; Suzuki, 2019). The rate $r_n^{S,L,D}$ we obtain in this work is exactly of the same order than the upper bound on the covering entropy number given in Lemma 5 in (Schmidt-Hieber, 2017) and in Lemma 3 in (Suzuki, 2019) which derive rates of convergence for the empirical risk minimizer using different proof techniques. Note that replacing a uniform by a Gaussian in the prior and variational distributions leads to the same rate of convergence, see Appendix G.

Nevertheless, deep neural networks are mainly used for their computational efficiency and their ability to approach complex functions, which makes the task of estimating a neural network not so popular in machine learning. As said earlier, (Imaizumi & Fukumizu, 2019) used neural networks for estimating non-smooth functions. In such a context where the neural network model is misspecified, our generalization error bound is robust and still holds, and satisfies the best possible balance between bias and variance.

Indeed, the upper bound on the generalization error on the right-hand-side of (3) is mainly divided in two parts: the approximation error of f_0 by a DNN f_{θ^*} in $\Theta_{S,L,D}$ (i.e. the bias) and the estimation error $r_n^{S,L,D}$ of a neural network f_{θ^*} in $\Theta_{S,L,D}$ (i.e. the variance). For instance, even if the generalization power is decreasing linearly with respect to the number of layers compared to the logarithmic dependence on the width due to the variance term, this effect is compensated by the benefits of depth in the approximation theory of deep learning. Then, as there exists relationships between the bias/the variance and the architecture of a neural network (respectively due to the approximation theory/the form of $r_n^{S,L,D}$), Theorem 2 gives both a general formula for deriving rates of convergence for variational approximations and insight on the way to choose the architecture. We choose the architecture that minimizes the right-hand-side of (3), which can lead to minimax estimators for smooth functions. It also connects the approximation and estimation theories following previous studies. This was done for instance by (Schmidt-Hieber, 2017; Suzuki, 2019; Imaizumi & Fukumizu, 2019) who exploited the effectiveness of ReLU activation function in terms of approximation ability (Yarotsky, 2016; Petersen & Voigtländer, 2017) for Hölder/Besov smooth and piecewise smooth generating functions.

Now we illustrate Theorem 2 on Hölder smooth functions. The following result shows that the variational approximation achieves the same rate of convergence than the posterior distribution it approximates, and even the minimax rate of convergence if the architecture is well chosen. We present both consistency and concentration results.

Corollary 3. *Let us fix $\alpha \in (0, 1)$. We consider the ReLU activation function. Assume that f_0 is β-Hölder smooth with*

$0 < \beta < d$. Then with L, D and S defined as in Theorem 1, the variational approximation of the tempered posterior distribution $\tilde{\pi}_{n,\alpha}$ is consistent and hence concentrates at the minimax rate $r_n = n^{\frac{-2\beta}{2\beta+d}}$ (up to a squared logarithmic factor):

$$\tilde{\pi}_{n,\alpha}\left(\theta \in \Theta_{S,L,D} \;/\; \|f_\theta - f_0\|_2^2 > M_n \cdot n^{\frac{-2\beta}{2\beta+d}} \cdot \log^2 n\right) \to 0$$

in probability as $n \to +\infty$ for any $M_n \to +\infty$.

4.3. Optimization error

In this subsection, we discuss the effect of an optimization error that is independent on the previous statistical error. Indeed, in the variational Bayes community, people use approximate algorithms in practice to solve the optimization problem (1) when the model is non-conjugate, i.e. the VB solution is not available in closed-form. This is the case here when considering a sparse spike-and-slab variational approximation in $\mathcal{F}_{S,L,D}$ for DNNs with hyperparameters $\phi = (q_\gamma, (\phi_t)_{1\leq t\leq T})$ and an algorithm that gives a sequence of hyperparameters $(\phi^k)_{k\geq 1}$ and associated variational approximations $(\tilde{\pi}^k_{n,\alpha})_{k\geq 1}$. The following theorem gives a statistical guarantee for any approximation $\tilde{\pi}^k_{n,\alpha}$, $k \geq 1$:

Theorem 4. *For any $\alpha \in (0,1)$,*

$$\mathbb{E}\left[\int \|f_\theta - f_0\|_2^2 \tilde{\pi}^k_{n,\alpha}(d\theta)\right] \leq \frac{2}{1-\alpha}\inf_{\theta^*}\|f_{\theta^*} - f_0\|_2^2 + \frac{2}{1-\alpha}\left(1 + \frac{\sigma^2}{\alpha}\right) r_n^{S,L,D} + \frac{2\sigma^2}{\alpha(1-\alpha)} \cdot \frac{\mathbb{E}[\mathcal{L}_n^* - \mathcal{L}_n^k]}{n},$$

where $\mathcal{L}_n^*$ is the maximum of the evidence lower bound i.e. the ELBO evaluated at $\tilde{\pi}_{n,\alpha}$, while $\mathcal{L}_n^k$ is the ELBO evaluated at $\tilde{\pi}^k_{n,\alpha}$.

We establish a clear connection between the convergence (in mean) of the ELBO $\mathcal{L}_n^k$ to $\mathcal{L}_n^*$ and the consistency of our algorithm $\tilde{\pi}^k_{n,\alpha}$. Indeed, as soon as the ELBO $\mathcal{L}_n^k$ converges at rate $c_{k,n}$, then our variational approximation $\tilde{\pi}^k_{n,\alpha}$ is consistent at rate:

$$\max\left(\frac{c_{k,n}}{n}, \frac{S\log(nL/S)}{n}, \frac{SL\log D}{n}\right).$$

In particular, as soon as k is such that $c_{k,n} \leq \max(S\log n, S\log D)$, then we obtain consistency of $\tilde{\pi}^k_{n,\alpha}$ at rate $r_n^{S,L,D}$, i.e. $\tilde{\pi}^k_{n,\alpha}$ and $\tilde{\pi}_{n,\alpha}$ have the same rate of convergence.

However, deriving the convergence of the ELBO is a hard task. For instance, when considering a simple Gaussian mean-field approximation without sparsity, the variational objective $\mathcal{L}_n$ can be maximized using either stochastic (Graves, 2011; Blundell et al., 2015) or natural gradient methods (Khan et al., 2018) on the parameters of the Gaussian approximation. The convergence of the ELBO is often met in practice (Buchholz et al., 2018; Mishkin et al., 2018) and the recent work of (Osawa et al., 2019) even showed that Bayesian deep learning enables practical deep learning and matches the performance of standard methods while preserving benefits of Bayesian principles. Nevertheless, the objective is nonconvex and hence it is difficult to prove the convergence to a global maximum in theory. Some recent papers studied global convergence properties of gradient descent algorithms for frequentist classification and regression losses (Du et al., 2019; Allen-Zhu et al., 2019) that we may extend to gradient descent algorithms for the ELBO objective such as Variational Online Gauss Newton or Vadam (Khan et al., 2018; Osawa et al., 2019).

Another point is to develop and study more complex algorithms than simple gradient descent that deal with spike-and-slab sparsity-inducing variational inference, as for instance (Titsias & Lázaro-Gredilla, 2011) did for multi-task and multiple kernel learning. Also, (Louizos et al., 2018) connected sparse spike-and-slab variational inference with L_0-norm regularization for neural networks and proposed a solution to the intractability of the L_0-penalty term through the use of non-negative stochastic gates, while (Bellec et al., 2018) proposed an algorithm preserving sparsity during training. Nevertheless, these optimization concerns fall beyond the scope of this paper and are left for further research.

5. Architecture design via ELBO maximization

We saw in Section 4 that the choice of the architecture of the neural network is crucial and can lead to faster convergence and better approximation. In this section, we formulate the architecture design of DNNs as a model selection problem and we investigate the ELBO maximization strategy which is very popular in the variational Bayes community. This approach is different from (Rockova & Polson, 2018) which is fully Bayesian and treats the parameters of the network architecture, namely the depth, the width and the sparsity, as random variables. We show that the ELBO criterion does not overfit and is adaptive: it provides a variational approximation with the optimal rate of convergence, and it does not require the knowledge of the unknown aspects of the regression function f_0 (e.g. the level of smoothness for smooth functions) to select the optimal variational approximation.

We denote $\mathcal{M}_{S,L,D}$ the statistical model associated with the parameter set $\Theta_{S,L,D}$. We consider a countable number of models, and we introduce prior beliefs $\pi_{S,L,D}$ over the sparsity, the depth and the width of the network, that can be defined hierarchically and that are known before-

hand. For instance, the prior beliefs can be chosen such that $\pi_L = 2^{-L}$, $\pi_{D|L}$ follows a uniform distribution over $\{d, ..., \max(e^L, d)\}$ given L, and $\pi_{S|L,D}$ a uniform distribution over $\{1, ..., T\}$ given L and D (we recall that T is 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 still consider spike-and-slab priors on $\theta_{S,L,D} \in \Theta_{S,L,D}$ given model $\mathcal{M}_{S,L,D}$.

Each tempered posterior associated with model $\mathcal{M}_{S,L,D}$ is denoted $\pi_{n,\alpha}^{S,L,D}$. We recall that the variational approximation $\tilde{\pi}_{n,\alpha}^{S,L,D}$ associated with model $\mathcal{M}_{S,L,D}$ is defined as the distribution into the variational set $\mathcal{F}_{S,L,D}$ that maximizes the Evidence Lower Bound:

$$\tilde{\pi}_{n,\alpha}^{S,L,D} = \underset{q^{S,L,D} \in \mathcal{F}_{S,L,D}}{\arg\max} \mathcal{L}_n(q^{S,L,D}).$$

We will simply denote in the following $\mathcal{L}_n^*(S, L, D)$ the closest approximation to the log-evidence i.e., the value of the ELBO evaluated at its maximum:

$$\mathcal{L}_n^*(S, L, D) = \mathcal{L}_n(\tilde{\pi}_{n,\alpha}^{S,L,D}).$$

The model selection criterion we use here to select the architecture of the network is a slight penalized variant of the classical ELBO criterion (Blei et al., 2017) with strong theoretical guarantees (Cherief-Abdellatif, 2019) :

$$(\hat{S}, \hat{L}, \hat{D}) = \underset{S,L,D}{\arg\max} \left\{ \mathcal{L}_n^*(S, L, D) - \log\left(\frac{1}{\pi_{S,L,D}}\right) \right\}.$$

For any choice of the prior beliefs $\pi_{S,L,D}$, compute the ELBO for each model $\mathcal{M}_{S,L,D}$ using an algorithm that will converge to $\mathcal{L}_n^*(S, L, D)$ and choose the architecture that maximizes the penalized ELBO criterion. It is possible to restrict to a finite number of layers in practice (for instance, a factor of n or $\log n$).

The following theorem shows that this ELBO criterion leads to a variational approximation with the optimal rate of convergence:

Theorem 5. *For any $\alpha \in (0, 1)$, for any S, L, D,*

$$\mathbb{E}\left[\int \|f_\theta - f_0\|_2^2 \tilde{\pi}_{n,\alpha}^{\hat{S},\hat{L},\hat{D}}(d\theta)\right] \leq \frac{2}{1-\alpha} \inf_{\theta^* \in \Theta_{S,L,D}} \|f_{\theta^*} - f_0\|_2^2 + \frac{2}{1-\alpha}\left(1 + \frac{\sigma^2}{\alpha}\right) r_n^{S,L,D} + \frac{2\sigma^2}{\alpha(1-\alpha)} \frac{\log(\frac{1}{\pi_{S,L,D}})}{n}.$$

This inequality shows that as soon as the complexity term $\log(1/\pi_{S,L,D})/n$ that reflects the prior beliefs is lower than the effective rate of convergence that balances the accuracy and the estimation error $r_n^{S,L,D}$, the selected variational approximation adaptively achieves the best possible rate. For instance, it leads to (near-)minimax rates for Hölder smooth functions and selects the optimal architecture even without the knowledge of β, which was required in the previous section. Note that for the previous choice of prior beliefs $\pi_L = 2^{-L}$, $\pi_{D|L} = 1/(\max(e^L, d) - d + 1)$, $\pi_{S|L,D} = 1/T$, we get:

$$\frac{\log(\frac{1}{\pi_{S,L,D}})}{n} \leq \frac{2\log(D+1) + \log L}{n} + \frac{\max(L, \log d) + L\log 2}{n}$$

that is lower than $r_n^{S,L,D}$ (up to a factor) and hence the ELBO criterion does not overfit.

6. Discussion

In this paper, we provided theoretical justifications for neural networks from a Bayesian point of view using sparse variational inference. We derived new generalization error bounds and we showed that sparse variational approximations of DNNs achieve (near-)minimax optimality when the regression function is Hölder smooth. All our results directly imply concentration of the approximation of the posterior distribution. We also proposed an automated method for selecting an architecture of the network with optimal consistency guarantees via the ELBO maximization framework.

We think that one of the main challenges here is the design of new computational algorithms for spike-and-slab deep learning in the wake of the work of (Titsias & Lázaro-Gredilla, 2011) for multi-task and multiple kernel learning, or those of (Louizos et al., 2018) and (Bellec et al., 2018). In the latter paper, the authors designed an algorithm for training deep networks while simultaneously learning their sparse connectivity allowing for fast and computationally efficient learning, whereas most approaches have focused on compressing already trained neural networks.

In the same time, a future point of interest is the study of the global convergence of these approximate algorithms in nonconvex settings i.e. study of the theoretical convergence of the ELBO. This work was conducted for frequentist gradient descent algorithms (Allen-Zhu et al., 2019; Du et al., 2019). Such studies should be investigated for Bayesian gradient descents, as well as for algorithms that preserve the sparsity of the network during training.

Acknowledgements

We would like to warmly thank Pierre Alquier for his helpful suggestions on early versions of this work.

References

Allen-Zhu, Z., Li, Y., and Song, Z. A convergence theory for deep learning via over-parameterization. In Chaudhuri, K. and Salakhutdinov, R. (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 242–252, Long Beach, California, USA, 09–15 Jun 2019. PMLR. URL http://proceedings.mlr.press/v97/allen-zhu19a.html.

Alquier, P. and Ridgway, J. Concentration of tempered posteriors and of their variational approximations. *arXiv preprint arXiv:1706.09293*, 2017.

Alquier, P., Ridgway, J., and Chopin, N. On the properties of variational approximations of Gibbs posteriors. *JMLR*, 17(239):1–41, 2016.

Baldi, P. and Hornik, K. Neural networks and principal component analysis: Learning from examples without local minima", ne. *Neural Networks*, 2:53–58, 12 1989. doi: 10.1016/0893-6080(89)90014-2.

Barron, A. Barron, a.e.: Universal approximation bounds for superpositions of a sigmoidal function. ieee trans. on information theory 39, 930-945. *Information Theory, IEEE Transactions on*, 39:930 – 945, 06 1993. doi: 10.1109/18.256500.

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

Bartlett, P. L., Foster, D. J., and Telgarsky, M. J. Spectrally-normalized margin bounds for neural networks. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 30*, pp. 6240–6249. Curran Associates, Inc., 2017. URL http://papers.nips.cc/paper/7204-spectrally-normalized-margin-bounds-for-neural-networks.pdf.

Behrens, G., Friel, N., and Hurn, M. Tuning tempered transitions. *Statistics and computing*, 22(1):65–78, 2012.

Bellec, G., Kappel, D., Maass, W., and Legenstein, R. Deep rewiring: Training very sparse deep networks. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=BJ_wN01C-.

Bengio, Y. and Delalleau, O. On the expressive power of deep architectures. In *Proceedings of the 22Nd International Conference on Algorithmic Learning Theory*, ALT'11, pp. 18–36, Berlin, Heidelberg, 2011. Springer-Verlag. ISBN 978-3-642-24411-7. URL http://dl.acm.org/citation.cfm?id=2050345.2050349.

Bhattacharya, A., Pati, D., and Yang, Y. Bayesian fractional posteriors. *arXiv preprint arXiv:1611.01125, to appear in the Annals of Statistics*, 2016.

Bhattacharya, A., Pati, D., and Yang, Y. On statistical optimality of variational Bayes. *Proceedings of Machine Learning Research*, 84 - AISTAT, 2018.

Blei, D. M., Kucukelbir, A., and McAuliffe, J. D. Variational inference: A review for statisticians. *Journal of the American Statistical Association*, 112(518):859–877, 2017.

Blundell, C., Cornebise, J., Kavukcuoglu, K., and Wierstra, D. Weight uncertainty in neural networks. In *Proceedings of the 32Nd International Conference on International Conference on Machine Learning - Volume 37*, ICML'15, pp. 1613–1622. JMLR.org, 2015. URL http://dl.acm.org/citation.cfm?id=3045118.3045290.

Buchholz, A., Wenzel, F., and Mandt, S. Quasi-Monte Carlo variational inference. In Dy, J. and Krause, A. (eds.), *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 668–677, Stockholmsmässan, Stockholm Sweden, 10–15 Jul 2018. PMLR. URL http://proceedings.mlr.press/v80/buchholz18a.html.

Campbell, T. and Li, X. Universal boosting variational inference. In Wallach, H., Larochelle, H., Beygelzimer, A., dAlché-Buc, F., Fox, E., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 32*, pp. 3479–3490. Curran Associates, Inc., 2019. URL http://papers.nips.cc/paper/8608-universal-boosting-variational-inference.pdf.

Castillo, I., Schmidt-Hieber, J., and van der Vaart, A. Bayesian linear regression with sparse priors. *Ann. Statist.*, 43(5):1986–2018, 10 2015. doi: 10.1214/15-AOS1334. URL https://doi.org/10.1214/15-AOS1334.

Catoni, O. *PAC-Bayesian supervised classification: the thermodynamics of statistical learning*. Institute of Mathematical Statistics Lecture Notes—Monograph Series, 56. Institute of Mathematical Statistics, Beachwood, OH, 2007.

Chérief-Abdellatif, B. and Alquier, P. Consistency of variational bayes inference for estimation and model selection in mixtures. *Electronic Journal of Statistics*, 12(2):2995–3035, 2018. ISSN 1935-7524. doi: 10.1214/18-EJS1475.

Cherief-Abdellatif, B.-E. Consistency of elbo maximization for model selection. In Ruiz, F., Zhang, C., Liang, D., and Bui, T. (eds.), *Proceedings of The 1st Symposium on Advances in Approximate Bayesian Inference*, volume 96 of *Proceedings of Machine Learning Research*, pp. 11–31. PMLR, 02 Dec 2019. URL http://proceedings.mlr.press/v96/cherief-abdellatif19a.html.

Chérief-Abdellatif, B.-E., Alquier, P., and Khan, M. A generalization bound for online variational inference. Preprint arXiv:1904.03920v1, 2019.

Cybenko, G. Approximation by superpositions of a sigmoidal function. *Mathematics of Control, Signals, and Systems (MCSS)*, 2(4):303–314, December 1989. ISSN 0932-4194. doi: 10.1007/BF02551274. URL http://dx.doi.org/10.1007/BF02551274.

Du, S., Lee, J., Li, H., Wang, L., and Zhai, X. Gradient descent finds global minima of deep neural networks. In Chaudhuri, K. and Salakhutdinov, R. (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 1675–1685, Long Beach, California, USA, 09–15 Jun 2019. PMLR. URL http://proceedings.mlr.press/v97/du19c.html.

Dziugaite, G. K. and Roy, D. M. Computing nonvacuous generalization bounds for deep (stochastic) neural networks with many more parameters than training data. In *Proceedings of the Thirty-Third Conference on Uncertainty in Artificial Intelligence, UAI 2017, Sydney, Australia, August 11-15, 2017*, 2017. URL http://auai.org/uai2017/proceedings/papers/173.pdf.

Gal, Y. *Uncertainty in Deep Learning*. PhD thesis, University of Cambridge, 2016.

Ghosal, S., Ghosh, J. K., and van der Vaart, A. W. Convergence rates of posterior distributions. *Ann. Statist.*, 28(2):500–531, 04 2000. doi: 10.1214/aos/1016218228. URL https://doi.org/10.1214/aos/1016218228.

Graves, A. Practical variational inference for neural networks. In Shawe-Taylor, J., Zemel, R. S., Bartlett, P. L., Pereira, F., and Weinberger, K. Q. (eds.), *Advances in Neural Information Processing Systems 24*, pp. 2348–2356. Curran Associates, Inc., 2011.

Grohs, P., Perekrestenko, D., Elbrächter, D., and Bölcskei, H. Deep neural network approximation theory. *arXiv preprint arXiv:1901.02220*, 01 2019.

Grünwald, P. D. and Van Ommen, T. Inconsistency of Bayesian inference for misspecified linear models, and a proposal for repairing it. *Bayesian Analysis*, 12(4): 1069–1103, 2017.

Guedj, B. A primer on pac-bayesian learning. *arXiv preprint arXiv:1901.05353*, 2019.

Hayakawa, S. and Suzuki, T. On the minimax optimality and superiority of deep neural network learning over sparse parameter spaces. *arXiv preprint arXiv:1905.09195*, 2019.

Hinton, G. E. and van Camp, D. Keeping the neural networks simple by minimizing the description length of the weights. In *Proceedings of the Sixth Annual Conference on Computational Learning Theory*, COLT '93, pp. 5–13, New York, NY, USA, 1993. ACM. ISBN 0-89791-611-5. doi: 10.1145/168304.168306. URL http://doi.acm.org/10.1145/168304.168306.

Huggins, J. H., Campbell, T., Kasprzak, M., and Broderick, T. Practical bounds on the error of bayesian posterior approximations: A nonasymptotic approach. *ArXiv*, abs/1809.09505, 2018.

Imaizumi, M. and Fukumizu, K. Deep neural networks learn non-smooth functions effectively. In Chaudhuri, K. and Sugiyama, M. (eds.), *Proceedings of Machine Learning Research*, volume 89 of *Proceedings of Machine Learning Research*, pp. 869–878. PMLR, 16–18 Apr 2019. URL http://proceedings.mlr.press/v89/imaizumi19a.html.

Jaiswal, P., Honnappa, H., and Rao, V. A. Risk-sensitive variational bayes: Formulations and bounds. *ArXiv*, arXiv:1906.01235, 2019a.

Jaiswal, P., Rao, V. A., and Honnappa, H. Asymptotic consistency of α-rényi-approximate posteriors. Preprint arXiv:1902.01902, 2019b.

Kawaguchi, K. Deep learning without poor local minima. In Lee, D. D., Sugiyama, M., Luxburg, U. V., Guyon, I., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 29*, pp. 586–594. Curran Associates, Inc., 2016. URL http://papers.nips.cc/paper/6112-deep-learning-without-poor-local-minima.pdf.

Kawaguchi, K., Huang, J., and Kaelbling, L. P. Effect of depth and width on local minima in deep learning. *Neural Computation*, 31(6):1462–1498, 2019.

Khan, M., Nielsen, D., Tangkaratt, V., Lin, W., Gal, Y., and Srivastava, A. Fast and scalable Bayesian deep learning by weight-perturbation in Adam. In Dy, J. and Krause, A. (eds.), *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 2611–2620, Stockholmsmässan, Stockholm Sweden, 10–15 Jul 2018. PMLR. URL http://proceedings.mlr.press/v80/khan18a.html.

Louizos, C., Welling, M., and Kingma, D. P. Learning sparse neural networks through l_0-regularization. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=H1Y8hhg0b.

MacKay, D. J. C. A practical bayesian framework for backpropagation networks. *Neural Computation*, 4 (3):448–472, 1992a. doi: 10.1162/neco.1992.4.3.448. URL https://doi.org/10.1162/neco.1992.4.3.448.

MacKay, D. J. C. *Bayesian methods for adaptive models*. PhD thesis, California Institute of Technology, 1992b.

Mishkin, A., Kunstner, F., Nielsen, D., Schmidt, M., and Khan, M. E. Slang: Fast structured covariance approximations for bayesian deep learning with natural gradient. In Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 31*, pp. 6245–6255. Curran Associates, Inc., 2018.

Neal, R. M. *Bayesian learning for neural networks*. PhD thesis, University of Toronto, 1995.

Neyshabur, B., Bhojanapalli, S., and Srebro, N. A PAC-bayesian approach to spectrally-normalized margin bounds for neural networks. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=Skz_WfbCZ.

Nguyen, Q., Mukkamala, M. C., and Hein, M. On the loss landscape of a class of deep neural networks with no bad local valleys. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=HJgXsjA5tQ.

Opper, M. and Archambeau, C. The variational gaussian approximation revisited. *Neural computation*, 21:786–92, 10 2008. doi: 10.1162/neco.2008.08-07-592.

Osawa, K., Swaroop, S., Khan, M. E., Jain, A., Eschenhagen, R., Turner, R. E., and Yokota, R. Practical deep learning with bayesian principles. In Wallach, H., Larochelle, H., Beygelzimer, A., dAlché-Buc, F., Fox, E., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 32*, pp. 4289–4301. Curran Associates, Inc., 2019.

Oymak, S. and Soltanolkotabi, M. Overparameterized nonlinear learning: Gradient descent takes the shortest path? In Chaudhuri, K. and Salakhutdinov, R. (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 4951–4960, Long Beach, California, USA, 09–15 Jun 2019. PMLR. URL http://proceedings.mlr.press/v97/oymak19a.html.

Petersen, P. and Voigtländer, F. Optimal approximation of piecewise smooth functions using deep relu neural networks. *Neural Networks*, 09 2017. doi: 10.1016/j.neunet.2018.08.019.

Rockova, V. and Polson, n. Posterior concentration for sparse deep learning. In Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 31*, pp. 930–941. Curran Associates, Inc., 2018.

Rolnick, D. and Tegmark, M. The power of deeper networks for expressing natural functions. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*, 2018. URL https://openreview.net/forum?id=SyProzZAW.

Schmidt-Hieber, J. Nonparametric regression using deep neural networks with relu activation function. *ArXiv*, arxiv:1708.06633, 2017.

Sheth, R. and Khardon, R. Excess risk bounds for the bayes risk using variational inference in latent gaussian models. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 30*, pp. 5151–5161. Curran Associates, Inc., 2017.

Soudry, D. and Carmon, Y. No bad local minima: Data independent training error guarantees for multilayer neural networks. *arXiv preprint arXiv:1605.08361*, 05 2016.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15:1929–1958, 2014. URL http://jmlr.org/papers/v15/srivastava14a.html.

Stanford, J., Giardina, K., Gerhardt, G., Fukumizu, K., and Amari, S.-i. Local minima and plateaus in hierarchical structures of multilayer perceptrons. *Neural Networks*, 13, 05 2000. doi: 10.1016/S0893-6080(00)00009-5.

Suzuki, T. Fast generalization error bound of deep learning from a kernel perspective. In Storkey, A. and Perez-Cruz, F. (eds.), *Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics*, volume 84 of *Proceedings of Machine Learning Research*, pp. 1397–1406, Playa Blanca, Lanzarote, Canary Islands, 09–11 Apr 2018. PMLR. URL http://proceedings.mlr.press/v84/suzuki18a.html.

Suzuki, T. Adaptivity of deep reLU network for learning in besov and mixed smooth besov spaces: optimal rate and curse of dimensionality. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=H1ebTsActm.

Titsias, M. K. and Lázaro-Gredilla, M. Spike and slab variational inference for multi-task and multiple kernel learning. In Shawe-Taylor, J., Zemel, R. S., Bartlett, P. L., Pereira, F., and Weinberger, K. Q. (eds.), *Advances in Neural Information Processing Systems 24*, pp. 2339–2347. Curran Associates, Inc., 2011.

Tonolini, F., Jensen, B. S., and Murray-Smith, R. Variational sparse coding, 2019. URL https://openreview.net/forum?id=SkeJ6iR9Km.

Tsybakov, A. B. *Introduction to Nonparametric Estimation*. Springer Publishing Company, Incorporated, 1st edition, 2008. ISBN 0387790519, 9780387790510.

Vladimirova, M., Verbeek, J., Mesejo, P., and Arbel, J. Understanding priors in Bayesian neural networks at the unit level. In Chaudhuri, K. and Salakhutdinov, R. (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 6458–6467, Long Beach, California, USA, 09–15 Jun 2019. PMLR. URL http://proceedings.mlr.press/v97/vladimirova19a.html.

Wang, Y. and Blei, D. M. Frequentist consistency of variational Bayes. Journal of the American Statistical Association (to appear), 2018.

Yarotsky, D. Error bounds for approximations with deep relu networks. *Neural Networks*, 94, 10 2016. doi: 10.1016/j.neunet.2017.07.002.

Zhang, C., Bengio, S., Hardt, M., Recht, B., and Vinyals, O. Understanding deep learning requires rethinking generalization. In *International Conference on Learning Representations*, 2017. URL https://openreview.net/forum?id=Sy8gdB9xx&noteId=Sy8gdB9xx.

Zhang, F. and Gao, C. Convergence rates of variational posterior distributions. *arXiv preprint arXiv:1712.02519v1*, 2017.