VARIATIONAL INFERENCE WITH THEORETICAL GUARANTEES

by

Jincheng Bai

A Dissertation

Submitted to the Faculty of Purdue University

In Partial Fulfillment of the Requirements for the degree of

Doctor of Philosophy

Department of Statistics

West Lafayette, Indiana

December 2020

THE PURDUE UNIVERSITY GRADUATE SCHOOL
STATEMENT OF COMMITTEE APPROVAL

Dr. Guang Cheng, Co-chair

Department of Statistics

Dr. Qifan Song, Co-chair

Department of Statistics

Dr. Vinayak Rao

Department of Statistics and Computer Science (by courtesy)

Dr. Faming Liang

Department of Statistics

Approved by:

Dr. Jun Xie

To my parents, Jingdong Bai and Wei Wang.

ACKNOWLEDGMENTS

I wish to express the deepest gratitude and appreciation to my advisors Prof. Qifan Song and Prof. Guang Cheng, for their invaluable guidance, encouragement and support throughout my PhD study. Prof. Cheng was the mentor who introduced me to the exciting field of machine learning research. Over the years, I gathered inspiration for research ideas from the discussions with Prof. Cheng. Prof. Song was my co-advisor who introduced me into the specific world of Bayesian learning. With uttermost patience, Prof. Song spent tremendous time and effort guiding and helping me, from tackling detailed research problems to writing research papers properly. I learned every aspects for being a qualified researcher from my advisors, without which it would be impossible for me to overcome the obstacles throughout my journey as a PhD student. Financially, I would like to thank Prof. Song for his generous support that allowed me to concentrate on my research.

I am grateful to Prof. Vinayak Rao and Prof. Faming Liang for serving as my committee members. I really appreciate them for providing insightful advice on my research projects. In particular, I would like to thank Prof. Rao for letting me attend his reading group and participating meaningful discussions. Besides, I also want to express my gratitude to all other faculty and staff members in the department of statistics, especially Prof. Jun Xie, for their great work and support throughout the years.

I would also like to thank all the members in Prof. Cheng's group, especially Botao Hao who introduced me to the group. The group meetings provided me the ideal platform to learn novel ideas and receive critical feedbacks from other group members. I also want to thank Ryan Murphy, Tim Keaton and Chuan Zuo for being my officemates. In addition, I owe many thanks to Zhou Shen, Boqian Zhang and Jiasen Yang for their advice on my career path. Besides, I want to thank all my friends and fellow graduate students, as well as everyone who helped and supported me. Without their care and support, my life at Purdue would not be such a journey of joy.

Last but not least, I would like to thank my parents Jingdong Bai and Wei Wang for their unconditional love and support. This dissertation is dedicated to them.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABBREVIATIONS

VI	variational inference
VB	variational Bayes
KL divergence	Kullback-Leibler divergence
ELBO	evidence lower bound
SGD	stochastic gradient descent
MFVB	mean-field variational Bayes
DNN	deep neural network
BNN	Bayesian neural network
CNN	convolutional neural network
ReLU	rectified linear unit
FDR	false discovery rate
TPR	true positive rate
RMSE	squared mean squared error

ABSTRACT

Variational inference (VI) or Variational Bayes (VB) is a popular alternative to MCMC, which doesn't scale well on complex Bayesian learning tasks with large datasets. Despite its huge empirical successes, the statistical properties of VI have not been carefully studied only until recently. In this dissertation, we are concerned with both the implementation and theoretical guarantee of VI.

In the first part of this dissertation, we propose a VI procedure for high-dimensional linear model inferences with heavy tail shrinkage priors, such as student-t prior. Theoretically, we establish the consistency of the proposed VI method and prove that under the proper choice of prior specifications, the contraction rate of the VB posterior is nearly optimal. It justifies the validity of VB inference as an alternative of MCMC sampling. Meanwhile, comparing to conventional MCMC methods, the VI procedure achieves much higher computational efficiency, which greatly alleviates the computing burden for modern machine learning applications such as massive data analysis. Through numerical studies, we demonstrate that the proposed VI method leads to shorter computing time, higher estimation accuracy, and lower variable selection error than competitive sparse Bayesian methods.

In the second part of this dissertation, we focus on sparse deep learning, which aims to address the challenge of huge storage consumption by deep neural networks, and to recover the sparse structure of target functions. We train sparse deep neural networks with a fully Bayesian treatment under two classes spike-and-slab priors, and develop sets of computationally efficient variational inferences via continuous relaxation of Bernoulli distribution. Given a pre-specified sparse DNN structure, the corresponding variational contraction rate is characterized that reveals a trade-off between the statistical estimation error, the variational error, and the approximation error, which are all determined by the network structural complexity (i.e., depth, width and sparsity). Note that the optimal network structure, which strikes the balance of the aforementioned trade-off and yields the best rate, is generally unknown. However, our methods could always achieve the best contraction rate as if the optimal network structure is known. In particular, when the true function is Hölder smooth, the variational inferences are capable to attain nearly minimax rate without

the knowledge of smoothness level. In addition, our empirical results demonstrate that the variational procedures provide uncertainty quantification in terms of Bayesian predictive distribution and are also capable to accomplish consistent variable selection by training a sparse multi-layer neural network.

1. INTRODUCTION

1.1 Variational Inference

Variational Inference (VI) or Variational Bayes (VB) (Jordan et al. 1999; Bishop 2006; Blei, Kucukelbir, et al. 2017) is an alternative to Markov Chain Monte Carlo (MCMC) for Bayesian learning. It approximates the true posterior distribution by a simpler family of distributions through an optimization problem.

Bayesian procedure makes statistical inferences from the posterior distribution $\pi(\theta|D) \propto \pi(\theta)p_\theta(D)$, where $\pi(\theta)$ is the prior distribution for θ, and $p_\theta(D)$ is the likelihood given θ and the dataset D. In the framework of variational inference, one seeks to find a good approximation of the posterior $\pi(\theta|D)$ via optimization rather than to simulate the posterior distribution by long-run MCMC. Given a variational family of distributions, denoted by $\mathcal{Q}$, the goal is to minimize the KL divergence between distributions in $\mathcal{Q}$ and true posterior distribution:

$$\widehat{q}(\theta) = \arg \min_{q(\theta)\in\mathcal{Q}} \mathrm{KL}(q(\theta)\|\pi(\theta|D)), \tag{1.1}$$

and the variational posterior $\widehat{q}(\theta)$ is subsequently used for approximated inference.

Unfortunately the optimization problem (1.1) is intractable, but we note that $\mathrm{KL}(q(\theta)\|\pi(\theta|D)) = C + \Omega$, where C is some constant depending on data D only, and

$$\Omega := -\mathbb{E}_{q(\theta)}[\log \frac{p_\theta(D)\pi(\theta)}{q(\theta)}]$$

is the so-called negative Evidence Lower Bound (ELBO). Then an equivalent optimization to (1.1) is

$$\widehat{q}(\theta) = \arg \min_{q(\theta)\in\mathcal{Q}} \Omega, \tag{1.2}$$

which is usually conducted via gradient descent type algorithms.

An inspiring representation of Ω is

$$\Omega = -\mathbb{E}_{q(\theta)}[\log p_\theta(D)] + \mathrm{KL}(q(\theta)\|\pi(\theta)), \tag{1.3}$$

where the first term in (1.3) can be viewed as the reconstruction error Kingma and Welling 2014 and the second term serves as regularization. Hence the variational inference procedure tends to be minimizing the reconstruction error while being penalized against prior distribution in the sense of KL divergence.

Alternative divergences Besides the KL divergence used in (1.1), some alternative divergences have also been considered: Minka 2001 proposed expectation propagation based on reciprocal KL divergence; Li et al. 2016; Jaiswal et al. 2019 investigated Rényi divergence, Dieng et al. 2017 considered χ^2 divergence. Those alternative variational inference may lead to better approximation, but could also cause difficulties in optimization or bring additional hyperparameters to tune. In this dissertation, we will stick to the KL-based variational inference.

1.1.1 Mean-field variational inference

For simplicity, it is commonly assumed that $\mathcal{Q}$ belongs to the mean-field family, i.e.

$$q(\theta) = \prod_{\mathrm{i}=1}^{T} q(\theta_{\mathrm{i}}).$$

Although conceptually simple and computationally convenient, the major drawback of mean-field variational interence is its inability to capture the covariance structure of the true posterior distribution. Specifically, it turns to underestimate the marginal posterior variance, which has long been observed in literature (Wang and Titterington 2004; Bishop 2006; Li et al. 2016; Wang and Blei 2019). Some attempts to correct the variance could be found in Giordano et al. 2015; 2018; Westling et al. 2019. Alternatively, beyond mean-field family, one could consider a structured variational family (Ranganath, Tran, et al. 2016) or use copula to model dependence (Tran et al. 2015). Since our primary goal is efficient point estimation, we will only consider the mean-field variational inference in the rest of this dissertation.

1.1.2 Stochastic optimization

Beyond the closed form coordinate ascent/descent update for variational inference with conditionally conjugate exponential families (Blei, Kucukelbir, et al. 2017), stochastic optimization has been widely used due to its flexibility with the choice of variational family and capability of handling large datasets. Hoffman et al. 2013 proposed traditional stochastic variational inference limited to conditionally conjugate models; Blei, Jordan, et al. 2012 introduced naive stochastic gradient estimator combined with a control variate approach to reduce the variance; Ranganath, Gerrish, et al. 2013 expressed the gradient as an expectation and then applied stochastic gradient descent with variance reduction.

More importantly, with the emergence of the reparameterization trick, stochastic variational inference is proposed for complex and deep generative models (Kingma and Welling 2014; Rezende et al. 2014), which paved the way for variational inference in deep learning. It is worth noting that the reparameterization trick could also help reduce Monte Carlo variance for nontrivial reasons (Rezende et al. 2014). Specifically, when the variational family is indexed by some hyperparameter ω, i.e., any $q \in \mathcal{Q}$ can be written as $q_\omega(\theta)$, then the negative ELBO is a function of ω as $\Omega(\omega)$. The KL divergence term in (1.3) could usually be integrated analytically, while the reconstruction error requires Monte Carlo estimation. Therefore, the optimization of $\Omega(\omega)$ can utilize the stochastic gradient approach (Kingma and Welling 2014). To be concrete, if all distributions in $\mathcal{Q}$ can be reparameterized as $q_\omega \overset{d}{=} g(\omega, \nu)$[1] for some differentiable function g and random variable ν, then the stochastic estimator of $\Omega(\omega)$ and its gradient are

$$\begin{aligned} \widetilde{\Omega}^m(\omega) &= -\frac{n}{m}\frac{1}{K}\sum_{\mathrm{i}=1}^{m}\sum_{k=1}^{K} \log p_{g(\omega,\nu_k)}(D_\mathrm{i}) + \mathrm{KL}(q_\omega(\theta)||\boldsymbol{\pi}(\theta)), \\ \nabla_\omega \widetilde{\Omega}^m(\omega) &= -\frac{n}{m}\frac{1}{K}\sum_{\mathrm{i}=1}^{m}\sum_{k=1}^{K} \nabla_\omega \log p_{g(\omega,\nu_k)}(D_\mathrm{i}) + \nabla_\omega \mathrm{KL}(q_\omega(\theta)||\boldsymbol{\pi}(\theta)), \end{aligned} \tag{1.4}$$

where D_i's are randomly sampled data points and ν_k's are iid copies of ν. Here, m and K are minibatch size and Monte Carlo sample size, respectively.

---

[1] "$\overset{d}{=}$" means equivalence in distribution

1.1.3 Theoretical developments

Despite its huge empirical successes, the statistical properties of VI have not been carefully studied only until recently. Early theoretical developments of variational inference centered around specific models by analyzing the iterative updating algorithms directly: You et al. 2014; Ormerod et al. 2017 studied Bayesian linear models; Hall, Ormerod, et al. 2011; Hall, Pham, et al. 2011 analyzed Poisson mixed-effects model; Celisse et al. 2012; Bickel et al. 2013 examined stochastic blockmodels.

Recently, some general frameworks for analyzing the theoretical properties of VI has been proposed: Westling et al. 2019 connected the consistency of VI to M-estimation; Wang and Blei 2019 established frequentist consistency and asymptotic normality of VB methods under LAN condition; Alquier et al. 2017; Pati et al. 2018; Gao et al. 2020; Yang, Pati, et al. 2020 examined the general conditions for deriving the variational contraction rate. Those general frameworks laid the foundation for our theoretical analyses under concrete models in this dissertation.

1.2 Deep Neural Networks

Deep Neural Networks (DNNs) have achieved tremendous successes in AI fields such as computer vision, natural language processing and reinforcement learning. One crucial factor for the successes of DNN is that it possesses highly complex and nonlinear model architecture, which allows it to approximate almost any complicated function Cybenko 1989; Mhasker et al. 2017; Rolnick et al. 2018.

1.2.1 Sparse neural networks

DNN may face various problems despite its huge successes. Large and deep fully connected networks are memory demanding (Srivastava et al. 2014) and also slow in inference for some real time tasks. Particularly, larger training sets and more complicated network structures improve accuracy in deep learning, but always incur huge storage and computation burdens. For example, small portable devices may have limited resources such as several megabyte memory, while a dense neural networks like ResNet-50 with 50 convolutional layers would need more than 95 megabytes of memory for storage and numerous floating number computation (Cheng et al. 2018). It is therefore necessary to compress deep learning models before deploying them on these hardware limited devices.

Meanwhile, sparse neural nets have been shown to have accurate approximation and strong generalization power (Glorot et al. 2011; Goodfellow et al. 2016). For example, the popular Dropout regularization Srivastava et al. 2014 could be interpreted as averaging over l_0 regularized sparse neural nets. From a nonparametric perspective, Schmidt-Hieber 2017 showed that sparse DNN with a ReLU activation function could achieve nearly minimax rate in the regression setup.

In addition, sparse neural networks may recover the potential sparsity structure of the target function, e.g., sparse teacher network in the teacher-student framework (Tian 2018; Goldt et al. 2019). Another example is from nonparametric regression with sparse target functions, i.e., only a portion of input variables are relevant to the response variable. A sparse network may serve the goal of variable selection (Feng et al. 2017; Liang et al. 2018; Ye et al. 2018), and is also known to be robust to adversarial samples against l_∞ and l_2 attacks (Guo et al. 2018).

1.2.2 Bayesian neural networks

Bayesian neural nets (BNN) are perceived to perform well against overfitting due to its regularization nature by enforcing a prior distribution. The study of Bayesian neural nets could date back to MacKay 1992, Neal 1992. Comparing to frequentist DNN, BNN possesses the advantages of robust prediction via model averaging and automatic uncertainty quantification (Blundell et al. 2015). Conceptually, BNN can easily induce sparse network selection by assigning discrete prior over all possible network structures. In particular, a spike-and-slab prior George et al. 1993 would switch a certain neuron off, and thus in nature imposes l_0 regularization and encourages network sparsity. Polson et al. 2018 introduced the Spike-and-Slab Deep Learning as a fully Bayesian alternative to Dropout for improving the generalizability of DNN with ReLU activation, where the posterior distribution is proven to concentrate at a nearly minimax rate.

However, a well-known obstacle for BNN is its high computational cost for drawing samples from posterior distribution via MCMC. Therefore, as a computationally efficient method, VI has been used widely for neural networks Graves 2011; Kingma and Welling 2014; Rezende et al. 2014; Blundell et al. 2015. Another challenge remains for sparse BNN is the lack of theoretical justification- the convergence property for variational BNN remains much less explored. Specifically, it would be interesting to examine whether the variational inference leads to the same rate of convergence compared to the Bayesian posterior distribution and frequentist estimators.

1.2.3 Related work

A plethora of methods on sparsifying or compressing neural networks have been proposed (Cheng et al. 2018; Gale et al. 2019). The majority of these methods are pruning-based (Han et al. 2016; Frankle et al. 2018; Zhu et al. 2018), which are ad-hoc on choosing the threshold of pruning and usually require additional training and fine tuning. Some other methods could achieve sparsity during training. For example, Louizos et al. 2018 introduced l_0 regularized learning and Mocanu et al. 2018 proposed sparse evolutionary training. However, the theoretical guarantee and the optimal choice of hyperparameters for these methods are unclear. As a more natural solution to enforce sparsity in DNN, Bayesian sparse neural network has been proposed by placing prior distributions on network weights: Blundell et al. 2015 and Deng et al. 2019 considered spike-and-slab priors with a Gaussian and Laplacian spike respectively; Log-uniform prior was used in Molchanov et al. 2017; Ghosh, Yao, et al. 2018 chose to use the popular horseshoe shrinkage prior. These existing works actually yield posteriors over the dense DNN model space despite applying sparsity induced priors. In order to derive explicit sparse inference results, users have to additionally determine certain pruning rules on the posterior. On the other hand, theoretical works regarding sparse deep learning have been studied in Schmidt-Hieber 2017, Polson et al. 2018 and Chérief-Abdellatif 2020, but finding an efficient implementation to close the gap between theory and practice remains a challenge for these mentioned methods.

1.2.4 Network structure

An L-hidden-layer neural network will be used to model the target function. The number of neurons in each hidden layer is denoted by p_{i} for $\mathrm{i} = 1, \ldots, L$. The weight matrix and bias parameter in each layer are denoted by $W_{\mathrm{i}} \in \mathbb{R}^{p_{\mathrm{i}-1} \times p_{\mathrm{i}}}$ and $b_{\mathrm{i}} \in \mathbb{R}^{p_{\mathrm{i}}}$ for $\mathrm{i} = 1, \ldots, L+1$. An example neural network is illustrated in Figure 1.1. Let $\sigma(x)$ be the activation function, and for any $r \in \mathbb{Z}^+$ and any $b \in \mathbb{R}^r$, we define $\sigma_b : \mathbb{R}^r \to \mathbb{R}^r$ as

$$\sigma_b \begin{bmatrix} y_1 \\ \vdots \\ y_r \end{bmatrix} = \begin{bmatrix} \sigma(y_1 - b_1) \\ \vdots \\ \sigma(y_r - b_r) \end{bmatrix}.$$

Then, given parameters $\boldsymbol{p} = (p_1, \ldots, p_L)$ and $\theta = \{W_1, b_1, \ldots, W_L, b_L, W_{L+1}, b_{L+1}\}$, the output of this DNN model can be written as

$$f_\theta(X) = W_{L+1}\sigma_{b_L}(W_L\sigma_{b_{L-1}} \ldots \sigma_{b_1}(W_1X)) + b_{L+1}. \tag{1.5}$$

In what follows, with slight abuse of notation, θ is also viewed as a vector that contains all the coefficients in W_{i}'s and b_{i}'s, , i.e., $\theta = (\theta_1, \ldots, \theta_T)$, where the length $T := \sum_{l=1}^{L-1} p_{l+1}(p_l + 1) + p_1(p+1) + (p_L + 1)$. The notation of the DNN will be used in the following chapters.

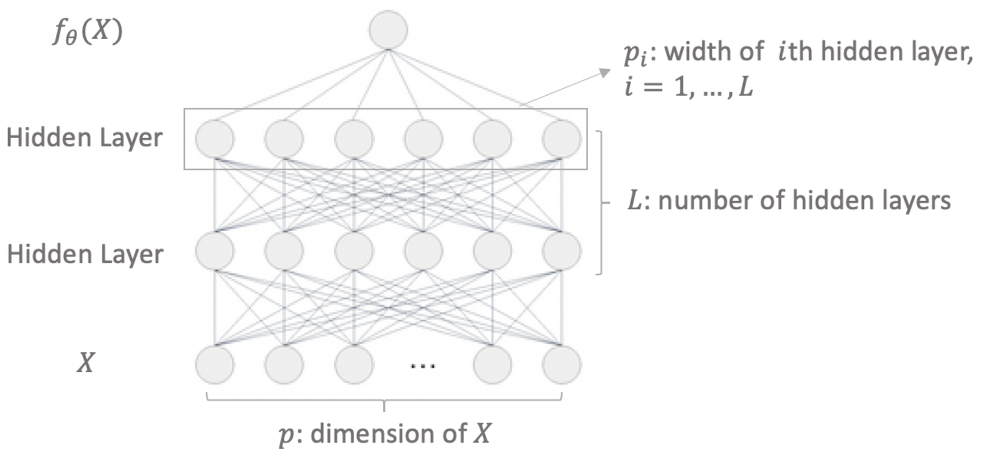

$f_\theta(X)$
p_i: width of ith hidden layer, $i = 1, \ldots, L$
Hidden Layer
L: number of hidden layers
Hidden Layer
X
…
p: dimension of X

Figure 1.1. Deep neural network

Instead of using a fully connected neural net, i.e., θ is a dense vector, we will consider a sparse NN $f_\theta \in \mathcal{F}(L, \boldsymbol{p}, s)$, where

$$\mathcal{F}(L, \boldsymbol{p}, s) = \{f_\theta \text{ as in } (1.5) : \|\theta\|_0 \leqslant s\},$$

$s \in \mathbb{N}$ controls the sparsity level of NN connectivity. The set of θ under the constraint $\mathcal{F}(L, \boldsymbol{p}, s)$ is denoted as $\Theta(L, \boldsymbol{p}, s)$.

1.3 Our Contribution and Dissertation Organization

1) Firstly, we propose a variational Bayesian (VB) procedure for high-dimensional linear model inferences with heavy tail shrinkage priors, such as student-t prior. Besides the superiority in computation efficiency, theory-wise we establish the consistency of the proposed VB method and prove that under the proper choice of prior specifications, the contraction rate of the VB posterior is nearly optimal. This part of work can be found in our paper Bai et al. 2020b.

2) Secondly, our work on sparse neural networks aims to resolve the aforementioned two important bottlenecks simultaneously by utilizing variational inference. On the computational side, it can reduce the ultra-high dimensional sampling problem of Bayesian computing, to an optimization task that can still be solved by a back-propagation algorithm. On the theoretical side, we provide a proper prior specification, under which the variational posterior distribution converges towards the truth. To the best of our knowledge, our work is the first one that provides a complete package of both theory and computation for sparse Bayesian DNN.

We achieve sparse deep learning by imposing a spike-and-slab prior (George et al. 1993; Ishwaran et al. 2005) on all the edges (weights and biases) of a neural network, where the spike component and slab component represent whether the corresponding edge is inactive or active, respectively. Our work distinguished itself from prior works on Bayesian sparse neural network by imposing the spike-and-slab prior with the Dirac spike function. Hence automatically, all posterior samples are from exact sparse DNN models. This part of work is published in our papers Bai et al. 2019; 2020a.

The rest of the dissertation will be organized as following: In Chapter 2, we apply variational inference to high dimensional linear regression problem under shrinkage priors. Chapter 3 and Chapter 4 focus on sparse deep learning via varitional infererence. In Chapter 3, the emphasis is on theoretical development. In Chapter 4, we further improve the computational efficiency as well as remaining theoretical validity via an alternative prior setting. Finally, all the results are summarized in Chapter 5 and the future directions are outlined.

2. HIGH DIMENSIONAL REGRESSION

2.1 Introduction

High dimensional sparse linear regression is one of the most commonly encountered problems in machine learning and statistics communities (Hastie et al. 2001). In the Bayesian paradigm, this problem is approached by placing sparsity-inducing priors on the regression coefficients. There are mainly two types of priors: the spike-and-slab prior (Mitchell et al. 1988; George et al. 1993; Ishwaran et al. 2005) and the shrinkage prior (Hans 2009; Carvalho et al. 2010; Griffin et al. 2012). The spike-and-slab prior has been considered as the gold standard for high dimensional linear regression, whose theoretical properties have been thoroughly studied (Johnson et al. 2012; Song and Liang 2014; Yang, Wainwright, et al. 2016; Gao et al. 2020). Although theoretically sound, the posterior sampling cost under spike-and-slab priors could be highly expensive, as it usually requires a tran-dimensional MCMC sampler such as reversible-jump MCMC. Alternatively, shrinkage priors could lead to equally good theoretical properties (Ghosal 1999; Armagan et al. 2013; Song and Liang 2017) while enjoying computational efficiency via the use of conjugate Gibbs sampler.

Although switching to shrinkage prior could reduce the computational burden to some extent, the nature of Bayesian computing (i.e., Markov chain Monte Carlo simulation) inevitably requires a huge number of iterations in order to achieve good mixing behavior and obtain accurate large-sample average. Consequently, people has sought to find frequentist shortcuts for Bayesian estimators. For example, Ročková et al. 2014 proposed EM algorithm to find posterior modes under the spike-and-slab prior; Ročková et al. 2018 obtained the posterior modes by using penalized likelihood estimation; Bhadra et al. 2019 searched posterior modes under horseshoe prior via optimization methods. Those approaches are computational-friendly, however completely ignore the distribution information of posterior and can not derive any Bayesian inferences beyond point estimation.

Another computationally convenient alternative to MCMC is the variational inference (VI or VB) (Jordan et al. 1999; Blei, Kucukelbir, et al. 2017). VI can provide an approximate posterior via frequentist optimization, thus it delivers (approximate) distributional inferences within a fairly small number of iterations. In the context of high dimensional linear regression, Carbonetto et al. 2012, Huang et al. 2016 and Ray et al. 2020 have proposed algorithms to carry out variational inferences under spike-and-slab priors. Besides their empirical successes, the theoretical properties were also justified. Specifically, Huang et al. 2016 showed their algorithm could achieve asymptotic consistency, and Ray et al. 2020 established the oracle inequalities for their VB approximation. Therefore, by employing the scalable variational inference, we would obtain the same theoretical guarantees as using MCMC while hugely reducing the computational cost.

In this chaper, we focus on the variational inference for Bayesian regression with shrinkage priors, which further improves the computational efficiency comparing to the one based on the spike-and-slab prior. Meanwhile, by showing the nearly optimal contraction rate of the proposed variational posterior, the validity of the proposed method is justified.

2.2 Preliminaries

2.2.1 High-dimensional Regression

Consider the linear regression model

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \sigma\boldsymbol{\epsilon}, \tag{2.1}$$

where $\boldsymbol{Y} \in \mathbb{R}^n$ is the response vector, $\boldsymbol{X} = (X_{\mathrm{ij}})$ is a $n \times p_n$ design matrix, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_{p_n}) \in \mathbb{R}^{p_n}$ is the coefficient vector and $\boldsymbol{\epsilon} \sim \mathcal{N}(0, I_n)$ is the Gaussian random noise. p_n denotes the dimension of coefficient parameter $\boldsymbol{\beta}$, and it can increase with the sample size n. The research objective is to make consistent variational Bayesian inferences on the coefficient $\boldsymbol{\beta}$. Note that we are particularly interested in the high dimensional setting, i.e. $p_n \gg n$, but our developed theory and methodology hold for general dimensional setting. For the simplicity of analysis, σ^2 is assumed to be known throughout our theoretical analysis, while in practice it can be estimated by frequentist methods (see Reid et al. 2016 for a comprehensive review), Empirical Bayesian approach (Castillo et al. 2015), or full Bayesian analysis (for example, placing inverse gamma prior on σ^2 (Ishwaran et al. 2005; Park et al. 2008)).

Let $\boldsymbol{\beta}^0$ denote the true coefficient vector, and we assume that $\boldsymbol{\beta}^0$ has certain sparsity structure. The corresponding true model is denoted as $\xi^0 = \{\mathrm{j} : \beta_{\mathrm{j}} \neq 0\}$, and true sparsity is denoted as $s = \|\boldsymbol{\beta}^0\|_0 = |\xi^0|$, which is the cardinality of the true subset model. Note that s is allowed to increase with n as well. Let $\xi \subseteq \{1, \ldots, p_n\}$ be the generic notation for any subset model, and $\boldsymbol{X}_\xi$ and $\boldsymbol{\beta}_\xi$ respectively denote the sub-matrix of $\boldsymbol{X}$ and sub-vector of $\boldsymbol{\beta}$ corresponding to ξ.

The following regularity conditions are required for the main results:

Condition 2.2.1 *The column norms of the design matrix are bounded by n, i.e.* $\sum_{\mathrm{i}} X_{\mathrm{ij}}^2 = \|\boldsymbol{X}_{\mathrm{j}}\|_2^2 \leqslant n$.

Condition 2.2.2 *There exist some integer $\overline{p}$ (depending on n and p_n) and fixed constant λ_0, such that $\overline{p} \succ s$*[1] *and the smallest eigenvalue of $\boldsymbol{X}_\xi^T \boldsymbol{X}_\xi$ is greater than $n\lambda_0$ for any subset model $|\xi| \leqslant \overline{p}$.*

Condition 2.2.3 $\log(\max_{\mathrm{j}} |\beta_{\mathrm{j}}^0|) = O(\log(p_n \vee n))$[2].

Remark: Condition 2.2.1 is trivially satisfied when the covariates X_{ij} are bounded by 1, or the design matrix is properly standardized. This bound condition is assumed for the technical simplicity, readers of interest can generalize this condition to that all covariates follow a sub-Gaussian distribution. Condition 2.2.2 imposes a regularity assumption on the eigen structure of the design matrix which controls the multicollinearity. Similar conditions are commonly used in the literature of high dimensional statistics (Zhang 2010; Narisetty et al. 2014; Song and Liang 2017). Under a random design scenario, if all entries of the design matrix are i.i.d. sub-Gaussian variables, then the random matrix theory (e.g., Vershynin 2012) guarantees that w.h.p., the eigen structure restriction holds with $\bar{p}$ being at least of order $n/\log p_n$, hence the condition $\overline{p} \succ s$ is met w.h.p. by assuming the common dimensionality condition $s \log p_n \prec n$. Condition 2.2.3 imposes an upper bound for the magnitude of true coefficients, it allows the magnitude of $\boldsymbol{\beta}$ increases polynomially with respect to $p_n \vee n$. Similar bounded conditions on true coefficient are common among Bayesian theoretical literature e.g., Yang, Wainwright, et al. 2016. Such conditions are necessary to ensure that the prior density around $\boldsymbol{\beta^0}$ is bounded away from zero, such that the domination of posterior around $\boldsymbol{\beta^0}$ becomes possible.

---

[1] $a_n \prec b_n$ means $\lim_n a_n/b_n = 0$.

[2] $a \vee b$ denotes $\max(a, b)$.

2.2.2 Heavy Tail Shrinkage Prior and Variational Inference

Prior Distribution

To resemble a spike-and-slab prior, a reasonable choice of shrinkage prior shall (1) allocate large probability mass around a small neighborhood of zero, i.e., a prior spike around 0; and (2) possess a very flat tail, i.e., a prior slab over real line. Following the suggestion by recent Bayesian literature (e.g., Ghosh and Chakrabarti 2015; Song and Liang 2017; Song 2020), our work will implement heavy-tailed prior distribution, i.e., polynomially decaying prior with properly tuning scale hyperparameter. For the simplicity of representation, this paper will only consider the theory and computation under student-t prior, however, the general insights obtained apply to any heavy tailed priors.

Consider an independent t prior for $\boldsymbol{\beta}$, which can be rewritten as a scaled mixture of Gaussian distribution with Inverse-Gamma scaling distributions, i.e., for $\mathrm{j} = 1, \dots, p_n$,

$$\pi(\beta_{\mathrm{j}}|\lambda_{\mathrm{j}}) = \mathcal{N}(0, \lambda_{\mathrm{j}}^{-1}), \quad \pi(\lambda_{\mathrm{j}}) = Gamma(a_0, b_n).$$

where a_0, b_n are user-specified hyperparameters. Thus, it yields a student-t prior of d.f. $2a_0$ with scale parameter $\sqrt{b_n/a_0}$. In other words, a_0 determines the polynomial degree of prior tail decay, i.e., the prior tail shape, while b_n controls the scale of prior distribution. As demonstrated by numerous Bayesian results (e.g., Van Der Pas, Kleijn, et al. 2014; Van Der Pas, Salomond, et al. 2016; Song 2020), the prior scale needs to converge to zero as dimensionality increases, hence we let a_0 be a constant, and b_n asymptotically decrease as n increases.

Variational Inference

In this chapter, we choose $\mathcal{Q}$ as independent student-t distribution to resemble the prior distribution, i.e.

$$q(\beta_{\mathrm{j}}|\lambda_{\mathrm{j}}) = \mathcal{N}(\mu_{\mathrm{j}}, \lambda_{\mathrm{j}}^{-1}), \quad q(\lambda_{\mathrm{j}}) = Gamma(a_{\mathrm{j}}, b_{\mathrm{j}}),$$

where $\mu_{\mathrm{j}} \in \mathbb{R}$, $a_{\mathrm{j}} > 0$, $b_{\mathrm{j}} > 0$ for $\mathrm{j} = 1, \dots, p_n$.

Remark: Choosing a different heavy tailed distribution as the prior distribution (e.g., horseshoe prior Carvalho et al. 2009) and variational family $\mathcal{Q}$ doesn't hurt the validity of the consistency result displayed in the next section, except that we require a different condition on the prior shape and scale hyperparameters. However, the difficulty of minimizing the negative ELBO varies from case to case, depending on the existence of closed form for the negative ELBO.

2.3 Theoretical Results

To establish consistency of variational Bayes posterior, we impose the following condition on the prior specification.

Condition 2.3.1 *$a_0 > 1$ and $(p_n \vee n)^{-K} \prec b_n/a_0 \prec s\log(p_n \vee n)/[np_n^{2+1/a_0}(p_n \vee n)^{\delta/a_0}]$ for some large constant K and small constant $\delta > 0$.*

$a_0 > 1$ ensures the existence of the second moment for the prior distribution, and the scale b_n is required to decrease polynomially w.r.t n and p_n, such that the prior contains a steep spike at 0.

First, we study the infimum of the negative ELBO Ω (up to a constant). Define the loglikelihood ratio as

$$l_n(P_0, P_{\boldsymbol{\beta}}) = \log\frac{p(\boldsymbol{Y}|\boldsymbol{\beta}^{\mathbf{0}})}{p(\boldsymbol{Y}|\boldsymbol{\beta})} = \sum_{\mathrm{i}=1}^{n}\log\frac{p(Y_{\mathrm{i}}|\boldsymbol{\beta}^{\mathbf{0}})}{p(Y_{\mathrm{i}}|\boldsymbol{\beta})},$$

then we have the following theorem.

Theorem 2.3.1 *With dominating probability for some $C > 0$, we have*

$$\inf_{q(\boldsymbol{\beta})\in\mathcal{Q}}\left\{KL(q(\boldsymbol{\beta})\|\boldsymbol{\pi}(\boldsymbol{\beta})) + \int l_n(P_0, P_{\boldsymbol{\beta}})q(\boldsymbol{\beta})d\boldsymbol{\beta}\right\} \leqslant Cs\log(p_n \vee n). \tag{2.2}$$

Remark: Theorem 2.3.1 establishes the upper bound of the loss function corresponding to the variational posterior.

Our next theorem studies how fast the variational posterior contrasts toward the true $\boldsymbol{\beta}^0$.

Theorem 2.3.2 *With dominating probability, for any slowly diverging sequence M_n, we have*

$$\widehat{q}(\|\boldsymbol{\beta}-\boldsymbol{\beta}^0\|_2 \geqslant M_n\sqrt{s\log(p_n \vee n)/n}) = o(1).$$

Remark: Theorem 2.3.2 implies that the contraction rate of the variational posterior $\widehat{q}(\boldsymbol{\beta})$ is of order $\sqrt{s\log(p_n \vee n)/n}$. Under low dimensional setting, it reduces to $\sqrt{s/n}\log^{0.5}(n)$ which is the optimal rate up to a logarithmic term; Under high dimensional setting, it reduces to $\sqrt{s\log(p_n)/n}$ which is the near-optimal convergence rate[3] commonly achieved in the literature. In other words, there is little loss in term of distributional convergence asymptotics by implementing variational approximation. The variational inference procedure delivers consistent Bayesian inferences.

2.4 Implementation

2.4.1 Updating Equations

The direct optimization of the negative ELBO requires stochastic gradient descent algorithm, since there is no closed form for the KL divergence between two student-t distributions. Therefore, for the purpose of efficient optimization, we instead consider minimizing the KL divergence of the joint distribution of $\boldsymbol{\beta}$ and $\boldsymbol{\lambda}$, where the negative ELBO is defined as

$$\Omega = -\int \log p(\boldsymbol{Y}|\boldsymbol{\beta},\lambda)q(\boldsymbol{\beta}|\boldsymbol{\lambda})q(\boldsymbol{\lambda})d\boldsymbol{\beta}d\boldsymbol{\lambda} + \int \mathrm{KL}(q(\boldsymbol{\beta}|\boldsymbol{\lambda})\|\pi(\boldsymbol{\beta}|\boldsymbol{\lambda}))q(\boldsymbol{\lambda})d\boldsymbol{\lambda} + \mathrm{KL}(q(\boldsymbol{\lambda})\|\pi(\boldsymbol{\lambda})). \tag{2.3}$$

As showed by the toy examples in the Appendix A, the variational inference results derived based on minimizing the KL divergence of the joint distribution of $(\boldsymbol{\beta},\boldsymbol{\lambda})$ has little difference to the ones based on minimizing the KL divergence of marginal distribution of $\boldsymbol{\beta}$.

---

[3] The optimal is of order $\sqrt{s\log(p_n/s)/n}$.

We minimize (2.3) by iteratively updating variational parameters in the fashion of coordinate descent. Specifically, the negative ELBO is

$$
\begin{aligned}
\Omega = & -\int \log p(\boldsymbol{Y}|\boldsymbol{\beta}, \lambda) q(\boldsymbol{\beta}|\boldsymbol{\lambda}) q(\boldsymbol{\lambda}) d\boldsymbol{\beta} d\boldsymbol{\lambda} + \int \mathrm{KL}(q(\boldsymbol{\beta}|\boldsymbol{\lambda}) \| \boldsymbol{\pi}(\boldsymbol{\beta}|\boldsymbol{\lambda})) q(\boldsymbol{\lambda}) d\boldsymbol{\lambda} + \mathrm{KL}(q(\boldsymbol{\lambda}) \| \boldsymbol{\pi}(\boldsymbol{\lambda})) \\
= & const + \int \Big\{ -\frac{\boldsymbol{Y}^T \boldsymbol{X} \mathbb{E}[\boldsymbol{\beta}|\boldsymbol{\lambda}]}{\sigma^2} + \frac{\mathbb{E}[\boldsymbol{\beta}^T X^T X \boldsymbol{\beta}|\boldsymbol{\lambda}]}{2\sigma^2} \Big\} q(\boldsymbol{\lambda}) d\boldsymbol{\lambda} + \sum_{\mathrm{j}=1}^{p_n} \int \Big[\log \frac{\lambda_{\mathrm{j}}}{\lambda_{\mathrm{j}}} + \frac{\lambda_{\mathrm{j}}^{-1} + \mu_{\mathrm{j}}^2}{2\lambda_{\mathrm{j}}^{-1}} \Big] q(\lambda_{\mathrm{j}}) d\lambda_{\mathrm{j}} \\
& + \sum_{\mathrm{j}=1}^{p_n} \Big[a_0 \log \frac{b_{\mathrm{j}}}{b_n} - \log \frac{\Gamma(a_{\mathrm{j}})}{\Gamma(a_0)} + (a_{\mathrm{j}} - a_0) \psi(a_{\mathrm{j}}) - (b_{\mathrm{j}} - b_n) \frac{a_{\mathrm{j}}}{b_{\mathrm{j}}} \Big] \\
= & const - \frac{\boldsymbol{Y}^T \boldsymbol{X} \boldsymbol{\mu}}{\sigma^2} + \frac{\boldsymbol{\mu}^T \boldsymbol{X}^T \boldsymbol{X} \boldsymbol{\mu}}{2\sigma^2} + \sum_{\mathrm{j}=1}^{p_n} \int \Big\{ \frac{n_{\mathrm{j}}}{2\sigma^2} \lambda_{\mathrm{j}}^{-1} \Big\} q(\lambda_{\mathrm{j}}) d\lambda_{\mathrm{j}} + \sum_{\mathrm{j}=1}^{p_n} \int \Big(\frac{\mu_{\mathrm{j}}^2 \lambda_{\mathrm{j}}}{2} \Big) q(\lambda_{\mathrm{j}}) d\lambda_{\mathrm{j}} \\
& + \sum_{\mathrm{j}=1}^{p_n} \Big[a_0 \log \frac{b_{\mathrm{j}}}{b_n} - \log \frac{\Gamma(a_{\mathrm{j}})}{\Gamma(a_0)} + (a_{\mathrm{j}} - a_0) \psi(a_{\mathrm{j}}) - (b_{\mathrm{j}} - b_n) \frac{a_{\mathrm{j}}}{b_{\mathrm{j}}} \Big],
\end{aligned}
\tag{2.4}
$$

where $\psi(x)$ is the digamma function, and $n_{\mathrm{j}} = [\boldsymbol{X}^T \boldsymbol{X}]_{\mathrm{j,j}}$. Therefore,

$$
\begin{aligned}
\Omega = & const - \frac{\boldsymbol{Y}^T \boldsymbol{X} \boldsymbol{\mu}}{\sigma^2} + \frac{\boldsymbol{\mu}^T \boldsymbol{X}^T \boldsymbol{X} \boldsymbol{\mu}}{2\sigma^2} + \frac{1}{2\sigma^2} \sum_{\mathrm{j}=1}^{p_n} \frac{n_{\mathrm{j}} b_{\mathrm{j}}}{a_{\mathrm{j}} - 1} + \sum_{\mathrm{j}=1}^{p_n} (\mu_{\mathrm{j}}^2/2 + b_n) \frac{a_{\mathrm{j}}}{b_{\mathrm{j}}} \\
& + \sum_{\mathrm{j}=1}^{p_n} \Big[a_0 \log \frac{b_{\mathrm{j}}}{b_n} - \log \frac{\Gamma(a_{\mathrm{j}})}{\Gamma(a_0)} + (a_{\mathrm{j}} - a_0) \psi(a_{\mathrm{j}}) - a_{\mathrm{j}} \Big],
\end{aligned}
$$

and the gradients are

$$
\begin{aligned}
\frac{d\Omega}{d\boldsymbol{\mu}} &= -\frac{\boldsymbol{X}^T \boldsymbol{Y}}{\sigma^2} + \frac{\boldsymbol{X}^T \boldsymbol{X} \boldsymbol{\mu}}{\sigma^2} + \boldsymbol{\Lambda} \mu, \\
\frac{d\Omega}{da_{\mathrm{j}}} &= -\frac{n_{\mathrm{j}}}{2\sigma^2} \frac{b_{\mathrm{j}}}{(a_{\mathrm{j}} - 1)^2} + \frac{\mu_{\mathrm{j}}^2/2 + b_n}{b_{\mathrm{j}}} + (a_{\mathrm{j}} - a_0) \psi_1(a_{\mathrm{j}}) - 1 \\
\frac{d\Omega}{db_{\mathrm{j}}} &= \frac{n_{\mathrm{j}}}{2\sigma^2} \frac{1}{a_{\mathrm{j}} - 1} - \frac{(\mu_{\mathrm{j}}^2/2 + b_n) a_{\mathrm{j}}}{b_{\mathrm{j}}^2} + \frac{a_0}{b_{\mathrm{j}}},
\end{aligned}
$$

where $\psi_1(x)$ is the trigamma function and $\boldsymbol{\Lambda} = \mathrm{diag}(a_1/b_1, \ldots, a_{p_n}/b_{p_n})$. Solve the above equations, we have

$$\begin{aligned}
\boldsymbol{\mu} &= (\boldsymbol{X}^T\boldsymbol{X} + \sigma^2\boldsymbol{\Lambda})^{-1}\boldsymbol{X}^T\boldsymbol{Y}, \\
a_{\mathrm{j}} &= solve(-\frac{n_{\mathrm{j}}}{2\sigma^2}\frac{b_{\mathrm{j}}}{(a_{\mathrm{j}}-1)^2} + \frac{\mu_{\mathrm{j}}^2/2 + b_n}{b_{\mathrm{j}}} + (a_{\mathrm{j}} - a_0)\psi_1(a_{\mathrm{j}}) - 1 = 0), \\
b_{\mathrm{j}} &= \frac{-a_0 + \sqrt{a_0^2 + 2n_{\mathrm{j}}a_{\mathrm{j}}(\mu_{\mathrm{j}}^2/2 + b_n)/\sigma^2(a_{\mathrm{j}}-1)}}{n_{\mathrm{j}}/\sigma^2(a_{\mathrm{j}}-1)}.
\end{aligned}$$

If σ is unknown, then the above derivation is modified as:

$$\begin{aligned}
\Omega =& const + n\log\sigma + \frac{\boldsymbol{Y}^T\boldsymbol{Y}}{2\sigma^2} - \frac{\boldsymbol{Y}^T\boldsymbol{X}\boldsymbol{\mu}}{\sigma^2} + \frac{\boldsymbol{\mu}^T\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{\mu}}{2\sigma^2} + \frac{1}{2\sigma^2}\sum_{\mathrm{j}=1}^{p_n}\frac{n_{\mathrm{j}}b_{\mathrm{j}}}{a_{\mathrm{j}}-1} + \sum_{\mathrm{j}=1}^{p_n}(\mu_{\mathrm{j}}^2/2 + b_n)\frac{a_{\mathrm{j}}}{b_{\mathrm{j}}} \\
&+ \sum_{\mathrm{j}=1}^{p_n}\Big[a_0\log\frac{b_{\mathrm{j}}}{b_n} - \log\frac{\Gamma(a_{\mathrm{j}})}{\Gamma(a_0)} + (a_{\mathrm{j}} - a_0)\psi(a_{\mathrm{j}}) - a_{\mathrm{j}}\Big],
\end{aligned}$$

and the additional partial derivative w.r.t. σ is

$$\frac{d\Omega}{d\sigma} = \frac{n}{\sigma} - \frac{(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\mu})^T(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\mu})}{\sigma^3} - \frac{1}{\sigma^3}\sum_{\mathrm{j}=1}^{p_n}\frac{n_{\mathrm{j}}b_{\mathrm{j}}}{a_{\mathrm{j}}-1}.$$

Thus, the updates of μ_{j}, a_{j} and b_{j} keep the same, and the update of σ follows

$$\sigma = \sqrt{\frac{(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\mu})^T(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\mu}) + \sum_{\mathrm{j}=1}^{p_n}\frac{n_{\mathrm{j}}b_{\mathrm{j}}}{a_{\mathrm{j}}-1}}{n}}.$$

To summarize, the updating equations are provided in below.

Updating $\boldsymbol{\mu}$ By fixing $a_{\mathrm{j}}s$ and $b_{\mathrm{j}}s$, the mean vector $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_{p_n})^T$ is updated by

$$\boldsymbol{\mu} = (\boldsymbol{X}^T\boldsymbol{X} + \sigma^2\boldsymbol{\Lambda})^{-1}\boldsymbol{X}^T\boldsymbol{Y}, \tag{2.5}$$

where $\boldsymbol{\Lambda} = \mathrm{diag}(a_1/b_1, \ldots, a_{p_n}/b_{p_n})$.

Updating a_{j} By fixing μ_{j} and b_{j}, a_{j} is updated by solving the following equation

$$-\frac{n_{\mathrm{j}}}{2\sigma^2}\frac{b_{\mathrm{j}}}{(a_{\mathrm{j}}-1)^2} + \frac{\mu_{\mathrm{j}}^2/2 + b_n}{b_{\mathrm{j}}} + (a_{\mathrm{j}} - a_0)\psi_1(a_{\mathrm{j}}) - 1 = 0. \tag{2.6}$$

Updating b_{j} By fixing μ_{j} and a_{j}, b_{j} is updated by

$$b_{\mathrm{j}} = \frac{-a_0 + \sqrt{a_0^2 + 2n_{\mathrm{j}}a_{\mathrm{j}}(\mu_{\mathrm{j}}^2/2 + b_n)/\sigma^2(a_{\mathrm{j}} - 1)}}{n_{\mathrm{j}}/\sigma^2(a_{\mathrm{j}} - 1)}. \tag{2.7}$$

2.4.2 Computation for Large p_n

The major computational bottleneck of the above updating rule is the inversion of the large $p_n \times p_n$ matrix $(\boldsymbol{X}^T\boldsymbol{X} + \sigma^2\boldsymbol{\Lambda})$ in (2.5), which could lead to huge computation cost.

Instead, (2.5) could be improved by using the blockwise update strategy introduced by Ishwaran et al. 2005. Specifically, decompose $\boldsymbol{\mu}$ as $(\boldsymbol{\mu}_{(1)}, \ldots, \boldsymbol{\mu}_{(B)})^T$, $\boldsymbol{\Lambda}$ as $\mathrm{diag}(\boldsymbol{\Lambda}_{(1)}, \ldots, \boldsymbol{\Lambda}_{(B)})$ and $\boldsymbol{X}$ as $[\boldsymbol{X}_{(1)}, \ldots, \boldsymbol{X}_{(B)}]$, where B is the number of blocks. Denote the exclusion of the kth block using subscript $(-k)$, then the blockwise update for $\boldsymbol{\mu}$ is

$$\boldsymbol{\mu}_{(k)} = (\boldsymbol{X}_{(k)}^T\boldsymbol{X}_{(k)} + \sigma^2\boldsymbol{\Lambda}_{(k)})^{-1}\boldsymbol{X}_{(k)}^T(\boldsymbol{Y} - \boldsymbol{X}_{(-k)}\boldsymbol{\mu}_{(-k)}), \tag{2.8}$$

for $k = 1, \ldots, B$. The blockwise update will reduce the order of computational complexity from $O(p_n^3)$ to $O(B^{-2}p_n^3)$ (ibid.), which could alleviate the computation burden when p_n is huge.

To summarize, the variational inference with Student-t prior is shown in Algorithm 1.

Algorithm 1 Variational inference with Student-t prior.

```
Hyperparameters: $a_0$, $b_n$
Initialize $\boldsymbol{\mu}$, $\{a_{\mathrm{j}}\}_{\mathrm{j}=1}^{p_n}$, $\{b_{\mathrm{j}}\}_{\mathrm{j}=1}^{p_n}$
repeat
    for $k = 1$ to $B$ do
        $\boldsymbol{\mu}_{(k)} \leftarrow$ apply equation (2.8)
    for all $\mathrm{j} \in \{1, \ldots, p_n\}$ do in parallel
        $a_{\mathrm{j}} \leftarrow$ solve equation (2.6)
        $b_{\mathrm{j}} \leftarrow$ apply equation (2.7)
    end for
    $\Omega \leftarrow \boldsymbol{\mu}, \{a_{\mathrm{j}}\}_{\mathrm{j}=1}^{p_n}, \{b_{\mathrm{j}}\}_{\mathrm{j}=1}^{p_n}$ using (2.3)
until convergence of $\Omega$
return $\boldsymbol{\mu}$, $\{a_{\mathrm{j}}\}_{\mathrm{j}=1}^{p_n}$, $\{b_{\mathrm{j}}\}_{\mathrm{j}=1}^{p_n}$
```

Note that it is crucial that the algorithm allows us to update the key variational parameter $\boldsymbol{\mu}$ blockwisely. Comparing to Algorithm 1 of Ray et al. 2020 which has to update variational parameter entrywisely, our algorithm has a much better convergence speed. In addition, Algorithm 1 of ibid. also has to conduct more iterations of univariate numerical optimizations. Therefore, as showed by our simulation studies, our algorithm has much faster computing speed.

2.5 Numerical Studies

In this section, we validate the effectiveness of our method via simulation experiments. To satisfy Condition 2.3.1, throughout this section, we let $a_0 = 2$ and $b_n/a_0 = \log(p_n \vee n)/[np_n^{2+1/a_0}(p_n \vee n)^{1/a_0}]$. We use Lasso estimator to initialize $\boldsymbol{\mu}$. a_{j} and b_{j} are initialized as $(a_0 + 0.5)$ and $(b_n + \mu_{\mathrm{j}}^2)$ respectively. The following rule is used to derive variable selection results: if the 95% credible interval of marginal t variational posterior contains 0, then the corresponding predictor is not selected, and vice versa. This method of Bayesian model selection under shrinkage priors is discussed by (Van Der Pas, Szabó, et al. 2017). More sophisticated approaches under variational Bayesian shrinkage for model selection could be a future study direction.

Both variational inference (t-VB) and MCMC (t-MCMC) are implemented under the same student-t prior for fair comparison, where t-MCMC is computed by Gibbs sampler (Song and Liang 2017). We also compare our method to the following competitive methods: variational Bayes for spike-and-slab priors with Laplace slabs (Laplace) (Ray et al. 2020), variational Bayes for spike-and-slab priors with Gaussian slabs (varbvs) (

2012), the spike-and-slab LASSO (SSLASSO) (Ročková et al. 2018), and the EM algorithm for spike-and-slab prior (EMVS) (Ročková et al. 2014).

For t-MCMC, we run Gibbs update for 1000 iterations with 200 burning in, and the initialization is the same as t-VB. We employ the blocklization (Ishwaran et al. 2005) for the Gibbs update. For Laplace, we use hyper-parameter $a_0 = 1, b_0 = n$ and $\lambda = 1$. The ridge estimator $(\boldsymbol{X}^T\boldsymbol{X} + \boldsymbol{I})^{-1}\boldsymbol{X}^T\boldsymbol{Y}$ is used for initialization and the unknown σ is estimated by selectiveInference package (Reid et al. 2016). For other methods, we use their associated R packages with default parameters. All the methods are implemented on the MacBook Pro with 2.7 GHz Intel Core i7.

The metrics reported are the Root Mean Squared Error between the posterior mean estimator $\widehat{\boldsymbol{\beta}}$ and $\boldsymbol{\beta}^0$ (RMSE), the False Discovery Rate (FDR), True Postitive Rate (TPR), and the run time. For Bayesian methods, the Coverage rates of 95% credible intervals for non-zero coefficients ξ^0 and zero coefficients $(\xi^0)^c$ are also calculated. All the experiments are repeated 100 times and the mean metric together with its standard deviation are reported.

2.5.1 Example 1: Moderate Dimension Case

This is an example similar to the one in Ray et al. 2020. Let $n = 100$, $p_n = 400$ and $s = 20$. All the nonzero coefficients are equal to $\log(n)$ (strong) or $\log(n)/2$ (weak) and their positions are randomly located within the p_n dimension coefficient vector. Take the design matrix $X_{\text{ij}} \overset{\text{iid}}{\sim} \mathcal{N}(0, 1)$ and assume σ is known that equals 4. Since p_n is moderate, we choose $B = 1$ when update $\boldsymbol{\mu}$ and use 5 blocks for Gibbs update.

Table 2.1 shows for relatively large signal, SSLASSO achieves the best estimation accuracy and the smallest selection error with the shortest run time, however it can not give second-order inferences. Among Bayesian methods, our method achieves estimation accuracy close to that of MCMC with the shortest run time. Meanwhile, the variable selection errors and the coverage rates of our method are also close to those of MCMC. Table 2.2 exhibits when the signal is relatively weak, our method obtains the estimation accurracy and selection error close to the best ones (Laplace) with much shorter run time. The MCMC is underperformed in this case probably due to insufficient number of Gibbs iterations. Note that the FDR for EMVS is undefined since none of the predictors is selected.

Table 2.2. Regression Results for Example 1 (b): Weak Signal Case.

	Bayesian				Non-Bayesian	
	t-VB	t-MCMC	Laplace	varbvs	SSLASSO	EMVS
RMSE	0.38 ± 0.07	0.46 ± 0.78	0.36 ± 0.07	0.45 ± 0.07	0.42 ± 0.10	0.46 ± 0.01
FDR	0.29 ± 0.15	0.12 ± 0.16	0.16 ± 0.15	0.12 ± 0.21	0.26 ± 0.29	-
TPR	0.57 ± 0.20	0.27 ± 0.10	0.62 ± 0.18	0.19 ± 0.22	0.47 ± 0.18	0.00 ± 0.00
Coverage of ξ^0	0.40 ± 0.05	0.22 ± 0.03	0.54 ± 0.04	0.14 ± 0.03	-	-
Coverage of $(\xi^0)^c$	0.99 ± 0.01	0.99 ± 0.00	0.99 ± 0.01	0.99 ± 0.00	-	-
Run time	0.54 ± 0.04	23.52 ± 0.38	13.78 ± 8.71	0.29 ± 0.13	0.07 ± 0.02	0.18 ± 0.02

Table 2.3. Regression Results for Example 2.

	Bayesian				Non-Bayesian	
	t-VB	t-MCMC	Laplace	varbvs	SSLASSO	EMVS
RMSE	0.01 ± 0.00	0.01 ± 0.00	0.06 ± 0.00	0.01 ± 0.00	0.01 ± 0.00	0.11 ± 0.00
FDR	0.00 ± 0.04	0.00 ± 0.00	0.07 ± 0.16	0.02 ± 0.07	0.00 ± 0.00	-
TPR	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	0.00 ± 0.00
Coverage of ξ^0	0.99 ± 0.01	0.95 ± 0.03	0.90 ± 0.03	0.16 ± 0.02	-	-
Coverage of $(\xi^0)^c$	1.00 ± 0.00	1.00 ± 0.00	0.99 ± 0.00	0.99 ± 0.00	-	-
Run time	0.48 ± 0.05	94.32 ± 3.40	37.61 ± 23.55	0.32 ± 0.11	0.75 ± 0.10	0.17 ± 0.01

2.5.2 Example 2: High Dimension Case

We consider an example similar to the one in Ročková et al. 2014. Let $n = 100$, $p = 1000$ and $\boldsymbol{\beta}^0 = (3, 2, 1, 0, \dots, 0)^T$. Generate the design matrix $X_{\mathrm{ij}} \overset{iid}{\sim} \mathcal{N}(0, 1)$. Assume $\sigma = 1$ and it is unknown in the experiment. For our method, we use the Empirical Bayes estimator for σ here. Specifically, by optimizing Ω w.r.t. σ, the Empirical Bayes (EB) update of σ follows

$$\sigma = \sqrt{\frac{(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\mu})^T(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\mu}) + \sum_{\mathrm{j}=1}^{p_n} \frac{n_{\mathrm{j}} b_{\mathrm{j}}}{a_{\mathrm{j}}-1}}{n}}. \tag{2.9}$$

Due to the high dimensionality, we choose $B = 10$ when updating $\boldsymbol{\mu}$ and also use 10 blocks for Gibbs update. The results are reported in Table 2.3.

Table 2.3 shows all the methods achieve good estimation accuracies expect Laplace and EMVS. Our method also obtains similar selection errors and coverage rates to those of MCMC with much shorter time. Again, the FDR for EMVS is undefined since all the estimated coefficients are not selected.

2.6 Conclusion and Discussion

We proposed a scalable variational inference algorithm for high dimensional linear regression under shrinkage priors. The established theoretical properties are justified by empirical studies. A possible future direction is to explore and compare efficient implementation for variational inference with other heavy tail shrinkage priors besides the Student-t.

2.7 Main Proofs

2.7.1 Proof of Theorem 2.3.1

Proof 1 The marginal prior distribution for β_{j} is

$$\pi(\beta_{\mathrm{j}}) = \frac{1}{\sqrt{\nu_0} s_0}\Big(1 + \nu_0^{-1}\Big(\frac{\beta_{\mathrm{j}}}{s_0}\Big)^2\Big)^{-\frac{\nu_0+1}{2}},$$

where $s_0 = \sqrt{b_n/a_0}$ and $\nu_0 = 2a_0$. We define $q^*(\beta_{\mathrm{j}})$ as follows

$$q^*(\beta_{\mathrm{j}}) = \frac{1}{\sqrt{\nu^*} s^*}\Big(1 + (\nu^*)^{-1}\Big(\frac{\beta_{\mathrm{j}} - \beta_{\mathrm{j}}^0}{s_{\mathrm{j}}}\Big)^2\Big)^{-\frac{\nu^*+1}{2}},$$

where $s^* = \sqrt{b_n/a_0}$ and $\nu^* = 2a_0$, and it is sufficient to show that $\mathrm{KL}(q^*(\boldsymbol{\beta}) \| \pi(\boldsymbol{\beta})) + \int l_n(P_0, P_{\boldsymbol{\beta}}) q^*(\boldsymbol{\beta}) d\boldsymbol{\beta} \leqslant C s \log(p_n \vee n)$.

i) We first show

$$\int l_n(P_0, P_{\boldsymbol{\beta}}) q^*(\boldsymbol{\beta}) d\boldsymbol{\beta} \leqslant C_1 s \log(p_n \vee n).,\tag{2.10}$$

for some $C_1 > 0$. Note that

$$\begin{aligned} l_n(P_0, P_{\boldsymbol{\beta}}) &= \frac{1}{2\sigma^2}(\|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2 - \|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}^0\|_2^2) \\ =& \frac{1}{2\sigma^2}(\|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}^0 + \boldsymbol{X}\boldsymbol{\beta}^0 - \boldsymbol{X}\boldsymbol{\beta})\|_2^2 - \|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}^0\|_2^2) \\ =& \frac{1}{2\sigma^2}(\|\boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{X}\boldsymbol{\beta}^0\|_2^2 + 2\langle \boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}^0, \boldsymbol{X}\boldsymbol{\beta}^0 - \boldsymbol{X}\boldsymbol{\beta}\rangle). \end{aligned}$$

Denote

$$\mathcal{R}_1 = \int \|\boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{X}\boldsymbol{\beta}^0\|_2^2 q^*(\boldsymbol{\beta})d\boldsymbol{\beta},$$
$$\mathcal{R}_2 = \int \langle \boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}^0, \boldsymbol{X}\boldsymbol{\beta}^0 - \boldsymbol{X}\boldsymbol{\beta}\rangle q^*(\boldsymbol{\beta})d\boldsymbol{\beta}.$$

Noting that $\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}^0 = \sigma\boldsymbol{\epsilon} \sim \mathcal{N}(0, \sigma^2 I_n)$, then

$$\begin{aligned}\mathcal{R}_2 &= \int \sigma\boldsymbol{\epsilon}^T(\boldsymbol{X}\boldsymbol{\beta}^0 - \boldsymbol{X}\boldsymbol{\beta})q^*(\boldsymbol{\beta})d\boldsymbol{\beta}\\ &= \sigma\boldsymbol{\epsilon}^T \int (\boldsymbol{X}\boldsymbol{\beta}^0 - \boldsymbol{X}\boldsymbol{\beta})q^*(\boldsymbol{\beta})d\boldsymbol{\beta} \sim \mathcal{N}(0, c_f\sigma^2),\end{aligned}$$

where $c_f = \| \int (\boldsymbol{X}\boldsymbol{\beta}^0 - \boldsymbol{X}\boldsymbol{\beta})q^*(\boldsymbol{\beta})d\boldsymbol{\beta}\|_2^2 \leqslant \mathcal{R}_1$ due to Cauchy-Schwarz inequality. Then by Gaussian tail bound

$$P_0(\mathcal{R}_2 \geqslant \mathcal{R}_1) \leqslant \exp(\frac{\mathcal{R}_1^2}{2\sigma^2\mathcal{R}_1}),$$

which implies $\mathcal{R}_2 \leqslant \mathcal{R}_1$ w.h.p.. Therefore, to prove (2.10) it suffices to establish that $\mathcal{R}_1 = O(s\log(p_n \vee n))$. Note that

$$\int \|\boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{X}\boldsymbol{\beta}^0\|_2^2 q^*(\boldsymbol{\beta})d\boldsymbol{\beta} \leqslant \|\boldsymbol{X}\|_2^2 \int \|\boldsymbol{\beta} - \boldsymbol{\beta}^0\|_2^2 q^*(\boldsymbol{\beta})d\boldsymbol{\beta}.$$

where $\|\boldsymbol{X}\|_2$ is the spectral norm of matrix $\boldsymbol{X}$. Since $\|\boldsymbol{X}\|_2^2 \leqslant tr(\boldsymbol{X}^T\boldsymbol{X}) = np_n$, and

$$\int \|\boldsymbol{\beta} - \boldsymbol{\beta}^0\|_2^2 q^*(\boldsymbol{\beta})d\boldsymbol{\beta} = \sum_{\text{j=1}}^{p_n} s^{*2}\frac{\nu^*}{\nu^* - 2} = p_n\frac{b_n}{a_0 - 1},$$

then

$$\int \|\boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{X}\boldsymbol{\beta}^0\|_2^2 q^*(\boldsymbol{\beta})d\boldsymbol{\beta} \leqslant np_n^2\frac{b_n}{a_0 - 1} = O(s\log(p_n \vee n)).$$

for sufficiently large n.

ii) We next show

$$\mathrm{KL}(q^*(\boldsymbol{\beta})\|\boldsymbol{\pi}(\boldsymbol{\beta})) \leqslant C_2 s\log(p_n \vee n), \tag{2.11}$$

for some $C_2 > 0$.

Note that

$$\begin{aligned}\mathrm{KL}(q^*(\boldsymbol{\beta})\|\boldsymbol{\pi}(\boldsymbol{\beta})) &= \sum_{\mathrm{j}=1}^{p_n} \mathrm{KL}(q^*(\beta_{\mathrm{j}})\|\boldsymbol{\pi}(\beta_{\mathrm{j}})) \\ &= \sum_{\mathrm{j}:\beta_{\mathrm{j}}^0 \neq 0} \mathrm{KL}(q^*(\beta_{\mathrm{j}})\|\boldsymbol{\pi}(\beta_{\mathrm{j}})).\end{aligned}$$

For each j,

$$\begin{aligned}&\mathrm{KL}(q^*(\beta_{\mathrm{j}})\|\boldsymbol{\pi}(\beta_{\mathrm{j}})) \\ =&\frac{\nu^*+1}{2}\int \log \frac{\nu^* s^{*2}+\beta_{\mathrm{j}}^2}{\nu^* s^{*2}+(\beta_{\mathrm{j}}-\beta_{\mathrm{j}}^0)^2} q^*(\beta_{\mathrm{j}})d\beta_{\mathrm{j}}.\end{aligned}$$

If $\beta_{\mathrm{j}}^0 > 0$, then $\frac{\nu^* s^{*2}+\beta_{\mathrm{j}}^2}{\nu^* s^{*2}+(\beta_{\mathrm{j}}-\beta_{\mathrm{j}}^0)^2}$ is maximized at $\widehat{\beta}_{\mathrm{j}} = \frac{\beta_{\mathrm{j}}^0+\sqrt{(\beta_{\mathrm{j}}^0)^2+4\nu^* s^{*2}}}{2}$, and the maximum is

$$\begin{aligned}\frac{\nu^* s^{*2}+\widehat{\beta}_{\mathrm{j}}^2}{\nu^* s^{*2}+(\widehat{\beta}_{\mathrm{j}}-\beta_{\mathrm{j}}^0)^2} &= \frac{(\beta_{\mathrm{j}}^0)^2+\beta_{\mathrm{j}}^0\sqrt{(\beta_{\mathrm{j}}^0)^2+4\nu^* s^{*2}}+4\nu^* s^{*2}}{(\beta_{\mathrm{j}}^0)^2-\beta_{\mathrm{j}}^0\sqrt{(\beta_{\mathrm{j}}^0)^2+4\nu^* s^{*2}}+4\nu^* s^{*2}} \\ \leqslant \frac{(\beta_{\mathrm{j}}^0)^2+\beta_{\mathrm{j}}^0\sqrt{(\beta_{\mathrm{j}}^0)^2+4\nu^* s^{*2}}+4\nu^* s^{*2}}{(\beta_{\mathrm{j}}^0)^2-\beta_{\mathrm{j}}^0(\beta_{\mathrm{j}}^0+\frac{4\nu^* s_0^2}{2\beta_{\mathrm{j}}^0})+4\nu^* s^{*2}} &= \frac{(\beta_{\mathrm{j}}^0)^2+\beta_{\mathrm{j}}^0\sqrt{(\beta_{\mathrm{j}}^0)^2+4\nu^* s^{*2}}+4\nu^* s^{*2}}{2\nu^* s^{*2}}.\end{aligned}$$

Therefore, for sufficiently large n,

$$\mathrm{KL}(q^*(\beta_{\mathrm{j}})\|\boldsymbol{\pi}(\beta_{\mathrm{j}})) \leqslant \frac{\nu^*+1}{2} \times O(\log(\beta_{\mathrm{j}}^0/s^*)) = O(\log(p_n \vee n)).$$

Similar result holds if $\beta_{\mathrm{j}}^0 < 0$ as well. This imples that $\mathrm{KL}(q^*(\boldsymbol{\beta})\|\boldsymbol{\pi}(\boldsymbol{\beta})) = O(s_n \log(p_n \vee n))$, and hence verifies (2.11).

Therefore, (2.2) immediately follows from (2.10) and (2.11). ■

The next lemma states the existence of testing condition. Define $\widetilde{p}$ as some sequence satisfying $s \leqslant \widetilde{p} \leqslant \bar{p} - s$, $\widetilde{p} \prec p_n$ and $\lim \widetilde{p} = \infty$. Let $\varepsilon_n = \sqrt{\widetilde{p}\log(p_n \vee n)/n}$. Denote B_n as the truncated parameter space

$$B_n = \{\boldsymbol{\beta} : \text{ at most } \widetilde{p} \text{ entries of } |\boldsymbol{\beta}/\sigma| \text{ is larger than } a_n\}$$

and

$$C_n = \{\boldsymbol{\beta} : B_n \cap \{\|\boldsymbol{\beta} - \boldsymbol{\beta}^0\|_2 \geqslant M_n\varepsilon_n\},$$

where $a_n \asymp \sqrt{s\log(p_n \vee n)/n}/p_n$, M_n is any diverging sequence as $M_n \to \infty$.

2.7.2 Proof of Theorem 2.3.2

Lemma 2.7.1 *There exists some testing function* $\phi_n \in [0,1]$ *and* $c_1 > 0$, $c_2 > 1/3$, *such that*

$$\mathbb{E}_{\beta^0}\phi_n \leqslant \exp(-c_1 n\varepsilon_n^2)$$
$$\sup_{\beta\in C_n}\mathbb{E}_{\beta}(1-\phi_n) \leqslant \exp(-c_2 nM_n^2\varepsilon_n^2)$$

Proof 2 The construction of the testing function is similar to that of Song and Liang 2017. Consider the following testing function

$$\phi_n = \max_{\{\xi\supset\xi^0, |\xi|\leqslant\widetilde{p}+s\}} 1\{\|(\boldsymbol{X}_\xi^T\boldsymbol{X}_\xi)^{-1}\boldsymbol{X}_\xi^T\boldsymbol{Y} - \boldsymbol{\beta}_\xi^0\|_2 \geqslant \sigma M\varepsilon_n\}$$

for some constant M.

i) For any ξ, such that $\xi \supset \xi^0$, $|\xi| \leqslant \widetilde{p} + s$,

$$\begin{aligned}&\mathbb{E}_{\beta^0}1\{\|(\boldsymbol{X}_\xi^T\boldsymbol{X}_\xi)^{-1}\boldsymbol{X}_\xi^T\boldsymbol{Y} - \boldsymbol{\beta}_\xi^0\|_2 \geqslant \sigma M\varepsilon_n\} = \mathbb{E}_{\beta^0}1\{\|(\boldsymbol{X}_\xi^T\boldsymbol{X}_\xi)^{-1}\boldsymbol{X}_\xi^T\boldsymbol{\epsilon}\|_2 \geqslant M\varepsilon_n\}\\ \leqslant & Pr(\|(\boldsymbol{X}_\xi^T\boldsymbol{X}_\xi)^{-1}\|_2(\boldsymbol{\epsilon}^T\boldsymbol{H}_\xi\boldsymbol{\epsilon}) \geqslant M^2\varepsilon_n^2) \leqslant Pr(\chi^2_{|\xi|} \geqslant n\lambda_0 M^2\varepsilon_n^2) \leqslant \exp(-c_1 M^2 n\varepsilon_n^2)\end{aligned}$$

for some constant c, where $\boldsymbol{H}_\xi = \boldsymbol{X}_\xi(\boldsymbol{X}_\xi^T\boldsymbol{X}_\xi)^{-1}\boldsymbol{X}_\xi^T$, and the last inequality is due to the sub-exponential properties of chi-square distribution and $|\xi| \ll n\epsilon_n^2$. This further implies that

$$\begin{aligned}\mathbb{E}_{\beta^0}\phi_n &\leqslant \sum_{\{\xi\supset\xi^0, |\xi|\leqslant\widetilde{p}+s\}} \mathbb{E}_{\beta^0}1\{\|(\boldsymbol{X}_\xi^T\boldsymbol{X}_\xi)^{-1}\boldsymbol{X}_\xi^T\boldsymbol{Y} - \boldsymbol{\beta}_\xi^0\|_2 \geqslant \sigma M\varepsilon_n\}\\ &\leqslant p_n^{\widetilde{p}+s}\exp(-c_1 M^2 n\varepsilon_n^2) \leqslant \exp(-c_1 n\varepsilon_n^2)\end{aligned}$$

when M is sufficiently large.

ii) Let $\widetilde{\xi} = \{k : |\beta_k/\sigma| > a_n\} \cup \xi^0$, then

$$
\begin{aligned}
&\sup_{\boldsymbol{\beta}\in C_n} \mathbb{E}_{\boldsymbol{\beta}}(1-\phi_n) = \sup_{\boldsymbol{\beta}\in C_n} \mathbb{E}_{\boldsymbol{\beta}} \min_{|\xi|\leqslant \widetilde{p}+s} 1\{\|(\boldsymbol{X}_{\xi}^T\boldsymbol{X}_{\xi})^{-1}\boldsymbol{X}_{\xi}^T\boldsymbol{Y} - \boldsymbol{\beta}^0_{\xi}\|_2 \leqslant \sigma\varepsilon_n\} \\
\leqslant &\sup_{\boldsymbol{\beta}\in C_n} \mathbb{E}_{\boldsymbol{\beta}} 1\{\|(\boldsymbol{X}_{\widetilde{\xi}}^T\boldsymbol{X}_{\widetilde{\xi}})^{-1}\boldsymbol{X}_{\widetilde{\xi}}^T\boldsymbol{Y} - \boldsymbol{\beta}^0_{\widetilde{\xi}}\|_2 \leqslant \sigma\varepsilon_n\} \\
= &\sup_{\boldsymbol{\beta}\in C_n} Pr\{\|(\boldsymbol{X}_{\widetilde{\xi}}^T\boldsymbol{X}_{\widetilde{\xi}})^{-1}\boldsymbol{X}_{\widetilde{\xi}}^T\boldsymbol{Y} - \boldsymbol{\beta}^0_{\widetilde{\xi}}\|_2 \leqslant \sigma\varepsilon_n\} \\
= &\sup_{\boldsymbol{\beta}\in C_n} Pr\{\|(\boldsymbol{X}_{\widetilde{\xi}}^T\boldsymbol{X}_{\widetilde{\xi}})^{-1}\boldsymbol{X}_{\widetilde{\xi}}^T\sigma\boldsymbol{\epsilon} + \boldsymbol{\beta}_{\widetilde{\xi}} + (\boldsymbol{X}_{\widetilde{\xi}}^T\boldsymbol{X}_{\widetilde{\xi}})^{-1}\boldsymbol{X}_{\widetilde{\xi}}^T\boldsymbol{X}_{\widetilde{\xi}^c}\boldsymbol{\beta}_{\widetilde{\xi}^c} - \boldsymbol{\beta}^0_{\widetilde{\xi}}\|_2 \leqslant \sigma\varepsilon_n\} \\
\leqslant &\sup_{\boldsymbol{\beta}\in C_n} Pr\{\|(\boldsymbol{X}_{\widetilde{\xi}}^T\boldsymbol{X}_{\widetilde{\xi}})^{-1}\boldsymbol{X}_{\widetilde{\xi}}^T\boldsymbol{\epsilon}\|_2 \geqslant (\|\boldsymbol{\beta}_{\widetilde{\xi}} - \boldsymbol{\beta}^0_{\widetilde{\xi}}\|_2 - \sigma\varepsilon_n - \|(\boldsymbol{X}_{\widetilde{\xi}}^T\boldsymbol{X}_{\widetilde{\xi}})^{-1}\boldsymbol{X}_{\widetilde{\xi}}^T\boldsymbol{X}_{\widetilde{\xi}^c}\boldsymbol{\beta}_{\widetilde{\xi}^c}\|_2)/\sigma\}.
\end{aligned}
$$

Note that $\|\boldsymbol{X}_{\widetilde{\xi}^c}\boldsymbol{\beta}_{\widetilde{\xi}^c}\|_2 \leqslant \sqrt{np_n}\|\boldsymbol{\beta}_{\widetilde{\xi}^c}\|_2 \leqslant \sqrt{np_n}\cdot\sqrt{p_n}\sigma a_n \leqslant c\sqrt{n}\sigma\varepsilon_n$ for some constant c, and

$$\|(\boldsymbol{X}_{\widetilde{\xi}}^T\boldsymbol{X}_{\widetilde{\xi}})^{-1}\boldsymbol{X}_{\widetilde{\xi}}^T\boldsymbol{X}_{\widetilde{\xi}^c}\boldsymbol{\beta}_{\widetilde{\xi}^c}\|_2/\sigma \leqslant \sqrt{\|(\boldsymbol{X}_{\widetilde{\xi}}^T\boldsymbol{X}_{\widetilde{\xi}})^{-1}\|_2}c\sqrt{n}\varepsilon_n \leqslant \sqrt{1/n\lambda_0}\sqrt{n}c\varepsilon_n \leqslant c\varepsilon_n/\sqrt{\lambda_0},$$

where the second inequality is due to $|\widetilde{\xi}| \leqslant \widetilde{p}+s \leqslant \overline{p}$. Besides,

$$\|\boldsymbol{\beta}_{\widetilde{\xi}} - \boldsymbol{\beta}^0_{\widetilde{\xi}}\|_2 \geqslant \|\boldsymbol{\beta} - \boldsymbol{\beta}^0\|_2 - \sqrt{p_n}\sigma a_n.$$

Therefore, $(\|\boldsymbol{\beta}_{\widetilde{\xi}} - \boldsymbol{\beta}^0_{\widetilde{\xi}}\|_2 - \sigma\varepsilon_n/2 - \|(\boldsymbol{X}_{\widetilde{\xi}}^T\boldsymbol{X}_{\widetilde{\xi}})^{-1}\boldsymbol{X}_{\widetilde{\xi}}^T\boldsymbol{X}_{\widetilde{\xi}^c}\boldsymbol{\beta}_{\widetilde{\xi}^c}\|_2)/\sigma \geqslant M_n\varepsilon_n/(2\sigma)$ when M_n is sufficiently large, and

$$\sup_{\boldsymbol{\beta}\in C_n} \mathbb{E}_{\boldsymbol{\beta}}(1-\phi_n) \leqslant \sup_{\boldsymbol{\beta}\in C_n} Pr\{\|(\boldsymbol{X}_{\widetilde{\xi}}^T\boldsymbol{X}_{\widetilde{\xi}})^{-1}\boldsymbol{X}_{\widetilde{\xi}}^T\boldsymbol{\epsilon}\|_2 \geqslant M_n\varepsilon_n/(2\sigma)\} \leqslant \exp(-c_2 n M_n^2\varepsilon_n^2).$$

■

As a technical tool, we restates the Donsker and Varadhan's representation for the KL divergence in the following lemma, whose proof can be found in Boucheron et al. 2013.

Lemma 2.7.2 *For any two probability measures P and Q, and any measurable function f such that $\int e^f dP < \infty$,*

$$\int f dQ \leqslant KL(Q\|P) + \log\int e^f dP.$$

The next two lemmas bound the contraction rate of $\widehat{q}(\boldsymbol{\beta})$ on B_n and B_n^c respectively.

Lemma 2.7.3 *With dominating probability,*

$$\widehat{q}(B_n \cap \{\|\boldsymbol{\beta} - \boldsymbol{\beta}^0\|_2 \geqslant M_n\varepsilon_n\}) = o(1),$$

where $\varepsilon_n = \sqrt{\widetilde{p}\log(p_n \vee n)/n}$ *and* M_n *is any diverging sequence as* $M_n \to \infty$*.*

Proof 3 We denote $\widetilde{\boldsymbol{\pi}}(\boldsymbol{\beta})$ and $\widetilde{q}(\boldsymbol{\beta})$ as the truncated distribution of $\boldsymbol{\pi}(\boldsymbol{\beta})$ and $\widehat{q}(\boldsymbol{\beta})$ on set B_n, i.e.

$$\begin{aligned}\widetilde{\boldsymbol{\pi}}(\boldsymbol{\beta}) &= \boldsymbol{\pi}(\boldsymbol{\beta})1(\boldsymbol{\beta} \in B_n)/\boldsymbol{\pi}(B_n),\\ \widetilde{q}(\boldsymbol{\beta}) &= \widehat{q}(\boldsymbol{\beta})1(\boldsymbol{\beta} \in B_n)/\widehat{q}(B_n).\end{aligned}$$

Define $V(P_{\boldsymbol{\beta}}, P_0) = M_n\varepsilon_n 1(\|\boldsymbol{\beta} - \boldsymbol{\beta}^0\|_2 \geqslant M_n\varepsilon_n)$ and

$$\log \eta(P_{\boldsymbol{\beta}}, P_0) = l_n(P_{\boldsymbol{\beta}}, P_0) + \frac{n}{3}V^2(P_{\boldsymbol{\beta}}, P_0).$$

Lemma 2.7.1 implies the existence of testing function within B_n and by the same argument used in Theorem 3.1 of Pati et al. 2018, it can be shown that w.h.p.,

$$\int_{B_n} \eta(P_{\boldsymbol{\beta}}, P_0)\widetilde{\boldsymbol{\pi}}(\boldsymbol{\beta})d\boldsymbol{\beta} \leqslant \mathrm{e}^{C_1 n\varepsilon_n^2}$$

for some $C_1 > 0$. By Lemma 2.7.2, it follows that w.h.p.,

$$\begin{aligned}&\frac{n}{3\widehat{q}(B_n)}M_n^2\varepsilon_n^2\widehat{q}(B_n \cap \{\|\boldsymbol{\beta} - \boldsymbol{\beta}^0\|_2 \geqslant M_n\varepsilon_n\})\\ =&\frac{n}{3\widehat{q}(B_n)}\int_{B_n} V^2(P_{\boldsymbol{\beta}}, P_0)\widehat{q}(\boldsymbol{\beta})d\boldsymbol{\beta}\\ =&\frac{n}{3}\int_{B_n} V^2(P_{\boldsymbol{\beta}}, P_0)\widetilde{q}(\boldsymbol{\beta})d\boldsymbol{\beta}\\ \leqslant&C_1 n\varepsilon_n^2 + \mathrm{KL}(\widetilde{q}(\boldsymbol{\beta})\|\widetilde{\boldsymbol{\pi}}(\boldsymbol{\beta})) - \int_{B_n} l_n(P_{\boldsymbol{\beta}}, P_0)\widetilde{q}(\boldsymbol{\beta})d\boldsymbol{\beta}.\end{aligned}$$

Noting that,

$$\begin{aligned}
&\mathrm{KL}(\widetilde{q}(\boldsymbol{\beta})\|\widetilde{\boldsymbol{\pi}}(\boldsymbol{\beta}))\\
=&\frac{1}{\widehat{q}(B_n)}\int_{B_n}\log\frac{\widehat{q}(\boldsymbol{\beta})}{\boldsymbol{\pi}(\boldsymbol{\beta})}\widehat{q}(\boldsymbol{\beta})d\boldsymbol{\beta}+\log\frac{\boldsymbol{\pi}(B_n)}{\widehat{q}(B_n)}\\
=&\frac{1}{\widehat{q}(B_n)}\mathrm{KL}(\widehat{q}(\boldsymbol{\beta})\|\boldsymbol{\pi}(\boldsymbol{\beta}))-\frac{1}{\widehat{q}(B_n)}\int_{B_n^c}\log\frac{\widehat{q}(\boldsymbol{\beta})}{\boldsymbol{\pi}(\boldsymbol{\beta})}\widehat{q}(\boldsymbol{\beta})d\boldsymbol{\beta}+\log\frac{\boldsymbol{\pi}(B_n)}{\widehat{q}(B_n)},
\end{aligned}$$

and similarly,

$$\int_{B_n} l_n(P_{\boldsymbol{\beta}},P_0)\widetilde{q}(\boldsymbol{\beta})d\boldsymbol{\beta}=\frac{1}{\widehat{q}(B_n)}\int l_n(P_{\boldsymbol{\beta}},P_0)\widehat{q}(\boldsymbol{\beta})d\boldsymbol{\beta}-\frac{1}{\widehat{q}(B_n)}\int_{B_n^c} l_n(P_{\boldsymbol{\beta}},P_0)\widehat{q}(\boldsymbol{\beta})d\boldsymbol{\beta}.$$

Combine the above three inequalities, we obtain that

$$\begin{aligned}
&M_n^2\varepsilon_n^2\widehat{q}(B_n\cap\{\|\boldsymbol{\beta}-\boldsymbol{\beta}^0\|_2\geqslant M_n\varepsilon_n\})\\
\leqslant& C\widehat{q}(B_n)\varepsilon_n^2+\frac{3}{n}\Big\{\mathrm{KL}(\widehat{q}(\boldsymbol{\beta})\|\boldsymbol{\pi}(\boldsymbol{\beta}))-\int l_n(P_{\boldsymbol{\beta}},P_0)\widehat{q}(\boldsymbol{\beta})d\boldsymbol{\beta}\Big\}\\
&+\frac{3}{n}\int_{B_n^c} l_n(P_{\boldsymbol{\beta}},P_0)\widehat{q}(\boldsymbol{\beta})d\boldsymbol{\beta}+\frac{3}{n}\int_{B_n^c}\log\frac{\boldsymbol{\pi}(\boldsymbol{\beta})}{\widehat{q}(\boldsymbol{\beta})}\widehat{q}(\boldsymbol{\beta})d\boldsymbol{\beta}+\frac{3\widehat{q}(B_n)}{n}\log\frac{\boldsymbol{\pi}(B_n)}{\widehat{q}(B_n)}. \qquad (2.12)
\end{aligned}$$

By Theorem 2.3.1, the second term in the RHS of (2.12) is bounded by $3\varepsilon_n^2$.

Apply the similar argument used in the proof of Theorem 2.3.1, the third term in the RHS of (2.12) is bounded by

$$\begin{aligned}
&\frac{3}{n}\int_{B_n^c} l_n(P_{\boldsymbol{\beta}},P_0)\widehat{q}(\boldsymbol{\beta})d\boldsymbol{\beta}\\
=&\frac{3}{2n\sigma^2}\Big\{-2\sigma\boldsymbol{\epsilon}^T\int_{B_n^c}(\boldsymbol{X}\boldsymbol{\beta}^0-\boldsymbol{X}\boldsymbol{\beta})\widehat{q}(\boldsymbol{\beta})d\boldsymbol{\beta}-\int_{B_n^c}\|\boldsymbol{X}\boldsymbol{\beta}^0-\boldsymbol{X}\boldsymbol{\beta}\|_2^2\widehat{q}(\boldsymbol{\beta})d\boldsymbol{\beta}\Big\}.
\end{aligned}$$

Note that $-2\sigma\boldsymbol{\epsilon}^T\int_{B_n^c}(\boldsymbol{X}\boldsymbol{\beta}^0-\boldsymbol{X}\boldsymbol{\beta})\widehat{q}(\boldsymbol{\beta})d\boldsymbol{\beta}$ follows a normal distribution $\mathcal{N}(0,V^2)$, where $V^2=4\sigma^2\|\int_{B_n^c}(\boldsymbol{X}\boldsymbol{\beta}^0-\boldsymbol{X}\boldsymbol{\beta})\widehat{q}(\boldsymbol{\beta})d\boldsymbol{\beta}\|^2\leqslant 4\sigma^2\int_{B_n^c}\|\boldsymbol{X}\boldsymbol{\beta}^0-\boldsymbol{X}\boldsymbol{\beta}\|_2^2\widehat{q}(\boldsymbol{\beta})d\boldsymbol{\beta}$. Thus the third term in the RHS of (2.12) is bounded by

$$\frac{3}{2n\sigma^2}\left[\mathcal{N}(0,V^2)-\frac{V^2}{4\sigma^2}\right]. \qquad (2.13)$$

Noting that $\mathcal{N}(0, V^2) = O_p(G_n V)$ for any diverging sequence G_n, (2.13) is further bounded, w.h.p., by

$$\frac{3}{2n\sigma^2}(G_n V - \frac{V^2}{4\sigma^2}) \leqslant \frac{3}{2n\sigma^2}\sigma^2 G_n^2.$$

Therefore, the third term in the RHS of (2.12) can be bounded by ε_n^2 w.h.p. (by choosing $G_n^2 \asymp n\varepsilon_n^2$).

The fourth term in the RHS of (2.12) is bounded by

$$\frac{3}{n}\int_{B_n^c} \log \frac{\pi(\boldsymbol{\beta})}{\widehat{q}(\boldsymbol{\beta})}\widehat{q}(\boldsymbol{\beta})d\boldsymbol{\beta} \leqslant \frac{3}{n}\widehat{q}(B_n^c)\log\frac{\pi(B_n^c)}{\widehat{q}(B_n^c)} \leqslant \frac{3}{n}\sup_{x\in(0,1)}[x\log(1/x)] = O(1/n).$$

Similarly, the fifth term in the RHS of (2.12) is bounded by $O(1/n)$.

Therefore, we have that w.h.p.,

$$M_n^2\varepsilon_n^2\widehat{q}(B_n \cap \{\|\boldsymbol{\beta} - \boldsymbol{\beta}^0\|_2 \geqslant M_n\varepsilon_n\}) \leqslant C\widehat{q}(B_n)\varepsilon_n^2 + 3\varepsilon_n^2 + \varepsilon_n^2 + 1/n,$$

that is, $\widehat{q}(B_n \cap \{\|\boldsymbol{\beta} - \boldsymbol{\beta}^0\|_2 \geqslant M_n\varepsilon_n\}) = O_p(1/M_n^2) = o_p(1)$. ■

Lemma 2.7.4 *With dominating probability,* $\widehat{q}(B_n^c) = o(1)$.

Proof 4 By Theorem 2.3.1, we have that w.h.p.,

$$\begin{aligned}&\mathrm{KL}(\widehat{q}(\boldsymbol{\beta})\|\pi(\boldsymbol{\beta})) + \int l_n(P_0, P_{\boldsymbol{\beta}})\widehat{q}(\boldsymbol{\beta})d\boldsymbol{\beta} = \inf_{q(\boldsymbol{\beta})\in\mathcal{Q}}\Big\{\mathrm{KL}(q(\boldsymbol{\beta})\|\pi(\boldsymbol{\beta})) + \int l_n(P_0, P_{\boldsymbol{\beta}})q(\boldsymbol{\beta})(d\boldsymbol{\beta})\Big\}\\ \leqslant &Cn\varepsilon_n^2,\end{aligned}$$

where C is some constant. By the similar argument used in the proof of Theorem 2.3.1 in the main text,

$$\int l_n(P_0, P_{\boldsymbol{\beta}})\widehat{q}(\boldsymbol{\beta})d\boldsymbol{\beta} \leqslant \frac{1}{2\sigma_\epsilon^2}\left(\int ||\boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{X}\boldsymbol{\beta}^{\mathbf{0}}||_2^2\widehat{q}(\boldsymbol{\beta})(d\boldsymbol{\beta}) + Z\right)$$

where Z is a normal distributed $\mathcal{N}(0, \sigma^2 c_0)$, where $c_0 \leqslant c_0 = \int ||\boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{X}\boldsymbol{\beta}^{\mathbf{0}}||_2^2\widehat{q}(\boldsymbol{\beta})(d\boldsymbol{\beta})$. Therefore, $-\int l_n(P_0, P_{\boldsymbol{\beta}})\widehat{q}(\boldsymbol{\beta})d\boldsymbol{\beta} = (1/2\sigma^2)[-c_0 + O_p(\sqrt{c_0})]$, and $\mathrm{KL}(\widehat{q}(\boldsymbol{\beta})\|\pi(\boldsymbol{\beta})) \leqslant Cn\varepsilon_n^2 + (1/2\sigma^2)[-c_0 + O_p(\sqrt{c_0})] = O_p(n\epsilon_n^2)$.

For any $\beta_{\mathrm{j}} \sim \widehat{q}(\beta_{\mathrm{j}})$, define $\gamma_{\mathrm{j}} = 1(|\beta_{\mathrm{j}}/\sigma| > a_n)$, then

$$\begin{aligned} &\mathrm{KL}(\widehat{q}(\boldsymbol{\beta}) \| \boldsymbol{\pi}(\boldsymbol{\beta})) \geqslant \mathrm{KL}(\widehat{q}(\boldsymbol{\gamma}) \| \boldsymbol{\pi}(\boldsymbol{\gamma})) \\ =&\sum_{\mathrm{j}=1}^{p_n} \Big[\widehat{q}(\gamma_{\mathrm{j}} = 1) \log \frac{\widehat{q}(\gamma_{\mathrm{j}} = 1)}{\boldsymbol{\pi}(\gamma_{\mathrm{j}} = 1)} + \widehat{q}(\gamma_{\mathrm{j}} = 0) \log \frac{\widehat{q}(\gamma_{\mathrm{j}} = 0)}{\boldsymbol{\pi}(\gamma_{\mathrm{j}} = 0)} \Big]. \end{aligned} \tag{2.14}$$

Choose $\alpha_0 = p_n^{-1}$ and let $A = \{\mathrm{j} : \widehat{q}(\gamma_{\mathrm{j}} = 1) \geqslant \alpha_0\}$, and denote $\alpha = \boldsymbol{\pi}(\gamma_{\mathrm{j}} = 1)$. Noting that by the condition of a_0 and b_n, we can obtain that

$$\alpha = \boldsymbol{\pi}(\gamma_{\mathrm{j}} = 1) \asymp \alpha_0/(n \vee p_n)^{\delta},$$

thus (2.14) implies $\sum_{\mathrm{j}\in A} \widehat{q}(\gamma_{\mathrm{j}} = 1) \log(\alpha_0/\alpha) \leqslant Cn\varepsilon_n^2/2$ for some C and $\sum_{\mathrm{j}\in A} \widehat{q}(\gamma_{\mathrm{j}} = 1) = O(\widetilde{p})$.

Under $\widehat{q}$, by Markov inequality,

$$Pr(\sum_{\mathrm{j}\in A} \gamma_{\mathrm{j}} \geqslant \widetilde{p}/2) \leqslant Pr(\sum_{\mathrm{j}\in A} \gamma_{\mathrm{j}} \geqslant \widetilde{p}/3 + \mathbb{E}\sum_{\mathrm{j}\in A} \gamma_{\mathrm{j}}) \leqslant 9\mathrm{Var}(\sum_{\mathrm{j}\in A} \gamma_{\mathrm{j}})/\widetilde{p}^2 \leqslant 9\mathbb{E}\sum_{\mathrm{j}\in A} \gamma_{\mathrm{j}}/\widetilde{p}^2 = o(1).$$

by Chernoff bound,

$$\begin{aligned} &Pr(\sum_{\mathrm{j}\notin A} \gamma_{\mathrm{j}} \geqslant \widetilde{p}/2) \leqslant Pr(Bin(p_n, \alpha_0) \geqslant \widetilde{p}/2) \\ \leqslant &\exp\Big\{-p_n\Big(\frac{\widetilde{p}}{2p_n} \log \frac{\widetilde{p}/(2p_n)}{\alpha_0} + \Big(1 - \frac{\widetilde{p}}{2p_n}\Big) \log \frac{1 - \widetilde{p}/(2p_n)}{1 - \alpha_0}\Big)\Big\} \leqslant \exp(-c\widetilde{p}) = o(1), \end{aligned}$$

for some constant c, since $\widetilde{p} \to \infty$.

Combine the above results together, it is trivial to conclude that $\widehat{q}(B_n^c) = o(1)$. ■

Proof of Theorem 2.3.2

Proof 5 Trivially combine Lemmas 2.7.3 and 2.7.4, we obtain that $\widehat{q}(\{\|\boldsymbol{\beta} - \boldsymbol{\beta}^0\|_2 \geqslant M_n\varepsilon_n\}) = o_p(1)$ for any diverging M_n, where $\varepsilon_n = \sqrt{\widetilde{p}\log(p_n \vee n)/n}$. Due to the arbitrariness of M_n and $\widetilde{p}$, we can let $M_n\sqrt{\widetilde{p}/s} \leqslant M_n$, and the theorem naturally holds. ■

3. SPARSE DEEP LEARNING

3.1 Introduction

In this chapter, we mainly focus on investigating the theoretical behavior of variational posterior for Bayesian DNN under spike-and-slab modeling. Our specific goals are to understand how fast the variational posterior converges to the truth and how accurate the prediction carried out by variational inferences is. It is not surprising that the choice of the network structure, i.e., network depth, width and sparsity level, plays a crucial role for the success of variational inference. Notably, there exists a trade-off phenomenon for the choice of network architecture: an overly complex structure leads to a large variational error, while an overly simplified network may not be able to capture the nonlinear feature of true underlying regression function (i.e., large approximation error).

The optimal network structure, which yields the best contraction rate, is generally unknown in reality. This motivates us to develop an *adaptive* variational inference procedure that performs automatic variational architecture selection based on the penalized ELBO criterion. The selection procedure could lead to a data-dependent network structure that achieves the same best rate as if it were derived under the optimal structure choice.

The developed general theory is further applied to two particular examples, where the true underlying function 1) is Hölder smooth, or 2) exactly corresponds to some unknown sparse DNN model. For the formal case, we show that if the smoothness level is known, the variational posterior possesses minimax contraction rate (up to a logarithm factor) when the network structure is carefully chosen based on the known smoothness level. Even when the smoothness level is unknown, the proposed adaptive variational inference procedure still leads to the same theoretical guarantee. For the latter case, we find that the rate of convergence doesn't suffer from the curse of dimensionality, in the sense that the input dimension has at most a logarithmic effect to the convergence rate.

It is worth noting that the focus of this chapter lies on the theory of variational inference on sparse DNN, and the prior used for deriving the theoretical results leads to intractable ELBO optimization. Although the variational inferences could be implemented by utilizing certain approximation, computation-friendly priors will be developed in the next chapter.

3.1.1 Notations

Throughout this Chapter as well as Chapter 4, the following notations are used. Denote $\mathrm{KL}(\cdot\|\cdot)$ and $d(\cdot,\cdot)$ as the KL divergence and Hellinger distance between two probability measures, respectively. For a vector $\boldsymbol{x} = (x_1, \ldots, x_m)^T$, we define $\|\boldsymbol{x}\|_\infty := \max_{\mathrm{i}=1}^m |x_\mathrm{i}|$, $\|\boldsymbol{x}\|_0 := \sum_{\mathrm{i}=1}^m I(x_\mathrm{i} \neq 0)$, $\|\boldsymbol{x}\|_p := (\sum_{\mathrm{i}=1}^m |x_\mathrm{i}|^p)^{1/p}$ for $p > 0$. For any Lebesgue integrable function f, we denote the L_p norm for f as $\|f\|_p := (\int f^p)^{1/p}$ and $\|f\|_\infty := \sup_{y\in\mathcal{Y}} |f(y)|$.

3.2 Nonparametric Regression Via Bayesian Deep Learning

Consider a nonparametric regression model with random covariates $X_\mathrm{i} \sim \mathcal{U}([-1,1]^p)$[1] and

$$Y_\mathrm{i} = f_0(X_\mathrm{i}) + \epsilon_\mathrm{i},\ \mathrm{i} = 1, \ldots, n \tag{3.1}$$

where $\mathcal{U}$ denotes the uniform distribution, $\epsilon_\mathrm{i} \overset{\mathrm{iid}}{\sim} \mathcal{N}(0, \sigma_\epsilon^2)$ is the noise term, and $f_0 : [-1,1]^p \to \mathbb{R}$ is the underlying true function. For simplicity of the analysis, we assume that σ_ϵ is a known constant, while in practice we could use the empirical Bayes method or full Bayes method (by placing an Inverse-Gamma prior on σ_ϵ) to estimate it.

3.2.1 Regularization via spike-and-slab prior

Given a specified sparse network configuration, we impose a fully Bayesian modeling with a spike-and-slab prior on θ. Denoting δ_0 as the Dirac at 0 and $\gamma = (\gamma_1, \ldots, \gamma_T)$ as a binary vector indicating the inclusion of each edge in the network. The prior distribution $\boldsymbol{\pi}(\theta)$ thus follows:

$$\boldsymbol{\pi}(\theta_\mathrm{i}|\gamma_\mathrm{i}) = \gamma_\mathrm{i}\mathcal{M}_0(\theta_\mathrm{i}) + (1-\gamma_\mathrm{i})\delta_0,\ \boldsymbol{\pi}(\gamma) \propto 1\{\sum \gamma_\mathrm{i} = s\} \tag{3.2}$$

for $1 \leqslant \mathrm{i} \leqslant T$, where we assign uniform prior over all possible s-sparse network structures, and the slab distribution $\mathcal{M}_0(\theta_\mathrm{i})$ is either a uniform distribution $\mathcal{U}([-B_0, B_0])$ or a Gaussian

---

[1] The bounded support assumption is common in the literature (Schmidt-Hieber 2017; Polson et al. 2018) and applies to standardized data.

distribution $\mathcal{N}(0, \sigma_0^2)$ with predetermined constant $B_0 > 1$ and $\sigma_0^2 > 0$. Our developed theory holds for both uniform slab and Gaussian slab modeling.

We denote $D_{\mathrm{i}} = (X_{\mathrm{i}}, Y_{\mathrm{i}})$ and $D = (D_1, \ldots, D_n)$ as the observations. Let P_0 denote the underlying probability measure of data, and p_0 denote the corresponding density function, i.e., $p_0(D_{\mathrm{i}}) = \psi([Y_{\mathrm{i}} - f_0(X_{\mathrm{i}})]/\sigma_\varepsilon)/\sigma_\varepsilon$ where ψ is the normal pdf. Similarly, let P_θ and p_θ be the distribution and density functions induced by the parametric NN model (1.5). Thus, the posterior distribution is written as $\pi(\theta|D) \propto \pi(\theta) \cdot p_\theta(D)$.

3.3 Variational Inference

Technically, the variational family $\mathcal{Q}$ can be chosen freely. But for the sake of efficient implementation and optimization, it is often selected as some simple distribution family. In our case, $\mathcal{Q}$ is chosen as the spike-and-slab distribution to resemble the prior distribution, i.e., for $\mathrm{i} = 1, \ldots, T$,

$$q(\theta_{\mathrm{i}}|\gamma_{\mathrm{i}}) = \gamma_{\mathrm{i}} \mathcal{M}(\theta_{\mathrm{i}}) + (1 - \gamma_{\mathrm{i}})\delta_0, \; q(\gamma_{\mathrm{i}}) = \mathrm{Bern}(\phi_{\mathrm{i}}), \tag{3.3}$$

where $\mathcal{M}(\theta_{\mathrm{i}})$ is either $\mathcal{U}(l_{\mathrm{i}}, u_{\mathrm{i}})$ with $-B_0 \leqslant l_{\mathrm{i}} \leqslant u_{\mathrm{i}} \leqslant B_0$ or $\mathcal{N}(\mu_{\mathrm{i}}, \sigma_{\mathrm{i}}^2)$ depending on the slab choice $\mathcal{M}_0$ in (3.2), and $0 \leqslant \phi_{\mathrm{i}} \leqslant 1$. Note that since the posterior can not have a larger support than the prior distribution, the ELBO optimizer must satisfy $\widehat{\phi}_{\mathrm{i}} \in \{0, 1\}$ and $\sum \widehat{\phi}_{\mathrm{i}} = s$.

3.4 VB Posterior Asymptotics

In this section, we establish the distributional convergence of the variational Bayes posterior $\widehat{q}(\theta)$, towards the true regression function f_0, under the squared Hellinger distance $d(\cdot, \cdot)$, which is

$$d^2(P_\theta, P_0) = \mathbb{E}_X \left(1 - \exp\left\{-\frac{[f_\theta(X) - f_0(X)]^2}{8\sigma_\epsilon^2}\right\}\right).$$

Note that in section 3.7, the results under L_2 norm will be studied.

Denote the log-likelihood ratio between p_0 and p_θ as

$$l_n(P_0, P_\theta) = \log \frac{p_0(D)}{p_\theta(D)} = \sum_{\mathrm{i}=1}^{n} \log \frac{p_0(D_{\mathrm{i}})}{p_\theta(D_{\mathrm{i}})},$$

then the negative ELBO can be expressed as

$$-\Omega = \mathrm{KL}(q(\theta)\|\pi(\theta)) + \int l_n(P_0, P_\theta)q(\theta)d\theta + C,$$

where $C = -\log p_0(D)$ is a constant with respect to $q(\theta)$.

Our first lemma provides an upper bound for the negative ELBO for sparse DNN model under the prior specification (3.2) and variational family $\mathcal{Q}$. Let $\Theta_B(L, \boldsymbol{p}, s) = \{\theta \in \Theta(L, \boldsymbol{p}, s) : \|\theta\|_\infty \leqslant B\}$ for some constant $B > 0$.

Lemma 3.4.1 *Given any network family $\mathcal{F}(L, \boldsymbol{p}, s)$ with an equal width $\boldsymbol{p} = (12pN, \ldots, 12pN)$, we have that, with dominating probability for some $C' > 0$,*

$$\inf_{q(\theta)\in\mathcal{Q}} \left\{ KL(q(\theta)\|\pi(\theta)) + \int l_n(P_0, P_\theta)q(\theta)d\theta \right\} \leqslant C'n(r_n + \xi_n)$$

holds, where

$$r_n := r_n(L, N, s) = \frac{(L+1)s}{n}\log(12BpN) + \frac{s}{n}\log(n(L+1)/s),$$

and

$$\xi_n := \xi_n(L, N, s) = \inf_{\theta\in\Theta_B(L,\boldsymbol{p},s)} \|f_\theta - f_0\|_\infty^2,$$

where $B = B_0$ under uniform prior setting, and $B \geqslant 2$ under normal prior setting.

The upper bound (3.4) consists of two terms: the first term r_n is the variational error caused by the variational Bayes approximation; the second term ξ_n is the approximation error of approximating f_0 by sparse ReLU DNN whose weight and bias parameters are bounded by B. Note that since B is a pre-specific constant, its value doesn't affect the rate of r_n

Our next lemma links the contraction rate of variational posterior with the negative ELBO discussed in Lemma 3.4.1.

Lemma 3.4.2 *Given network family $\mathcal{F}(L, \boldsymbol{p}, s)$ with equal width $\boldsymbol{p} = (12pN, \ldots, 12pN)$, if $\max\{s\log(n(L+1)/s), (L+1)s\log(pN)\} = o(n)$, then with probability at least $(1 - \mathrm{e}^{-Cn\varepsilon_n^2})$ for some $C > 0$, we have*

$$\int d^2(P_\theta, P_0)\widehat{q}(\theta)d\theta \leqslant C\varepsilon_n^2 + \frac{3}{n}\inf_{q(\theta)\in\mathcal{Q}}\Big\{KL(q(\theta)\|\boldsymbol{\pi}(\theta)) + \int l_n(P_0, P_\theta)q(\theta)d\theta\Big\},$$

where

$$\varepsilon_n := \varepsilon_n(L, N, s) = M\sqrt{\frac{s\log(n(L+1)/s) + (L+1)s\log(pN)}{n}}\log^\delta(n)$$

for any $\delta \geqslant 1$ and some large constant M.

Note that Lemma 3.4.2 holds regardless of the choice of prior specification $\boldsymbol{\pi}(\theta)$ and variational family $\mathcal{Q}$.

The LHS of (3.4) is the variational Bayes posterior mean of the squared Hellinger distance. On the RHS, the first term ε_n represents the estimation error under Hellinger metric, such that it is possible to test the true distribution P_0 versus all alternatives $\{P_\theta : d(P_\theta, P_0) \geqslant \varepsilon_n, \theta \in \Theta(L, \boldsymbol{p}, s)\}$ with exponentially small error probability (refer to Lemma 1.2 in the supplementary material); the second term, as discussed above, is the negative ELBO (up to a constant), which has been elaborated in Lemma 3.4.1.

Combining the above two lemmas together, one can easily obtain the following theorem:

Theorem 3.4.1 *Given any network family $\mathcal{F}(L, \boldsymbol{p}, s)$ with equal width $\boldsymbol{p} = (12pN, \ldots, 12pN)$, if the conditions of Lemmas 3.4.1 and 3.4.2 hold, then*

$$\int d^2(P_\theta, P_0)\widehat{q}(\theta)d\theta \leqslant C\varepsilon_n^2 + 3C'r_n + 3C'\xi_n. \tag{3.4}$$

The three terms in the RHS of (3.4) correspond to estimation error, variational error and approximation error respectively. All the three terms depend on the complexity of network structure. Specifically,

$$\varepsilon_n^2 \sim r_n \sim \max\left(\frac{s\log(n(L+1)/s)}{n}, \frac{(L+1)s\log(pN)}{n}\right),$$

up to only logarithmic difference. Thus both ε_n^2 and r_n are nearly linearly dependent on the sparsity and depth of the network structure specification. On the other hand, the approximation error ξ_n generally decreases as one increases the complexity of networks configuration (i.e., the values of N, L and s). Therefore, it reveals a trade-off phenomenon on the choice of network structure. Note that such trade-off echoes with those observed in the literature of nonparametric statistics: as one increases the domain of parameter space (e.g., increases the number of basis functions in spline regression modeling), it usually leads to smaller bias but larger variance.

As mentioned in Chérief-Abdellatif 2020, we would like to bring out the concept of the bias-variance trade-off in the variational inference, where we name the third and second term in RHS of (3.4) by bias and variance respectively. The variance component is controlled by r_n with an order that is always linearly dependent on the sparsity level of the DNN, which is consistent with our perception. However, its linear dependence on the depth L versus the logarithmic dependence on the width N conflicts with the result that a deeper neural net generalizes better than a shallower one as often empirically observed. In the meantime, a deeper neural net could yield a smaller approximation error with fixed neurons (Rolnick et al. 2018), which would then compensate for the increased variance caused by a deeper neural net. This reveals an interesting bias-variance trade-off phenomenon.

3.5 Adaptive Architecture Search

In Section 3.4, we establish the distributional convergence of VB posterior (3.4) under the Hellinger metric, with a pre-specified DNN architecture, say depth L, width N and sparsity s. Ideally, one would like to choose the network structure that minimizes the RHS of (3.4), thus leading to a better convergence guarantee. However, this best choice is generally not available due to the fact that the approximation error ξ_n critically depends on the nature (e.g., continuity and smoothness) of the unknown f_0. Therefore, in this section, we will develop an adaptive variational Bayes inference procedure, under which the variational posterior contraction achieves the same convergence rate as if the optimal choice of network structure was given.

To simplify our analysis, we assume that the network depth L is already well specified, and are only concerned about the adaptivity with respect to the network width and sparsity. Note that for a certain family of f_0, e.g., f_0 is Hölder smooth, the optimal choice of L can indeed be specified without additional knowledge of f_0 (refer to Section 3.6 for detail).

To be more specific, we define

$$(N^*, s^*) = \arg\min_{N,s}\{r_n(s, L, N) + \xi_n(s, L, N)\},$$

and consider $12pN^*$ and s^* to be the optimal network structure configuration for width and sparsity respectively. Such a choice strikes an optimal balance between variational error and approximation error. It is worth mentioning that the estimation error term ε_n^2 is of the same order as r_n (up to a logarithmic term). Therefore, the optimal choice (s^*, N^*) does minimize the RHS of (3.4) (up to a logarithmic term). We further define

$$\varepsilon_n^* = M'\sqrt{\frac{(L+1)s^*\log N^* + s^*\log((L+1)n/s^*)}{n}}\log^\delta(n)$$

for some constant M', $r_n^* = r_n(L, N^*, s^*)$ and $\xi_n^* = \xi_n(L, N^*, s^*)$. They represent the estimation error, variational error and approximation error respectively, under optimal choices N^* and s^*.

In addition, the following conditions are imposed on the optimal values N^* and s^*:

Condition 3.5.1 $1 \prec \max\{(L+1)s^*\log(pN^*), s^*\log(n(L+1)/s^*)\} = o(n^\alpha)$ *for some* $\alpha < 1$.

Condition 3.5.2 $r_n^* \asymp \xi_n^*$.

Condition 3.5.3 $s^* \geqslant 12pN^* + L + 1$.

Condition 3.5.1 assumes that the optimal network structure, in the asymptotic sense, is a sparse one. This is reasonable as it essentially requires that the data can be well approximated by a sparse DNN model. If this condition fails, there will be no basis for conducting sparse DNN modeling. Condition 3.5.2 implies that the choice (N^*, s^*), which minimizes $r_n + \xi_n$, also strikes the balance between r_n and ξ_n. Condition 3.5.3 avoids the redundancy of network width. If this condition is violated, then there must be redundant node (i.e., node without connection) in every hidden layers. In such a situation, all these redundant nodes shall be removed from the network configuration, leading to a narrower network.

In the Bayesian paradigm, the adaptivity can be achieved by impose a reasonable prior on (N, s). In other words, we expand the prior support to

$$\mathcal{F} = \bigcup_{N=1}^{\infty} \bigcup_{s=0}^{H_N} \mathcal{F}(L, \boldsymbol{p}_N^L, s),$$

where $\boldsymbol{p}_N^L = (12pN, \ldots, 12pN) \in \mathbb{R}^L$ and T_N is the total possible number of edges in the L-hidden-layer network with layer width $12pN$. The prior specification on the network structure is similar to Polson et al. 2018, that is

$$\begin{aligned} \pi(N) &= \frac{\lambda^N}{(\mathrm{e}^{\lambda} - 1)N!} \quad \text{for } N \geqslant 1, \\ \pi(s) &\propto \mathrm{e}^{-\lambda_s s} \quad \text{for } s \geqslant 0, \end{aligned} \tag{3.5}$$

where λ_s satisfies $n\varepsilon_n^{*2}/s^* \succ \lambda_s \geqslant a(L+1)\log n$ for some $a > 0$.

To implement variational inference, we consider the variational family $\mathcal{Q}_{N,s}$ that restricts the VB marginal posterior of N and s to be a degenerate measure: every distribution $q(\theta, N, s)$ in $\mathcal{Q}_{N,s}$ follows

$$\begin{aligned} &q(N) = \delta_{\bar{N}}, \quad q(s) = \delta_{\bar{s}}, \quad q(\gamma_{\mathrm{i}}|N, s) = \mathrm{Bern}(\phi_{\mathrm{i}}), \\ &q(\theta_{\mathrm{i}}|\gamma_{\mathrm{i}}) = \gamma_{\mathrm{i}}\mathcal{M}(\theta_{\mathrm{i}}) + (1-\gamma_{\mathrm{i}})\delta_0, \end{aligned} \tag{3.6}$$

for some $\bar{N} \in \mathbb{Z}^+$ and $\bar{s} \in \mathbb{Z}^{\geqslant 0}$. This choice of variational family means that the VB posterior will adaptively select one particular network structure $(\widehat{N}, \hat{s})$ by minimizing

$$\widehat{q}(\theta, N, s) = \underset{q(\theta, N, s) \in \mathcal{Q}_{N,s}}{\arg\max} \ \mathrm{KL}(q(\theta, N, s) \| \pi(\theta, N, s|D)).$$

Note that $\mathrm{KL}(q(\theta, N, s)\|\pi(\theta, N, s|D)) = -\log \pi(\bar{N}, \bar{s}) + \mathrm{KL}(q(\theta|\bar{N}, \bar{s})\|p(\theta, D|\bar{N}, \bar{s})) + C$, for some constant C. Let

$$\Omega(\bar{N}, \bar{s}) = \max_{q(\theta|\bar{N}, \bar{s})} [-\mathrm{KL}(q(\theta|\bar{N}, \bar{s})\|p(\theta, D|\bar{N}, \bar{s}))]$$

be the maximized ELBO given the network structure determined by parameters $\bar{N}$ and $\bar{s}$. Then

$$(\widehat{N}, \widehat{s}) = \arg\max_{\bar{N},\bar{s}}[\Omega(\bar{N}, \bar{s}) + \log \pi(\bar{N}, \bar{s})]. \tag{3.7}$$

In other words, the above VB modeling leads to a variational network structure selection based on a penalized ELBO criterion, where the penalty term is the logarithm of the prior of $\bar{N}$ and $\bar{s}$.

In Bayesian analysis, model selection relies on the (log-)posterior: $\log \pi(D|\bar{N}, \bar{s}) + \log \pi(\bar{N}, \bar{s})$. Thus, the proposed variational structure selection procedure is an approximation to maximum a posteriori (MAP) estimator, by replacing the model evidence term $\log \pi(D|\bar{N}, \bar{s})$ with the ELBO $\Omega(\bar{N}, \bar{s})$.

Our next theorem shows that the proposed variational modeling attains the best rate of convergence without the knowledge of optimal network architecture N^* and s^*.

Theorem 3.5.1 *Under the adaptive variational Bayes modeling described above, we achieve that*

$$\int d^2(P_\theta, P_0)\widehat{q}(\theta)d\theta \leqslant C'[\varepsilon_n^{*2} + r_n^* + \xi_n^*] \tag{3.8}$$

holds with dominating probability for some constant $C' > 0$.

It is worth mentioning that the above result doesn't imply the adaptive variational procedure exactly finds the optimal choice such that $\widehat{N} \approx N^*$ and $\widehat{s} \approx s^*$. The proof of Theorem 3.5.1 only shows that the adaptive VB procedure avoids over-complicated network structures, such that $\widehat{N}$ and $\widehat{s}$ will not be overwhelmingly larger than the N^* and s^* respectively. Note that (N^*, s^*) is the universal optimal choice, in the sense that it ensures that for any data set generated from the underlying model (3.1), the corresponding variational inference is the best. Note that $(\widehat{N}, \widehat{s})$ is a data-dependent choice, which differs from data to data and may be quite different from (N^*, s^*).

3.6 Applications

In this section, we will apply the general theoretical results to two important types of ground truth: 1) f_0 is some unknown Hölder smooth function and 2) f_0 exactly corresponds to an unknown sparse DNN model, i.e., the teacher-student framework Tian 2018; Goldt et al. 2019.

3.6.1 Hölder smooth function

we assume the unknown f_0 belongs to the class of α-Hölder smooth functions $\mathcal{H}_p^\alpha$, defined as

$$\mathcal{H}_p^\alpha = \Big\{ f : \|f\|_{\mathcal{H}}^\alpha := \sum_{\kappa:|\kappa|<\alpha} \|\partial^\kappa f\|_\infty + \sum_{\kappa:|\kappa|=\lfloor\alpha\rfloor} \sup_{\substack{x,y\in[-1,1]^p \\ x\neq y}} \frac{|\partial^\kappa f(x) - \partial^\kappa f(y)|}{|x-y|_\infty^{\alpha-\lfloor\alpha\rfloor}} \leqslant \infty \Big\}.$$

To quantify the approximation error ξ_n, certain knowledge of approximation theory is required. There is rich literature on the approximation properties of neural networks. For instance, Cheang and Barron 2000 and Cheang 2010 provided tight approximation error bound for simple indicator functions; Ismailov 2017 studied approximation efficiency of shallow neural network. Some recent works characterize the approximation accuracy of sparsely connected deep nets Schmidt-Hieber 2017; Bauler et al. 2019; Bölcskei et al. 2019 as well.

The following lemma is due to Schmidt-Hieber 2017, Theorem 3.

Lemma 3.6.1 *Assume $f_0 \in \mathcal{H}_p^\alpha$ for some $\alpha > 0$, then there exists a neural net $\widehat{f} \in \mathcal{F}(L, \boldsymbol{p}, s)$ with $\boldsymbol{p} = (12pN, \ldots, 12pN) \in \mathbb{R}^L$ whose bias and weight parameters are bounded by 1, and*

$$\begin{aligned} L &= 8 + (\lfloor \log_2 n \rfloor + 5)(1 + \lceil \log_2 p \rceil), \\ s &\leqslant 94p^2(\alpha+1)^{2p} N(L + \lceil \log_2 p \rceil), \\ N &= C_N \lfloor n^{p/(2\alpha+p)} / \log(n) \rfloor, \end{aligned} \tag{3.9}$$

for some positive constant C_N, such that

$$\|\widehat{f} - f_0\|_\infty \leqslant (2\|f_0\|_{\mathcal{H}}^\alpha + 1)3^{p+1}\frac{N}{n} + \|f_0\|_{\mathcal{H}}^\alpha 2^\alpha (N)^{-\alpha/p}. \tag{3.10}$$

Lemma 3.6.1 summarizes the expressibility of sparse ReLU DNN in terms of its depth, width and sparsity. It trivially implies that if L, N, s satisfy (3.9) and $p = O(1)$, then

$\max(\xi_n, r_n, \epsilon_n^2) = O(n^{2\alpha/(2\alpha+p)} \log^\delta n)$ for some $\delta > 1$. Therefore, Theorem 3.4.1 implies the following corollary.

Corollary 3.6.1 *Assume* $f_0 \in \mathcal{H}_p^\alpha$ *for some known* $\alpha > 0$, *where* $p = O(1)$. *Choose* L, s *and* N *as in (3.9). Then, our variational modeling satisfies that*

$$\int d^2(P_\theta, P_0)\widehat{q}(\theta)d\theta \leqslant C'[n^{-\alpha/(2\alpha+p)} \log^\delta(n)]^2, \tag{3.11}$$

with dominating probability, for some $\delta > 1$ *and some constant* $C' > 0$.

Corollary 3.6.1 establishes the rate minimaxity (up to a logarithmic factor) of variational sparse DNN inference. The established rate matches the contraction rate of the true Bayesian posterior (Polson et al. 2018) and therefore implies that there is no sacrifice in statistical rate with variational inference. Note that (3.11) also implies that the VB posterior mass of $\{d(P_\theta, P_0) \geqslant C'n^{-\alpha/(2\alpha+p)} \log^\delta(n)\}$ converges to zero in probability, hence almost all of the VB posterior mass contracts towards a small Hellinger ball with (near-) minimax radius centered at P_0.

The choices of N and s in (3.9), although lead to rate-minimaxity, relies on the smoothness parameter α which is usually unknown in practice. Therefore, the adaptive variational modeling discussed in Section 3.5 can be implemented here to select a reasonable N and s adaptively, such that the rate (near-)minimax convergence still holds.

Corollary 3.6.2 *Assume* $f_0 \in \mathcal{H}_p^\alpha$ *for some unknown* $\alpha > 0$, *where* $p = O(1)$. *Choose* L *as in (3.9) and let* N *and* s *follow the prior (3.5). Then result (3.11) still holds for the adaptive variational approach.*

3.6.2 Teacher-student framework

Under the Hölder smooth assumption, the rate of convergence $n^{-\alpha/(2\alpha+p)}$ suffers from the curse of dimensionality. Note that this rate merely represents the worse-case analysis among all Hölder smooth functions, which may not be suitable for real structured dataset. Hence, in this section, we are interested in the teacher-student framework, i.e., the underlying f_0 is exactly an unknown fixed sparse ReLU network (so-called teacher network), that is,

$f_0 \in \mathcal{F}(L_0, \boldsymbol{p}_0, s_0)$ for some L_0, $\boldsymbol{p}_0 = (p_{0,1}, \ldots, p_{0,L_0})'$ and s_0, and its network parameter is denoted by θ_0.

Our variational Bayes modeling with spike and slab prior can be used to train the so-called student network, based on data generated by the teacher network. Adopting this teacher-student framework can better facilitate the understanding of how deep neural networks work in high-dimensional data as it provides an explicit target function with bounded complexity.

When certain information of teacher network structure is available, we have the following result.

Corollary 3.6.3 *Under the teacher-student framework, if we choose* $L = L_0$*,* $s \geqslant s_0$ *and* $N \geqslant \max_{1\leqslant \mathrm{i} \leqslant L_0} p_{0,\mathrm{i}}/(12p)$*,* $B_0 \geqslant \|\theta_0\|_\infty$ *(under uniform prior) and* $\max\{(L+1)s\log(pN), s\log(n(L+1)/s)\} = o(n)$ *holds, then our variational Bayes approach satisfies*

$$\int d^2(P_\theta, P_0)\widehat{q}(\theta)d\theta \leqslant C'\left(\frac{s\log(n(L+1)/s) + (L+1)s\log(pN)}{n}\log^{2\delta}(n)\right), \tag{3.12}$$

with dominating probability, for some constant $C' > 0$ *and any* $\delta > 1$*.*

The choice of (N, s) means that we deliberately choose a wider and denser network structure, which ensures that the approximation error $\xi_n = 0$.

When the information of s_0 and $\boldsymbol{p}_0$ is not available, by adopting the adaptive variational modeling we also have the following result:

Corollary 3.6.4 *If the teacher network structure satisfies that* $\max\{(L_0+1)s_0\log(p\max p_{0,\mathrm{i}}), s_0\log(n(L_0+1)/s_0)\} = o(n^\alpha)$ *for some* $\alpha \in (0,1)$*, and we choose* $L = L_0$*, and let* N *and* s *follow the prior (3.5),* $B_0 \geqslant \|\theta_0\|_\infty$ *(under uniform prior), then our adaptive variational Bayes approach satisfies*

$$\int d^2(P_\theta, P_0)\widehat{q}(\theta)d\theta \leqslant C'\left(\frac{s_0\log(n(L_0+1)/s_0) + (L_0+1)s_0\log(p\max p_{0,\mathrm{i}})}{n}\log^{2\delta}(n)\right), \tag{3.13}$$

with dominating probability, for any $\delta > 1$ *and some constant* $C' > 0$*.*

The above two corollaries show that, under the teacher-student framework, the input dimension p (i.e., input layer width) and hidden layer width $\boldsymbol{p_0}$ have at most logarithmic effect on the VB posterior convergence rate. Therefore, it doesn't suffer from the curse of dimensionality.

3.7 Convergence under L_2 Norm

Our main theorems 3.4.1 and 3.5.1 concern the posterior convergence with respect to the Hellinger metric. Although commonly used in the Bayesian literature (Ghosal and Van Der Vaart 2007; Pati et al. 2018; Zhang and Gao 2019), Hellinger distance is of less practical interest than L_2 norm, i.e., $\mathbb{E}_X|f_\theta(X) - f_0(X)|^2$, for regression problems. However, a result directly addressing the L_2 convergence may not be reasonable due to the extreme flexibility of DNN models. For instance, given $p = 1$, two ReLU DNN networks $f_\theta(x) \equiv 0$ and $f_{\theta'}(x) \equiv M\sigma(x - 1 + \varepsilon)$ can have arbitrarily large L_2 distance when M is sufficiently huge, but are impossible to be discriminated when ε is so tiny that no sampled X_i visits the interval $[1 - \varepsilon, 1]$.

Accordingly, our L_2 convergence result will exclude the "irregular" DNN model f_θ's whose L_2 distances from f_0 are mostly contributed by the integral of $[f_\theta(x) - f_0(x)]^2$ over some tiny-measure subset of $[-1, 1]^p$. To be more precise, we define the L_2 distance between f_θ and f_0 as $L_2^2(f_\theta, f_0) = \mathbb{E}_X|f_\theta(X) - f_0(X)|^2$, and let $\mathcal{G} \subset \mathcal{F}(L, \boldsymbol{p}, s)$ be the subset class of all "regular" DNNs that satisfy

$$\mathbb{E}_X\{|f_\theta(X) - f_0(X)|^2 1(X \in \mathcal{S})\} \geqslant \kappa L_2^2(f_0, f_\theta),$$

for some constant $0 < \kappa \leqslant 1$, where

$$\mathcal{S} = \{X : |f_\theta(X) - f_0(X)|^2 \leqslant \gamma_n L_2^2(f_0, f_\theta)\},$$

for some $\gamma_n \to \infty$. $\mathcal{G}$ represents the DNNs that possesses a large enough expected square L_2 distance between f_θ and f_0 on a set $\mathcal{S}$ where $|f_\theta(X) - f_0(X)|^2$ is upper bounded, and the integral of $[f_\theta(x) - f_0(x)]^2$ over $\mathcal{S}^c$ doesn't make dominating contribution to $L_2^2(f_0, f_\theta)$. Naturally, $\mathcal{G}$ excludes the cases when $L_2^2(f_\theta, f_0)$ is mainly determined by the data from only a small set of the support of X.

Let $\widetilde{\varepsilon}_n^2$ denote the Hellinger convergence rate in Theorem 3.4.1 or 3.5.1, i.e., $\widetilde{\varepsilon}_n^2$ is of the same order as the RHS of equation (3.4) or (3.8). We have the following convergence result regarding L_2 metric, which states that the variational posterior mass over the irregular DNNs, which have L_2 error greater than $M_n\widetilde{\varepsilon}_n^2$, is negligible.

Theorem 3.7.1 *Given any pre-specified network family as Theorem 3.4.1 or under the adaptive variational Bayes modeling as Theorem 3.5.1, if $\gamma_n\widetilde{\varepsilon}_n^2 = o(1)$, then we have that w.h.p.*

$$\int_{\mathcal{G}\cap\{L_2^2(f_0,f_\theta)\geqslant M_n\widetilde{\varepsilon}_n^2\}} \widehat{q}(\theta)d\theta = o(1),$$

for any sequence $M_n \to \infty$.

Remark In the literature, there do exist some direct results regarding L_2 convergence rate of DNN learning and these results usually rely on some regularity condition such as the L_∞ boundedness of DNNs in the model space (Schmidt-Hieber 2017; Polson et al. 2018). However, in practice, it is usually infeasible to ensure that the trained DNN models meet the pre-specified bound, since the relationship between the magnitude of θ and $|f_\theta|_\infty$ is rather complicated.

3.8 Gumbel-softmax Approximation

To conduct optimization of negative ELBO via stochastic gradient optimization, we need to find certain reparameterization for any distribution in $\mathcal{Q}$. One solution is to use the inverse CDF sampling technique. Specifically, if $\theta \sim q \in \mathcal{Q}$, its marginal $\theta_{\rm i}$'s are independent mixture of (3.3). Let $F_{(\mu_{\rm i},\sigma_{\rm i},\phi_{\rm i})}$ be the CDF of $\theta_{\rm i}$, then $\theta_{\rm i} \overset{d}{=} F^{-1}_{(\mu_{\rm i},\sigma_{\rm i},\phi_{\rm i})}(u_{\rm i})$ holds where $u_{\rm i} \sim \mathcal{U}(0,1)$. This inverse CDF reparameterization, although valid, can not be conveniently implemented within the state-of-art python packages like PyTorch. Rather, a more popular way in VB is to utilize the Gumbel-softmax approximation.

In particular, since it is impossible to reparameterize the discrete variable γ by a continuous system, we apply the continuous relaxation - Gumbel-softmax approximation (Jang et al. 2017; Maddison et al. 2017) for the binary variable $\gamma_{\mathrm{i}} \sim \mathrm{Bern}(\phi_{\mathrm{i}})$, that is

$$\widetilde{\gamma}_{\mathrm{i}} = g_{\tau}(\phi_{\mathrm{i}}; u_{\mathrm{i}}) = \frac{1}{1+\exp(-(\log\frac{\phi_{\mathrm{i}}}{1-\phi_{\mathrm{i}}} + \log\frac{u_{\mathrm{i}}}{1-u_{\mathrm{i}}})/\tau)}, \quad u_{\mathrm{i}} \sim \mathcal{U}(0,1),$$

where τ is called the temperature, and as it approaches 0, $\tilde{\gamma}_{\mathrm{i}}$ converges to γ_{i} in distribution (refer to Figure 3.1). In addition, one can show that

$$P(\widetilde{\gamma}_{\mathrm{i}} > 0.5) = \phi_{\mathrm{i}},$$

which implies

$$\gamma_{\mathrm{i}} \overset{d}{=} 1(\widetilde{\gamma}_{\mathrm{i}} > 0.5).$$

Thus, $\widetilde{\gamma}_{\mathrm{i}}$ is viewed as a soft version of γ_{i}, and will be used in the backward pass to enable the calculation for gradients, while the hard version γ_{i} will be used in the forward pass to obtain a sparse network structure. In practice, τ is usually chosen no smaller than 0.5 for numerical stability.

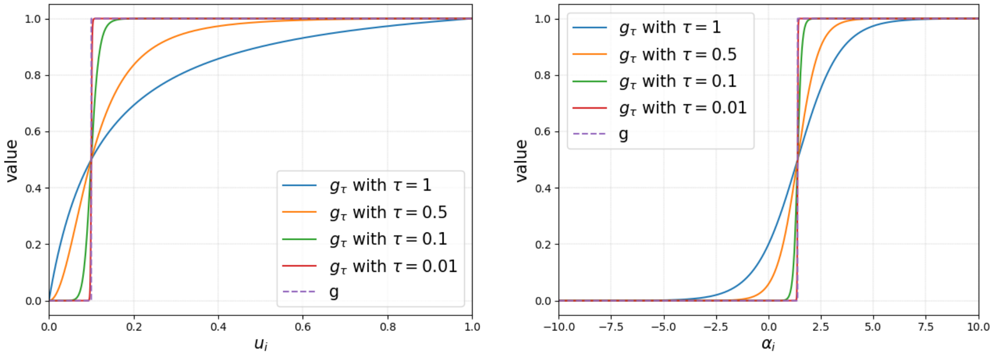
The Gumbel-softmax approximation introduces an additional error that may jeopardize the validity of our theorems. Our exploratory studies (refer to Appendix B) demonstrates little differences between the results of using inverse-CDF reparameterization and using Gumbel-softmax approximation in some simple model. Therefore, we conjecture that Gumbel-softmax approximation doesn't hurt the VB convergence, and thus will be implemented in our numerical studies.

3.8.1 Comparison between Bernoulli variable and the Gumbel softmax approximation

Denote $\gamma_{\mathrm{i}} \sim \mathrm{Bern}(\phi_{\mathrm{i}})$ and $\widetilde{\gamma}_{\mathrm{i}} = g_{\tau}(\phi_{\mathrm{i}}; u_{\mathrm{i}})$, rewrite γ_{i} as

$$\gamma_{\mathrm{i}} := g(\phi_{\mathrm{i}}; u_{\mathrm{i}}) = 1(u_{\mathrm{i}} \leqslant \phi_{\mathrm{i}}) \quad \text{where } u_{\mathrm{i}} \sim \mathcal{U}(0,1).$$

Fig 3.1 demonstrates the functional convergence of g_τ towards g as τ goes to zero. In Fig 3.1(a), by fixing $\phi_i(= 0.9)$, we show g_τ converges to g as a function of u_i. Fig 3.1 (b) demonstrates that g_τ converges to g as a function of $\alpha_i = \log(\phi_i/(1-\phi_i))$ when $u_i(= 0.2)$ is fixed. These two figures show that as $\tau \to 0$, $g_\tau \to g$. Formally, Maddison et al. 2017 rigorously proved that $\widetilde{\gamma}_i$ converges to γ_i in distribution as τ approaches 0.



(a) Fix $\phi_i = 0.9$. (b) Fix $u_i = 0.2$.

Figure 3.1. The convergence of g_τ towards g as τ approaches 0.

3.9 Implementation

In this section, the implementation details of Adaptive Sparse Variational Inference (ASVI) are provided.

3.9.1 Approximated negative ELBO

The exact AVSI algorithm requires one to figure out $\Omega(N, s)$ and compare $\Omega(N, s)$ across different choices of N and s. Our approximation integrates out the sparsity variable s in the hierarchical modeling, i.e., we consider the prior

$$\begin{aligned}
\pi(N) &= \frac{\lambda^N}{(e^\lambda - 1)N!}, \text{ for some } N \in \mathbb{Z}^+, \\
\pi(\gamma|N) &= c_1 e^{-\lambda_s \Gamma} \Big/ \binom{T}{\Gamma}, \text{ with } \Gamma = \sum_{i=1}^{T} \gamma_i, \text{ for } c_1 > 0, \qquad (3.14) \\
\pi(\theta_i|\gamma_i) &= \gamma_i \mathcal{M}_0(\theta_i) + (1-\gamma_i)\delta_0,
\end{aligned}$$

where T is the total number of possible connections given width multiplier N. The corresponding VB family is

$$\begin{aligned}
&q(N) = \delta_{\bar{N}}, \quad q(\gamma_{\rm i}|N) = \text{Bern}(\phi_{\rm i}),\\
&q(\theta_{\rm i}|\gamma_{\rm i}) = \gamma_{\rm i}\mathcal{M}(\theta_{\rm i}) + (1-\gamma_{\rm i})\delta_0,
\end{aligned}$$

for some $\bar{N} \in \mathbb{Z}^+$.

Under Gaussian slab distribution, the negative ELBO (up to a constant) corresponding to the above VB modeling is a function of $\bar{N}, \mu_{\rm i}, \sigma_{\rm i}$ and $\phi_{\rm i}$'s,

$$\begin{aligned}
-\Omega = &-\int \log p(D|\theta,\gamma,\bar{N}) d\theta d\gamma\\
&+\sum_{\rm i=1}^{T} q(\gamma_{\rm i}=1)\text{KL}(\mathcal{N}(\theta_{\rm i};\mu_{\rm i},\sigma_{\rm i}^2)\|\mathcal{N}(\theta_{\rm i};0,\sigma_0^2))\\
&+\text{KL}(q(\gamma|\bar{N})\|\boldsymbol{\pi}(\gamma|\bar{N})) - \log\boldsymbol{\pi}(\bar{N}).
\end{aligned}$$

Let

$$\begin{aligned}
\mathcal{L} = &-\int \log p(D|\theta,\gamma,\bar{N}) d\theta d\gamma\\
&+\sum_{\rm i=1}^{T} q(\gamma_{\rm i}=1)\text{KL}(\mathcal{N}(\theta_{\rm i};\mu_{\rm i},\sigma_{\rm i}^2)\|\mathcal{N}(\theta_{\rm i};0,\sigma_0^2))\\
&+\text{KL}(q(\gamma|\bar{N})\|\boldsymbol{\pi}(\gamma|\bar{N}))\\
:=&\mathcal{L}_1 + \mathcal{L}_2 + \mathcal{L}_3,
\end{aligned}$$

and

$$-\Omega(\bar{N}) = \underset{\{\mu_{\rm i},\sigma_{\rm i},\phi_{\rm i}\}}{\arg\min}\mathcal{L}. \tag{3.15}$$

Thus the optimal N value $\widehat{N}$ maximizes the penalized ELBO: $\Omega_p(\bar{N}) = \Omega(\bar{N}) + \log\boldsymbol{\pi}(\bar{N})$.

To approximate and optimize $\mathcal{L}$, we study each of the three terms:

i) $\mathcal{L}_1 = -\int \log p(D|\theta,\gamma)q(\theta|\gamma)q(\gamma|\bar{N})d\theta d\gamma$ requires Monte Carlo estimation. We use reparameterization trick (Kingma and Welling 2014; Rezende et al. 2014) for the normal slab distribution $\mathcal{M}(\theta)$, i.e., $\mathcal{M}(\theta_{\mathrm{i}})$ is equivalent in distribution to $\mu_{\mathrm{i}} + \sigma_{\mathrm{i}}\epsilon_{\mathrm{i}}$ for $\epsilon_{\mathrm{i}} \sim \mathcal{N}(0,1)$. In other words, let $\eta(\mu_{\mathrm{i}}, \sigma_{\mathrm{i}}, \phi_{\mathrm{i}}; \epsilon_{\mathrm{i}}, u_{\mathrm{i}}) = 1(g_\tau(\phi_{\mathrm{i}}; u_{\mathrm{i}}) > 0.5)(\mu_{\mathrm{i}} + \sigma_{\mathrm{i}}\epsilon_{\mathrm{i}})$ and $\widetilde{\eta}_(\mu_{\mathrm{i}}, \sigma_{\mathrm{i}}, \phi_{\mathrm{i}}; \epsilon_{\mathrm{i}}, u_{\mathrm{i}}) = g_\tau(\phi_{\mathrm{i}}; u_{\mathrm{i}})(\mu_{\mathrm{i}} + \sigma_{\mathrm{i}}\epsilon_{\mathrm{i}})$, then the stochastic estimator (Kingma and Welling 2014) for $\mathcal{L}_1$ (used for forward pass) is

$$\widetilde{\mathcal{L}_1} = -\frac{n}{m}\frac{1}{K}\sum_{\mathrm{j}=1}^{m}\sum_{k=1}^{K}\log p(D_{\mathrm{j}}|\theta^{(k)}), \tag{3.16}$$

where $\theta^{(k)} = (\theta_1^{(k)}, \ldots, \theta_T^{(k)})', \theta_{\mathrm{i}}^{(k)} = \eta(\mu_{\mathrm{i}}, \sigma_{\mathrm{i}}, \phi_{\mathrm{i}}; \epsilon_{\mathrm{i}}^{(k)}, u_{\mathrm{i}}^{(k)})$. D_{j}'s are randomly drawn from D, $\epsilon_{\mathrm{i}}^{(k)}$'s and $u_{\mathrm{i}}^{(k)}$'s are randomly drawn from $\mathcal{N}(0,1)$ and $\mathcal{U}(0,1)$ respectively, n is the sample size, m is the minibatch size and K is the Monte Carlo sample size. The stochastic estimator for $\nabla\mathcal{L}_1$ (used for backward pass) is

$$\begin{aligned}
\widetilde{\nabla}_{\mu_{\mathrm{i}}}\mathcal{L}_1 &= -\frac{n}{m}\frac{1}{K}\sum_{\mathrm{j}=1}^{m}\sum_{k=1}^{K}\nabla_{\mu_{\mathrm{i}}}\log p(D_{\mathrm{j}}|\widetilde{\theta}^{(k)}),\\
\widetilde{\nabla}_{\sigma_{\mathrm{i}}}\mathcal{L}_1 &= -\frac{n}{m}\frac{1}{K}\sum_{\mathrm{j}=1}^{m}\sum_{k=1}^{K}\nabla_{\sigma_{\mathrm{i}}}\log p(D_{\mathrm{j}}|\widetilde{\theta}^{(k)}),\\
\widetilde{\nabla}_{\phi_{\mathrm{i}}}\mathcal{L}_1 &= -\frac{n}{m}\frac{1}{K}\sum_{\mathrm{j}=1}^{m}\sum_{k=1}^{K}\nabla_{\phi_{\mathrm{i}}}\log p(D_{\mathrm{j}}|\widetilde{\theta}^{(k)}).
\end{aligned} \tag{3.17}$$

where $\widetilde{\theta}^{(k)} = (\widetilde{\theta}_1^{(k)}, \ldots, \widetilde{\theta}_T^{(k)})'$, $\widetilde{\theta}_{\mathrm{i}}^{(k)} = \widetilde{\eta}(\mu_{\mathrm{i}}, \sigma_{\mathrm{i}}, \phi_{\mathrm{i}}; \epsilon_{\mathrm{i}}^{(k)}, u_{\mathrm{i}}^{(k)})$.

ii) $\mathcal{L}_2$ is straightforward that

$$\begin{aligned}
&\sum_{\mathrm{i}=1}^{T} q(\gamma_{\mathrm{i}} = 1)\mathrm{KL}(\mathcal{N}(\theta_{\mathrm{i}}; \mu_{\mathrm{i}}, \sigma_{\mathrm{i}}^2)\|\mathcal{N}(\theta_{\mathrm{i}}; 0, \sigma_0^2))\\
=&\sum_{\mathrm{i}=1}^{T}\phi_{\mathrm{i}}(\log\frac{\sigma_0}{\sigma_{\mathrm{i}}} + \frac{\sigma_{\mathrm{i}}^2 + \mu_{\mathrm{i}}^2}{2\sigma_0^2} - 0.5).
\end{aligned} \tag{3.18}$$

iii) To compute $\mathcal{L}_3$, certain approximation is needed. Denote Γ^T as the set of all possible $\gamma = (\gamma_1, \dots, \gamma_T)$, then

$$\begin{aligned}
&\mathrm{KL}(q(\gamma|\bar{N}) \| \boldsymbol{\pi}(\gamma|\bar{N})) \\
=&\sum_{\gamma\in\Gamma^T} \log \frac{q(\gamma_1, \dots, \gamma_T)}{\boldsymbol{\pi}(\gamma_1, \dots, \gamma_T)} q(\gamma_1, \dots, \gamma_T) \\
=&\sum_{t=0}^{T} \sum_{\Gamma=t} \log \frac{q(\gamma_1, \dots, \gamma_T)}{\boldsymbol{\pi}(\gamma_1, \dots, \gamma_T)} q(\gamma_1, \dots, \gamma_T)
\end{aligned}$$

For the sake of fast computation, we approximate the VB distribution $q(\gamma)$ by iid Bernoulli distribution $q(\gamma) \approx \prod \widetilde{\phi}^{\gamma_\mathrm{i}}(1-\widetilde{\phi})^{1-\gamma_\mathrm{i}}$, where $\widetilde{\phi} = \frac{1}{T}\sum_{\mathrm{i}=1}^{T} \phi_\mathrm{i}$. Under this approximation:

$$\begin{aligned}
&\sum_{\Gamma=t} \log \frac{q(\gamma_1, \dots, \gamma_T)}{\boldsymbol{\pi}(\gamma_1, \dots, \gamma_T)} q(\gamma_1, \dots, \gamma_T) \\
\approx&\sum_{\Gamma=t} \log \frac{\widetilde{\phi}^t(1-\widetilde{\phi})^{T-t}}{\boldsymbol{\pi}(\gamma|\Gamma=t)} \widetilde{\phi}^t(1-\widetilde{\phi})^{T-t} \\
=&\binom{T}{t} \log \frac{\widetilde{\phi}^t(1-\widetilde{\phi})^{T-t}}{\boldsymbol{\pi}(\gamma|\Gamma=t)} \widetilde{\phi}^t(1-\widetilde{\phi})^{T-t} \\
=&\log \frac{\binom{T}{t}\widetilde{\phi}^t(1-\widetilde{\phi})^{T-t}}{\binom{T}{t}\boldsymbol{\pi}(\gamma|\Gamma=t)} \binom{T}{t} \widetilde{\phi}^t(1-\widetilde{\phi})^{T-t} \\
=&\log Pr(\Gamma=t) Pr(\Gamma=t) + \lambda_s t Pr(\Gamma=t) + C_1
\end{aligned}$$

where C_1 is some constant. Therefore, $\mathrm{KL}(q(\gamma)\|\boldsymbol{\pi}(\gamma)))$ is approximated by

$$
\begin{aligned}
&\sum_{\gamma\in\Gamma^T} \log \frac{q(\gamma_1,\dots,\gamma_T)}{\pi(\gamma_1,\dots,\gamma_T)} q(\gamma_1,\dots,\gamma_T)\\
=&\sum_{t=0}^{T}\sum_{\Gamma=t} \log \frac{q(\gamma_1,\dots,\gamma_T)}{\pi(\gamma_1,\dots,\gamma_T)} q(\gamma_1,\dots,\gamma_T)\\
=&\sum_{t=0}^{T} \log Pr(\Gamma=t)P(\Gamma=t) + \lambda_s \sum_{t=0}^{T} tPr(\Gamma=t) + C_2\\
=&-\mathbb{H}(\Gamma) + \lambda_s \mathbb{E}(\Gamma) + C_2\\
\approx&-0.5\log_2(2\pi e \sum \phi_{\mathrm{i}}(T - \sum \phi_{\mathrm{i}})/T) + \lambda_s \sum_{\mathrm{i}=1}^{T} \phi_{\mathrm{i}} + C_2\\
:=&\widetilde{\mathcal{L}_3}
\end{aligned}
\tag{3.19}
$$

where $\mathbb{H}(\Gamma)$ is the entropy of Γ and C_2 is some constant.

3.9.2 Algorithm

An additional re-parametrization transformation for σ and ϕ is used,

$$
\sigma'_{\mathrm{i}} = \log(\exp(\sigma_{\mathrm{i}}) - 1),\ \phi'_{\mathrm{i}} = \log(\frac{1-\phi_{\mathrm{i}}}{\phi_{\mathrm{i}}}),
$$

such that σ'_{i} and $\phi'_{\mathrm{i}} \in \mathbb{R}$. Let $\widetilde{\mathcal{L}}$ and $\widetilde{\nabla}\mathcal{L}$ denote the working approximations of $\mathcal{L}$ and $\nabla\mathcal{L}$, then $\widetilde{\mathcal{L}} = \widetilde{\mathcal{L}_1} + \mathcal{L}_2 + \widetilde{\mathcal{L}_3}$ using (3.16), (3.18) and (3.19). Furthermore, there exist explicit gradients of $\mathcal{L}_2$ and $\widetilde{\mathcal{L}_3}$ with respect to ϕ'_{i}, μ_{i} and σ'_{i}, which facilitates the calculation of the approximate gradient $\widetilde{\nabla}\mathcal{L}$ along with (3.17).

The complete adaptive sparse variational inference is described in Algorithm 2, where we use $\widetilde{\Omega}(\bar{N})$ and $\widetilde{\Omega}_p(\bar{N})$ to denote the working approximations of $\Omega(\bar{N})$ and $\Omega_p(\bar{N})$ respectively.

3.10 Experiments

In this section, we investigate the performance of the proposed Adaptive Sparse Variational Inference (ASVI) with Gaussian slab prior through empirical studies. To implement ASVI, after pre-specifying the depth L, one needs to assign prior distributions for N and s according

Algorithm 2 Adaptive sparse variational inference with normal slab distribution.

```
Hyperparameters: $\lambda$, $\lambda_s$, $\sigma_0$
Parameters: $\mu, \sigma', \phi'$
Candidate set of $\bar{N}$: $N_A$
for all $\bar{N} \in N_A$ do in parallel
    repeat
        $\{D_{\mathrm{j}}\}_{\mathrm{j}=1}^{m} \leftarrow$ Sample a minibatch of size $m$
        $\{\epsilon_{\mathrm{i}}^{(k)}\}_{1 \leqslant k \leqslant K, 1 \leqslant \mathrm{i} \leqslant T} \leftarrow$ i.i.d. samples from $\mathcal{N}(0,1)$
        $\{u_{\mathrm{i}}^{(k)}\}_{1 \leqslant k \leqslant K, 1 \leqslant \mathrm{i} \leqslant T} \leftarrow$ i.i.d. samples from $\mathcal{U}(0,1)$
        $\widetilde{\mathcal{L}} \leftarrow$ (3.16), (3.18) and (3.19)
        $\widetilde{\nabla}_{\mu_{\mathrm{i}}}\mathcal{L}$, $\widetilde{\nabla}_{\sigma_{\mathrm{i}}}\mathcal{L}$, $\widetilde{\nabla}_{\phi_{\mathrm{i}}}\mathcal{L} \leftarrow$ Gradients of $\mathcal{L}_2$ and $\widetilde{\mathcal{L}_3}$ together with (3.17)
        $\widetilde{\nabla}_{\sigma'_{\mathrm{i}}}\mathcal{L} \leftarrow \widetilde{\nabla}_{\sigma_{\mathrm{i}}}\mathcal{L} \cdot \nabla_{\sigma'_{\mathrm{i}}}\sigma_{\mathrm{i}}$
        $\widetilde{\nabla}_{\phi'_{\mathrm{i}}}\mathcal{L} \leftarrow \widetilde{\nabla}_{\phi_{\mathrm{i}}}\mathcal{L} \cdot \nabla_{\phi'_{\mathrm{i}}}\phi_{\mathrm{i}}$
        $\mu_{\mathrm{i}}, \sigma'_{\mathrm{i}}, \phi'_{\mathrm{i}} \leftarrow$ Update with $\widetilde{\nabla}_{\mu_{\mathrm{i}}}\mathcal{L}$, $\widetilde{\nabla}_{\sigma'_{\mathrm{i}}}\mathcal{L}$, $\widetilde{\nabla}_{\phi'_{\mathrm{i}}}\mathcal{L}$ using gradient descent algorithms
                        (e.g. RMSprop or Adam)
    until convergence of $\widetilde{\mathcal{L}}$
    $-\widetilde{\Omega}(\bar{N}) \leftarrow \widetilde{\mathcal{L}}$
    $-\widetilde{\Omega}_p(\bar{N}) \leftarrow -\widetilde{\Omega}(\bar{N}) - \log \boldsymbol{\pi}(\bar{N})$ with $(\bar{N}, \lambda)$
end for
$\widehat{N} = \arg\min_{\bar{N} \in N_A}(-\widetilde{\Omega}_p(\bar{N}))$
return $\widehat{N}$ and $(\mu, \sigma', \phi' | \widehat{N})$
```

Table 3.1. Results for teacher network experiment. The average test RMSE with standard error and average posterior number of edges with standard error are exhibited.

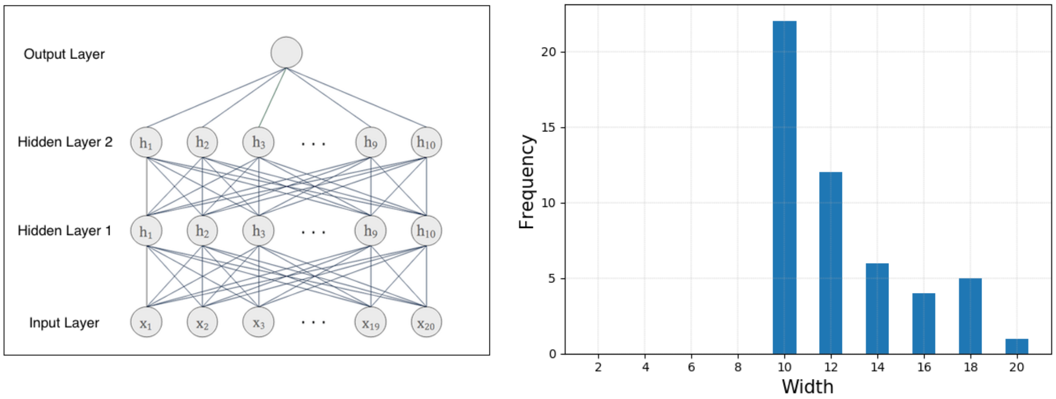
	Test RMSE				# of edges			
Width	**ASVI**	**SVI**	**HS-BNN**	**Dense-BNN**	**ASVI**	**SVI**	**HS-BNN**	**Dense-BNN**
2	-	2.193 ± 0.195	2.193 ± 0.163	2.131 ± 0.097	-	48.28 ± 2.099	51.00 ± 0.000	51.00 ± 0.000
4	-	1.636 ± 0.069	1.715 ± 0.160	1.591 ± 0.087	-	94.43 ± 4.499	109.0 ± 0.000	109.0 ± 0.000
6	-	1.210 ± 0.049	1.322 ± 0.179	1.190 ± 0.033	-	125.7 ± 8.805	175.0 ± 0.000	175.0 ± 0.000
8	-	1.065 ± 0.038	1.108 ± 0.048	1.046 ± 0.021	-	135.5 ± 10.87	249.0 ± 0.000	249.0 ± 0.000
10	-	1.014 ± 0.023	1.058 ± 0.029	1.014 ± 0.010	-	151.1 ± 13.25	331.0 ± 0.000	331.0 ± 0.000
12	-	1.019 ± 0.085	1.035 ± 0.016	1.010 ± 0.007	-	166.1 ± 14.41	421.0 ± 0.000	421.0 ± 0.000
14	-	1.018 ± 0.093	1.034 ± 0.010	1.011 ± 0.009	-	177.3 ± 15.62	519.0 ± 0.000	519.0 ± 0.000
16	-	1.011 ± 0.037	1.032 ± 0.010	1.009 ± 0.005	-	186.1 ± 16.48	625.0 ± 0.000	625.0 ± 0.000
18	-	1.005 ± 0.008	1.030 ± 0.010	1.010 ± 0.005	-	190.3 ± 15.87	739.0 ± 0.000	739.0 ± 0.000
20	-	1.003 ± 0.006	1.029 ± 0.008	1.010 ± 0.007	-	192.5 ± 13.78	861.0 ± 0.000	861.0 ± 0.000
Adaptive	1.003 ± 0.010	-	-	-	155.9 ± 15.58	-	-	-

to (3.5), and assign uniform prior (3.2) over the network structure γ given s. However, as emphasized in the introduction, it is not computationally feasible to solve ASVI, since the exact minimization of negative ELBO requires exhaustively search over all possible sparse network structures. As a consequence, in this numerical studies section, an approximated solution of $\widehat{q}$ is used instead. The details of the approximation and implementation of ASVI are presented in Section 2 of the Appendix. In short words, we integrate out the sparsity variable s in the hierarchical prior (3.5), and only consider the marginal modelling of N and θ. Given the width multiplier N, the maximized ELBO $\Omega(N)$ is obtained by back propagation with the help of some approximation and binary relaxation. The optimal structure is then selected by the penalized ELBO criterion similar to (3.7). In this simulation, we typically specify 5-10 levels of width choices and compute $\Omega(N)$ for different N in parallel.

For all the numerical studies, we use the VB posterior mean estimator $\widehat{f} = \sum_{\mathrm{i}=1}^{30} f_{\theta_\mathrm{i}}/30$ to assess the prediction accuracy, where θ_i's are randomly drawn from the VB posterior $\widehat{q}(\theta)$. We use $\widehat{s} = \sum_{\mathrm{i}=1}^{H} \phi_\mathrm{i}/H$ to measure the posterior network sparsity. We compare our method to Horseshoe BNN (HS-BNN) (Ghosh, Yao, et al. 2018) and dense BNN (Blundell et al. 2015).

3.10.1 Simulation study

We consider a simulated experiment under the teacher-student framework. As shown in Fig 3.2 (a), we use a 2-hidden-layer teacher network with ReLU activation, where the specific structure is 20-10-10-1. The edges of the teacher network are first randomly generated from $\mathcal{U}(0.5, 1.5)$ and then randomly set to 0 by a rate of 50% to ensure a sparse structure. We



(a) Teacher network

(b) Selected width

Figure 3.2. (a) Teacher network with structure 20-10-10-1, where 50% of the edges are set to 0 randomly. (b) Frequency of the selected width in 50 replications.

Table 3.2. Average test RMSE with standard error for UCI regression datasets.

Dataset	**n (p)**	**SVI**	**HS-BNN**	**PBP**
Kin8nm	8192 (8)	0.08±0.00	0.08±0.00	0.10±0.00
Naval	11934 (16)	0.00±0.00	0.00±0.00	0.01±0.00
Power Plant	9568 (4)	4.02±0.18	4.03±0.15	4.12±0.03
Protein	45730 (9)	4.36±0.04	4.39±0.04	4.73±0.01
Wine	1599 (11)	0.62±0.03	0.63±0.04	0.64±0.01
Year	515345 (90)	8.85±NA	9.26±NA	8.88±NA

fix the depth L of student net to 2 in the experiment, and consider the width of student net to range from 2 to 20 with a increment of 2. We randomly generate 50 datasets of size 10000 from the teacher network with random noise variance $\sigma_\epsilon = 1$ for training, and the adaptive variational inference is performed on each of these datasets to select the best network structure. The remaining implementation details can be found in the Appendix.

Fig 3.2 (b) plots the frequency of the selected width among the 50 replications. It shows that in most time the ASVI selects width 10 or 12, which is close to the true width. We compare the test Root Mean Squared Error (RMSE) of ASVI against non-adaptive SVI (i.e., ASVI without width selection), HS-BNN and Dense-BNN with all the choices of width. The result is displayed in Table 3.1. It shows that ASVI achieves best test Root Mean Squared

Error (RMSE), which is quite close to the random noise ($\sigma_\epsilon = 1$). In addition, the number of edges selected by ASVI is also close to the ground truth (around 165.5).

3.10.2 Real data

We compare the performance of our method to others on UCI regression tasks and MNIST data. For UCI datasets, following the same experimental protocol as Hernández-Lobato et al. 2015, a single layer neural network of 50 units with ReLU activation is used for all the datasets, except for the larger ones "Protein" and "Year", where 100 units are used. For the smaller datasets, we randomly select 90% and 10% for training and testing respectively, and the process is repeated for 20 times. For "Protein", only 5 replication is performed. For "Year", where the training and testing datasets are predefined, the process is only done once. We compare our method to HS-BNN and probabilistic backpropagation (PBP) of ibid. For MNIST, we use a two hidden layer ReLU network with width of $\{400, 500, 600, 700, 800\}$. Other Implementation details can be found in the Appendix.

Table 3.2 shows our method (SVI) performs as well as or better than the other methods on UCI datasets with pre-determined architecture. Figure 3.3 shows our method achieves best test accuracy for MNIST data, with a selected width of 700 and posterior sparsity of 6.01% (62855 edges) at epoch 300.

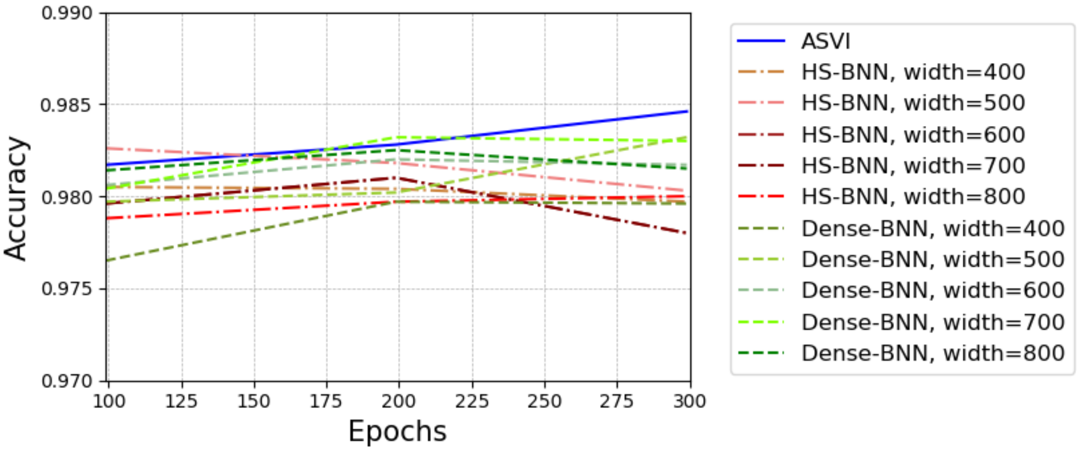


Figure 3.3. Test accuracy for MNIST data

3.10.3 Remaining implementation details

Teacher network The batch size is set as $m = 1024$, and Monte Carlo size $K = 1$ during training. Adam is used for optimization with a learning rate of 5×10^{-3}, and the number of epochs is 7000. λ_s is chosen as 3 ($a = 0.1$) and λ is chosen as 10, σ_0 is fixed at 0.8.

UCI datasets For all the datasets, the batch size is set as $m = 256$, Monte Carlo size K is set as 1 during training, and Adam is used for optimization with a learning rate of 1×10^{-3}. The number of epochs is 1000 for ”Naval”, ”Power Plant” and ”Protein”, 2000 for ”Kin8nm” and 100 for ”Year”. σ_0 and σ_ϵ are determined by a grid search that yields the best prediction accuracy.

MNIST The batch size is set as $m = 512$, and Monte Carlo size $K = 1$ during training. RMSprop is used for optimization with a learning rate of 5×10^{-3}, and the number of epochs is 300. λ_s is chosen as 50 ($a = 1.5$) and λ is chosen as 600, σ_0 is fixed at 2. MNIST data is standardized by mean of 0.1307 and standard deviation of 0.3081.

3.11 Conclusion and Discussion

In this chapter, we investigate the theoretical aspects of variational inference for sparse DNN models. Although theoretically sound, the spike and slab modeling with Dirac spike is difficult to implement in practice, and some continuous relaxation is required that deserves further theoretical investigation. In addition, despite the fact that the proposed uniform prior distribution for s guarantees good theoretical properties, it is also not practical and some approximation is involved in our implementation. Therefore, some alternative choice of prior distribution could be investigated in the future.

3.12 Main Proofs

The detailed proofs for our lemmas and theorems are included in this section.

3.12.1 Proof of Lemma 3.4.1

Lemma 3.12.1 restates the Donsker and Varadhan's representation for the KL divergence, its proof can be found in Boucheron et al. 2013.

Lemma 3.12.1 *For any probability measure λ and any measurable function h with $\mathrm{e}^h \in L_1(\lambda)$,*

$$\log \int \mathrm{e}^{h(\eta)} \lambda(d\eta) = \sup_{\rho} \left[\int h(\eta)\rho(d\eta) - KL(\rho \| \lambda). \right]$$

The next lemma proves the existence of a testing function which can exponentially separate P_0 and $\{P_\theta : d(P_0, P_\theta) \geqslant \varepsilon_n, P_\theta \in \mathcal{F}(L, \boldsymbol{p}, s)\}$. The existence of such testing function is crucial for Lemma 3.4.1.

Lemma 3.12.2 *Let $\varepsilon_n = M\sqrt{\frac{s\log(nL/s)+Ls\log(pN)}{n}}\log^{\delta}(n)$ for any $\delta \geqslant 1$ and some large constant M. Then there exists some testing function $\phi \in [0,1]$ and $C_1 > 0$, $C_2 > 1/3$, such that*

$$\mathbb{E}_{P_0}(\phi) \leqslant \exp\{-C_1 n\varepsilon_n^2\},$$

$$\sup_{\substack{P_\theta \in \mathcal{F}(L,\boldsymbol{p},s) \\ d(P_\theta, P_0) > \varepsilon_n}} \mathbb{E}_{P_\theta}(1-\phi) \leqslant \exp\{-C_2 n d^2(P_0, P_\theta)\}.$$

Proof 6 Due to the well-known result (e.g., Le Cam 1986, page 491 or Ghosal and Van Der Vaart 2007, Lemma 2), there always exists a function $\psi \in [0,1]$, such that

$$\mathbb{E}_{P_0}(\psi) \leqslant \exp\{-nd^2(P_{\theta_1}, P_0)/2\},$$

$$\mathbb{E}_{P_\theta}(1-\psi) \leqslant \exp\{-nd^2(P_{\theta_1}, P_0)/2\},$$

for all $P_\theta \in \mathcal{F}(L, \boldsymbol{p}, s)$ satisfying that $d(P_\theta, P_{\theta_1}) \leqslant d(P_0, P_{\theta_1})/18$.

Let $K = N(\varepsilon_n/19, \mathcal{F}(L, \boldsymbol{p}, s), d(\cdot,\cdot))$ denote the covering number of set $\mathcal{F}(L, \boldsymbol{p}, s)$, i.e., there exists K Hellinger-balls with radius $\varepsilon_n/19$, that completely cover $\mathcal{F}(L, \boldsymbol{p}, s)$. For any $\theta \in \mathcal{F}(L, \boldsymbol{p}, s)$ (W.O.L.G, we assume P_θ belongs to the kth Hellinger ball centered at P_{θ_k}), if $d(P_\theta, P_0) > \varepsilon_n$, then we must have that $d(P_0, P_{\theta_k}) > (18/19)\varepsilon_n$ and there exists a testing function ψ_k, such that

$$\begin{aligned}
\mathbb{E}_{P_0}(\psi_k) &\leqslant \exp\{-nd^2(P_{\theta_k}, P_0)/2\} \\
&\leqslant \exp\{-(18^2/19^2/2)n\varepsilon_n^2\}, \\
\mathbb{E}_{P_\theta}(1-\psi_k) &\leqslant \exp\{-nd^2(P_{\theta_k}, P_0)/2\} \\
&\leqslant \exp\{-n(d(P_0, P_\theta) - \varepsilon_n/19)^2/2\} \\
&\leqslant \exp\{-(18^2/19^2/2)nd^2(P_0, P_\theta)\}.
\end{aligned}$$

Now we define $\phi = \max_{k=1,\ldots,K} \psi_k$. Thus we must have

$$\begin{aligned}
\mathbb{E}_{P_0}(\phi) \leqslant \sum_k \mathbb{E}_{P_0}(\psi_k) &\leqslant K \exp\{-(18^2/19^2/2)n\varepsilon_n^2\} \\
&\leqslant \exp\{-((18^2/19^2/2)n\varepsilon_n^2 - \log K)\}.
\end{aligned}$$

Note that

$$\begin{aligned}
\log K &= \log N(\varepsilon_n/19, \mathcal{F}(L, \boldsymbol{p}, s), d(\cdot,\cdot)) \\
&\leqslant \log N(\sqrt{8}\sigma_\varepsilon\varepsilon_n/19, \mathcal{F}(L, \boldsymbol{p}, s), \|\cdot\|_\infty) \\
&\leqslant (s+1)\log\left(\frac{38}{\sqrt{8}\sigma_\varepsilon\varepsilon_n}(L+1)(12pN+1)^{2(L+1)}\right) \\
&\leqslant s\log\frac{1}{\varepsilon_n} + s\log(n(L+1)/s) + s(L+1)\log(pN) \\
&\leqslant n\varepsilon_n^2/4, \quad \text{for sufficiently large n,}
\end{aligned} \tag{3.20}$$

where the first inequality is due to the fact

$$d^2(P_\theta, P_0) \leqslant 1 - \exp\{-\frac{1}{8\sigma_\epsilon^2}\|f_0 - f_\theta\|_\infty^2\}$$

and $\varepsilon_n = o(1)$, the second inequality is due to Lemma 10 of Schmidt-Hieber 2017. Therefore,

$$\mathbb{E}_{P_0}(\phi) \leqslant \sum_k P_0(\psi_k) \leqslant \exp\{-C_1 n\varepsilon_n^2\},$$

for some $C_1 = 18^2/19^2/2 - 1/4$. On the other hand, for any θ, such that $d(P_\theta, P_0) \geqslant \varepsilon_n$, say P_θ belongs to the kth Hellinger ball, then we have

$$\mathbb{E}_{P_\theta}(1-\phi) \leqslant \mathbb{E}_{P_\theta}(1-\psi_k) \leqslant \exp\{-C_2 n d^2(P_0, P_\theta)\},$$

where $C_2 = 18^2/19^2/2$. Hence we conclude the proof. ■

Now, we are ready to prove Lemma 3.4.1.

Proof 7 It suffices to construct some $q^*(\theta) \in \mathcal{Q}$, such that w.h.p,

$$\begin{aligned}
&\mathrm{KL}(q^*(\theta)\|\pi(\theta)) + \int l_n(P_0, P_\theta) q^*(\theta) d\theta \\
\leqslant & n r_n + \frac{3n}{2\sigma_\varepsilon^2} \inf_\theta \|f_\theta - f_0\|_\infty^2 + \frac{3 n r_n}{2\sigma_\epsilon^2}.
\end{aligned} \tag{3.21}$$

Let $\theta^* = \arg\min_{\theta \in \Theta_B(L, \boldsymbol{p}, s)} \|f_\theta - f_0\|_2^2$ and we choose the same $q^*(\theta)$ that has been used in the proof of Theorem 2 of Chérief-Abdellatif 2020. Specifically, for all $h = 1, \ldots, T$, $\gamma_h^* = \mathbb{I}(\theta_h^* \neq 0)$, and

i) For uniform slab distribution,

$$\theta_h \sim \gamma_h^* \mathcal{U}([\theta_h^* - a_n, \theta_h^* + a_n]) + (1 - \gamma_h^*)\delta_0, \tag{3.22}$$

where $a_n = \frac{s}{4n}(12BpN)^{-2L}\{(p+1+\frac{1}{12BpN-1})^2\frac{L^2}{(12BpN)^2} + \frac{1}{(12BpN)^2-1} + \frac{2}{(12BpN-1)^2}\}^{-1}$.

ii) For Gaussian slab distribution,

$$\theta_h \sim \gamma_h^* \mathcal{N}(\theta_h^*, \sigma_n^2) + (1 - \gamma_h^*)\delta_0, \tag{3.23}$$

where $\sigma_n^2 = \frac{s}{16n}\log(36pN)^{-1}(24BpN)^{-2L}\{(p+1+\frac{1}{12BpN-1})^2 + \frac{1}{(24BpN)^2-1} + \frac{2}{(24BpN-1)^2}\}^{-1}$.

According to the proof of Theorem 2 in ibid.,

$$\mathrm{KL}(q^*(\theta)\|\pi(\theta)) \leqslant nr_n, \tag{3.24}$$

$$\int \|f_\theta - f_{\theta*}\|_\infty^2 q^*(\theta)d\theta \leqslant r_n, \tag{3.25}$$

and the first term on L.H.S of (3.21) is bounded.

To upper bound the second term on L.H.S of (3.21), note that

$$\begin{aligned}
l_n(P_0, P_\theta) =& \frac{1}{2\sigma_\epsilon^2}(\|Y - f_\theta(X)\|_2^2 - \|Y - f_0(X)\|_2^2) \\
=& \frac{1}{2\sigma_\epsilon^2}(\|Y - f_0(X) + f_0(X) - f_\theta(X))\|_2^2 - \|Y - f_0(X)\|_2^2) \\
=& \frac{1}{2\sigma_\epsilon^2}(\|f_\theta(X) - f_0(X)\|_2^2 + 2\langle Y - f_0(X), f_0(X) - f_\theta(X)\rangle).
\end{aligned}$$

Denote

$$\begin{aligned}
\mathcal{R}_1 &= \int \|f_\theta(X) - f_0(X)\|_2^2 q^*(\theta)d\theta, \\
\mathcal{R}_2 &= \int \langle Y - f_0(X), f_0(X) - f_\theta(X)\rangle q^*(\theta)d\theta.
\end{aligned}$$

Since $\|f_\theta(X) - f_0(X)\|_2^2 \leqslant n\|f_\theta - f_0\|_\infty^2 \leqslant n\|f_\theta - f_{\theta*}\|_\infty^2 + n\|f_{\theta*} - f_0\|_\infty^2$,

$$\mathcal{R}_1 \leqslant nr_n + n\|f_{\theta*} - f_0\|_\infty^2.$$

Noting that $Y - f_0(X) = \epsilon \sim \mathcal{N}(0, \sigma_\epsilon^2 I)$, then

$$\begin{aligned}
\mathcal{R}_2 &= \int \epsilon^T (f_0(X) - f_\theta(X))q^*(\theta)d\theta \\
&= \epsilon^T \int (f_0(X) - f_\theta(X))q^*(\theta)d\theta \\
&\sim \mathcal{N}(0, c_f \sigma_\epsilon^2),
\end{aligned}$$

where $c_f = \|\int (f_0(X) - f_\theta(X))q^*(\theta)d\theta\|_2^2 \leqslant \mathcal{R}_1$ due to Cauchy-Schwarz inequality. Then by Gaussian tail bound

$$P_0(\mathcal{R}_2 \geqslant \mathcal{R}_1) \leqslant \exp(-\frac{\mathcal{R}_1^2}{2\sigma_\epsilon^2 \mathcal{R}_1}),$$

which implies $\mathcal{R}_2 \leqslant \mathcal{R}_1$ w.h.p.. Therefore,

$$\int l_n(P_0, P_\theta) q^*(\theta) d\theta = \mathcal{R}_1/2\sigma_\epsilon^2 + \mathcal{R}_2/\sigma_\epsilon^2 \leqslant 3n(r_n + \|f_{\theta*} - f_0\|_\infty^2)/2\sigma_\varepsilon^2, \text{ w.h.p.},$$

which concludes this lemma together with (3.24).

■

3.12.2 Proof of Lemma 3.4.1

The proof is adapted from the proof of Theorem 3.1 in Pati et al. 2018.

Proof 8 We claim that with high probability (w.h.p),

$$M = \int_\Theta \eta(P_\theta, P_0) \boldsymbol{\pi}(\theta) d\theta \leqslant \mathrm{e}^{Cn\varepsilon_n^2} \tag{3.26}$$

for some $C > 0$, where $\log \eta(P_\theta, P_0) = l_n(P_\theta, P_0) + \frac{n}{3} d^2(P_\theta, P_0)$. Thus by Lemma 3.12.1, w.h.p.,

$$\begin{aligned}
&\frac{n}{3} \int d^2(P_\theta, P_0) \widehat{q}(\theta) d\theta \\
\leqslant & Cn\varepsilon_n^2 + \mathrm{KL}(\widehat{q}(\theta) \| \boldsymbol{\pi}(\theta)) - \int l_n(P_\theta, P_0) \widehat{q}(\theta) d\theta \\
\leqslant & Cn\varepsilon_n^2 + \mathrm{KL}(q(\theta) \| \boldsymbol{\pi}(\theta)) - \int l_n(P_\theta, P_0) q(\theta) d\theta
\end{aligned}$$

holds for any distribution q_θ. The last inequality holds since that $\mathrm{KL}(q(\theta) \| \boldsymbol{\pi}(\theta)) - \int l_n(P_\theta, P_0) q(\theta) d\theta$ is the negative ELBO function up to a constant, which is minimized at $\widehat{q}(\theta)$. This concludes Lemma 4.3.

To prove (3.26), we define

$$\begin{aligned}
M_1 &= \int_{d(P_\theta, P_0) \leqslant \varepsilon_n} \eta(P_\theta, P_0) \boldsymbol{\pi}(\theta) d\theta, \\
M_2 &= \int_{d(P_\theta, P_0) > \varepsilon_n} \eta(P_\theta, P_0) \boldsymbol{\pi}(\theta) d\theta,
\end{aligned}$$

and will bound both M_1 and M_2.

For M_1, by Fubini's theorem,

$$\begin{aligned}\mathbb{E}_{P_0} M_1 =& \int_{d(P_\theta,P_0)\leqslant\varepsilon_n} \int \frac{p_\theta(D)}{p_0(D)} \mathrm{e}^{\frac{n}{3}d^2(P_\theta,P_0)} \\ & dP_0(D)\pi(\theta)d\theta \\ =& \int_{d(P_\theta,P_0)\leqslant\varepsilon_n} \mathrm{e}^{\frac{n}{3}d^2(P_\theta,P_0)}\pi(\theta)d\theta \\ \leqslant& \mathrm{e}^{\frac{n}{3}\varepsilon_n^2}.\end{aligned}$$

It follows from Markov inequality that $M_1 \leqslant \mathrm{e}^{Cn\varepsilon_n^2}$ w.h.p..

For M_2, we further decompose it as $M_2 = M_{21} + M_{22}$,

$$\begin{aligned}M_{21} &= \int_{d(P_\theta,P_0)>\varepsilon_n} \phi\eta(P_\theta,P_0)\pi(\theta)d\theta, \\ M_{22} &= \int_{d(P_\theta,P_0)>\varepsilon_n} (1-\phi)\eta(P_\theta,P_0)\pi(\theta)d\theta,\end{aligned}$$

where the testing function ϕ is defined in Lemma 3.12.2.

For M_{21}, since $\mathbb{E}_{P_0}[\phi] \leqslant \mathrm{e}^{-C_1 n\varepsilon_n^2}$, $\phi \leqslant \mathrm{e}^{-C_1' n\varepsilon_n^2}$ for some $C_1' > 0$ w.h.p., thus $M_{21} \leqslant \mathrm{e}^{-C_1' n\varepsilon_n^2} M_2$ w.h.p.

For M_{22}, by Fubini's theorem and Lemma 3.12.2,

$$\begin{aligned}\mathbb{E}_{P_0} M_{22} &= \int_{d(P_\theta,P_0)>\varepsilon_n} \mathbb{E}_{P_\theta}(1-\phi)\mathrm{e}^{\frac{n}{3}d^2(P_\theta,P_0)}\pi(\theta)d\theta \\ &\leqslant \mathrm{e}^{-(C_2-1/3)n\varepsilon_n^2} := \mathrm{e}^{-C_2' n\varepsilon_n^2}.\end{aligned}$$

Thus, $M_2 \leqslant \mathrm{e}^{-C_1' n\varepsilon_n^2} M_2 + \mathrm{e}^{-C_2' n\varepsilon_n^2}$ w.h.p., which implies that $M_2 \leqslant \mathrm{e}^{-C_2 n\varepsilon_n^2}$ w.h.p. for some $C_2 > 0$.

Combine the boundedness results for both M_1 and M_2, we conclude (3.26).

■

3.12.3 Proof of Theorem 3.5.1

The following Lemmas 3.12.3 and 3.12.4 consider the situation that the network width N and s are not specified. These two lemmas prepares our proof for Theorem 3.5.1.

Lemma 3.12.3 *Let* $N_n = c_N[(L+1)s^* \log N^* + s^* \log((L+1)n/s^*)] \log^{2\delta}(n) \asymp n\varepsilon_n^{*2}$ *and* $s_n\lambda_s =$
$c_s[(L+1)s^* \log N^* + s^* \log((L+1)n/s^*)] \log^{2\delta}(n) \asymp n\varepsilon_n^{*2}$ *for some constant* c_N *and* c_s *(*N^**,* s^* *and* ε_n^* *are defined in Section 5). If the neural network width* N *and sparsity* s *follow some truncated priors with support* $\{1, \ldots, N_n\}$ *and* $\{0, \ldots, s_n\}$ *respectively, and this prior satisfies* $-\log \pi(N = N^*, s = s^*) = O(n\varepsilon_n^2)$*. Then similar results of Lemma 3.4.1 and Lemma 3.4.1 holds, that is for some* $C > 0$ *and* $C' > 0$*, we have*

$$\begin{aligned} &\int d^2(P_\theta, P_0)\widehat{q}(\theta)d\theta \leqslant C\varepsilon_n^{*2} + \frac{3}{n}\inf_{q(\theta)\in\mathcal{Q}}\Big\{KL(q(\theta)\|\pi(\theta)) + \int l_n(P_0, P_\theta)q(\theta)d\theta\Big\}, \text{ and} \\ &\inf_{q(\theta)\in\mathcal{Q}}\Big\{KL(q(\theta)\|\pi(\theta)) + \int l_n(P_0, P_\theta)q(\theta)d\theta\Big\} \leqslant C'n(\varepsilon_n^{*2} + r_n^* + \xi_n^*) \end{aligned} \tag{3.27}$$

hold with dominating probability.

Proof 9 To prove the first result of (3.27), similarly to the proof of Lemma 3.4.1, it is essential to show that there exists some testing function that achieves exponentially small error probability. This further requires a bounded covering number of $N(\varepsilon_n^*/19, \bigcup_{N=1}^{N_n}\bigcup_{s=0}^{s_n}\mathcal{F}(L, \boldsymbol{p}_N^L, s), d(\cdot,\cdot))$. Similar to (3.20), we have that

$$\begin{aligned} &N(\varepsilon_n^*/19, \bigcup_{N=1}^{N_n}\bigcup_{s=0}^{s_n}\mathcal{F}(L, \boldsymbol{p}_N^L, s), d(\cdot,\cdot)) \\ \leqslant &\log N(\sqrt{8}\sigma_\epsilon\varepsilon_n^*/19, \bigcup_{N=1}^{N_n}\bigcup_{s=0}^{s_n}\mathcal{F}(L, \boldsymbol{p}_N^L, s), \|\cdot\|_\infty) \\ \leqslant &\log(s_n) + \log(N_n) + (s_n+1)\log(\frac{38}{\sqrt{8}\sigma_\epsilon\varepsilon_n^*}(L+1)(12pN_n+1)^{2(L+1)}) \\ \leqslant &n\varepsilon_n^{*2}/4, \quad \text{given a large n,} \end{aligned}$$

where the last inequality holds due to the fact that $\log(N_n) \asymp \log n$, $s_n \log(1/\varepsilon_n^*) \asymp s_n \log n$ and $\lambda_s \geqslant a(L+1)\log n$ for some $a > 0$. Therefore, by the argument of Lemma 3.12.2, there still exists a testing function that separate P_0 and $\{P_\theta : d(P_0, P_\theta) \geqslant \varepsilon_n, P_\theta \in \bigcup_{N=1}^{N_n}\bigcup_{s=0}^{s_n}\mathcal{F}(L^*, \boldsymbol{p}_N^{L^*}, s)\}$ with exponentially small error probability. By the argument used in the proof of Lemma 3.4.1, implies that first result of (3.27) holds.

The proof of the second result of (3.27) follows the same argument used in Lemma 3.4.1. We can choose the $q^*(\theta, N, s) \in \mathcal{Q}_{N,s}$ as $q^*(N) = \delta_{N*}$, $q^*(s) = \delta_{s*}$, and $q^*(\theta|N^*, s^*) = q^*(\theta)$ as defined in (3.22). Trivially, (3.25) still holds, and $\text{KL}(q^*(\theta, N, s)\|\pi(\theta, N, s)) \leqslant nr_n^* - \log \pi(N = N^*, s = s^*) = O(n\varepsilon_n^{*2} + nr_n^*)$. It hence concludes the result. ■

The next Lemma is an improved result of Corollary 6.1 in Polson et al. 2018.

Lemma 3.12.4 *Under prior specification (13),*

$$\pi(N \geqslant N_n \text{ } or \text{ } s \geqslant s_n|D) \leqslant \exp\{-c_0 n\varepsilon_n^{*2}\},$$

where constant c_0 increases to infinity as c_s (defined in Lemma 3.12.3) increases.

Proof 10 Due to Lemma A.4 in Song and Liang 2017, it suffice to show that

$$\pi(N \geqslant N_n \text{ or } s \geqslant s_n) \leqslant \exp\{-c_1 n\varepsilon_n^{*2}\} \tag{3.28}$$

$$\log \frac{m(D)}{p_0(D)} \geqslant \exp\{-c_2 n\varepsilon_n^{*2}\}, \quad \text{w.h.p.} \tag{3.29}$$

where c_1 increases to infinity as c_s increases, $c_2 > 0$ is an absolute constant, $m(D) = \int p_\theta(D)d\pi(\theta)$ is the marginal density.

Inequality (3.28) is true, since

$$-\log \pi(N > N_n) \asymp N_n \log N_n \succ n\varepsilon_n^{*2} \text{ and}$$

$$-\log \pi(s > s_n) \geqslant C\lambda_s s_n \asymp n\varepsilon_n^{*2},$$

hold for some constant C.

To prove (3.29), it is suffice to find a subset $\mathcal{F}_s \subset \mathcal{F}$, such that $\pi(\mathcal{F}_s) \geqslant \exp\{-c_3 n\varepsilon_n^{*2}\}$ and w.h.p. $p_\theta(D)/p_0(D) \geqslant \exp\{-c_4 n\varepsilon_n^{*2}\}$ for any $p_\theta \in \mathcal{F}_s$. Such $\mathcal{F}_s$ can be defined as $\{f_\theta \in \mathcal{F}(L, \boldsymbol{p}^* = (12pN^*, \ldots, 12pN^*)', s^*) : \|f_\theta - f_0\|_\infty \leqslant \varepsilon_n^*\}$, First, we show that $p_\theta(D)/p_0(D) \geqslant \exp\{-c_4 n\varepsilon_n^{*2}\}$ for any $p_\theta \in \mathcal{F}_s$. Note that

$$\begin{aligned}
&-\log p_\theta(D)/p_0(D) \\
=&-\frac{1}{2\sigma_\epsilon^2}\sum_{\mathrm{i}=1}^{n}[(Y_{\mathrm{i}}-f_0(X_{\mathrm{i}}))^2-(Y_{\mathrm{i}}-f_\theta(X_{\mathrm{i}}))^2] \\
\leqslant&\frac{1}{2\sigma_\epsilon^2}[n\|f_\theta-f_0\|_\infty^2+2|\langle Y-f_0(X),f_\theta(X)-f_0(X)\rangle|].
\end{aligned}$$

Note that $Y-f_0(X)$ is a vector of i.i.d. normal $N(0,\sigma_\epsilon^2)$, then by concentration inequality, w.h.p,

$$|\langle X-f_0(X),f_\theta(X)-f_0(X)\rangle| \leqslant cn\varepsilon_n^{*2}$$

for some $c>0$, and we can conclude that w.h.p.,

$$\frac{p_\theta(D)}{p_0(D)} \geqslant \exp\{-c_4 n\varepsilon_n^{*2}\}$$

Second, we prove that $\boldsymbol{\pi}(\mathcal{F}_s) \geqslant \exp\{-c_3 n\varepsilon_n^{*2}\}$ in the following. By Condition 3.5.2, $\xi_n^* \asymp r_n^* = o(\varepsilon_n^{*2})$, hence there must exists a NN $\widehat{f}_{\widehat{\theta}} \in \mathcal{F}(L,\boldsymbol{p}^* s^*,\widehat{\gamma})$, where $\widehat{\gamma}$ denotes a specific pattern of nonzero links among $\widehat{\theta}$, s.t.

$$\|\hat{f}_{\hat{\theta}}-f_0\|_\infty \lesssim \varepsilon_n^*/2.$$

By triangle inequality,

$$\begin{aligned}
&\{f_\theta \in \mathcal{F}(L,\boldsymbol{p}^*,s^*):\|f_\theta-f_0\|_\infty \leqslant \varepsilon_n^*\} \\
\supset&\{f_\theta \in \mathcal{F}(L,\boldsymbol{p}^*,s^*,\widehat{\gamma}):\|f_\theta-\widehat{f}_{\widehat{\theta}}\|_\infty \leqslant \frac{\varepsilon_n^*}{2}\}.
\end{aligned}$$

Furthermore, from the proof of Lemma 10 of Schmidt-Hieber 2017, we have

$$\begin{aligned}
&\{f_\theta \in \mathcal{F}(L,\boldsymbol{p}^*,s^*,\widehat{\gamma}):\|f_\theta-\hat{f}_{\widehat{\theta}}\|_\infty \leqslant \frac{\varepsilon_n^*}{2}\} \\
\supset&\{f_\theta:\|\theta\|_\infty \leqslant 1 \text{ and } \|\theta-\hat{\theta}\|_\infty \leqslant \frac{\varepsilon_n^*}{2V(L+1)}\},
\end{aligned}$$

where $V=(L+1)(12pN^*+1)$.

Therefore,

$$\begin{aligned}&\pi\{f_\theta \in \mathcal{F}(L, \boldsymbol{p}^*, s^*) : \|f_\theta - f_0\|_\infty \leqslant \varepsilon_n^*\}\\>&\frac{\pi\{f_\theta \in \mathcal{F}(L, \boldsymbol{p}^*, s^*, \widehat{\gamma}) : \|f_\theta - \widehat{f}_{\widehat{\theta}}\|_\infty \leqslant \frac{\varepsilon_n^*}{2}\}}{\binom{T}{s*}}\\>&e^{-(L+1)s^*\log(12pN^*)}\pi\{\theta : \|\theta\|_\infty \leqslant 1 \text{ and } \|\theta - \hat{\theta}\|_\infty \leqslant \frac{\varepsilon_n^*}{2V(L+1)}\},\end{aligned}$$

where T denotes the total number of edge in network $\mathcal{F}(L, \boldsymbol{p}^*, s^*)$. Note that

$$\begin{aligned}&\pi\{\theta : \|\theta\|_\infty \leqslant 1 \text{ and } \|\theta - \hat{\theta}\|_\infty \leqslant \frac{\varepsilon_n^*}{2V(L+1)})\}\\\approx&\exp\{-s^*\log(\frac{2V(L+1)}{\varepsilon_n^*})\}.\end{aligned}$$

Therefore, it is sufficient to show that

$$\begin{aligned}&(L+1)s^*\log(12pN^*) + s^*\log(\frac{2(L+1)^2(12pN^*+1)}{\varepsilon_n^*})\\\leqslant&c_3 n\varepsilon_n^{*2},\end{aligned}$$

which hold trivially due to the definition of ε_n^*.

■

We are ready to prove Theorem 3.5.1.

Proof 11 Denote $\delta_{\widehat{N}}$ and $\delta_{\widehat{s}}$ be the degenerate VB posterior of N and s. We claim that with dominating probability,

$$\widehat{N} < N_n \text{ and } \widehat{s} < s_n. \tag{3.30}$$

Therefore, it will be equivalent to consider the truncated prior $\widetilde{\pi}(N) \propto \pi(N)1(N < N_n)$ and $\widetilde{\pi}(s) \propto \pi(s)1(s < s_n)$.

Note that

$$\begin{aligned}&-\log \pi(N = N^*) \leqslant -\log \widetilde{\pi}(N = N^*)\\ \leqslant&\lambda + \log N^*! - N^* \log \lambda \asymp N^* \log N^*\\ \leqslant&s^* \log N^* = O(n\varepsilon_n^{*2}),\end{aligned}$$

and

$$-\log \pi(s = s^*) = O(\lambda_s s^*) = O(n\varepsilon_n^{*2}).$$

Therefore, the conditions of Lemma 3.12.3 hold and we conclude the proof.

Recall $q^*(\theta, N, s) \in \mathcal{Q}_{N,s}$ which is defined in the proof of Lemma 3.12.3, and we prove (3.30) by showing that w.h.p.,

$$\mathrm{KL}(q^*(\theta, N, s)\|\pi(\theta, N, s|D)) \leqslant \mathrm{KL}(q(\theta, N, s)\|\pi(\theta, N, s|D)), \tag{3.31}$$

for any $q \in \mathcal{Q}_{N,s}$ whose marginal degenerate distribution of N is large than N_n or marginal degenerate distribution of s is greater than s_n. Note that

$$\begin{aligned}&\frac{1}{n}\mathrm{KL}(q^*(\theta, N, s)\|\pi(\theta, N, s|D))\\ =&\frac{1}{n}\mathrm{KL}(q^*(\theta, N, s)\|\pi(\theta, N, s)) + \frac{1}{n}\mathbb{E}_{q^*} \log \frac{p_0(D)}{p_\theta(D)}\\ +&\frac{1}{n} \log \frac{m(D)}{p_0(D)}.\end{aligned}$$

The sum of the first two terms in above equation, as shown in the proof of Lemma 3.12.3, is $O(\varepsilon_n^{*2} + r_n^*) = O(\varepsilon_n^{*2})$. For the third term, by LLN, it converges to constant $-\mathrm{KL}(P_0\|m) \leqslant 0$.

Due to Lemma 3.12.4, $\mathrm{KL}(q(\theta, N, s)\|\pi(\theta, N, s|D)) \geqslant c_0 n\varepsilon_n^{*2}$, and the constant c_0 increases to infinity as c_s increases. Therefore, providing a sufficiently large c_s, (3.31) holds.

■

3.12.4 Remarks for proofs of Corollaries 3.6.1-3.6.4.

The proofs for Corollaries 3.6.1 and 3.6.3 are straightforward, and they are directly implied by Theorem 3.4.1.

For the proofs of Corollaries 3.6.2 and 3.6.4, we comment that Theorem 3.5.1 actually holds for any (N^*, s^*) which satisfies Conditions 3.5.1, 3.5.3 and $\xi_n^* = O(r_n^*)$, but is not necessarily the exact minimization of $r_n^* + \xi_n^*$. Therefore, in this case we can still use Theorem 3.5.1 to prove Corollaries 3.6.2 and 3.6.4.

3.12.5 Proof of Theorem 3.7.1

Proof 12 For any $M_n \to \infty$, there always exists some $\widetilde{M}_n$ satistfying that $1 < \widetilde{M}_n = O(M_n)$ and $\gamma_n \widetilde{M}_n \widetilde{\varepsilon}_n^2 = o(1)$.

Then, for any $\theta \in \mathcal{G} \cap \{\theta : L_2^2(f_0, f_\theta) \geqslant \widetilde{M}_n \widetilde{\varepsilon}_n^2\}$,

$$\begin{aligned} d^2(P_\theta, P_0) &\geqslant \int_S (1 - \exp\{-(f_\theta(x) - f_0(x))^2/8\sigma_\epsilon^2\}) dP(x) \\ &\geqslant \frac{(1 - \exp\{-\gamma_n L_2^2(f_0, f_\theta)/8\sigma_\epsilon^2\})}{\gamma_n L_2^2(f_0, f_\theta)} \int_S (f_\theta(x) - f_0(x))^2 dP(x) \\ &\geqslant \frac{(1 - \exp\{-\gamma_n L_2^2(f_0, f_\theta)/8\sigma_\epsilon^2\})}{\gamma_n} \kappa \\ &\geqslant \frac{(1 - \exp\{-\gamma_n \widetilde{M}_n \widetilde{\varepsilon}_n^2/8\sigma_\epsilon^2\})}{\gamma_n} \kappa \geqslant c_M \widetilde{M}_n \widetilde{\varepsilon}_n^2, \end{aligned} \tag{3.32}$$

for some constant $c_M > 0$, where the second inequality holds since $|f_\theta(X) - f_0(X)|^2$ is upper bounded by $\gamma_n L_2^2(f_0, f_\theta)$ on $\mathcal{S}$, and the last inequality is due to the fact that $\gamma_n \widetilde{M}_n \widetilde{\varepsilon}_n^2 = o(1)$. (3.32) implies

$$\mathcal{G} \cap \{L_2^2(f_0, f_\theta) \geqslant \widetilde{M}_n \widetilde{\varepsilon}_n^2\} \subset \{d^2(P_\theta, P_0) \geqslant c_M \widetilde{M}_n \widetilde{\varepsilon}_n^2\}. \tag{3.33}$$

By Theorem 3.4.1, w.h.p.,

$$\int d^2(P_\theta, P_0) \widehat{q}(\theta) = O(\widetilde{\varepsilon}_n^2),$$

which implies that

$$\int_{d^2(P_\theta, P_0) \geqslant c_M \widetilde{M}_n \tilde{\varepsilon}_n^2} \widehat{q}(\theta) = O(1/\widetilde{M}_n) = o(1).$$

Combined with (3.33)

$$\begin{aligned} \int_{\mathcal{G} \cap \{L_2^2(f_0, f_\theta) > M_n \tilde{\varepsilon}_n^2\}} \widehat{q}(\theta) &\leqslant \int_{\mathcal{G} \cap \{L_2^2(f_0, f_\theta) > \widetilde{M}_n \tilde{\varepsilon}_n^2\}} \widehat{q}(\theta) \\ \leqslant \int_{d^2(P_\theta, P_0) \geqslant c_M \widetilde{M}_n \tilde{\varepsilon}_n^2} \widehat{q}(\theta) &= O(1/\widetilde{M}_n) = o(1), w.h.p. \end{aligned}$$

■

4. COMPUTATIONALLY EFFICIENT SPIKE AND SLAB PRIOR

4.1 Introduction

In this chapter, we improve the variational inference dicussed in Chapter 3 by placing a computationally efficient prior while remaining theoretically soundness. Specifically, the prior distribution for the inclusion variable γ_{i} follows independent Bernoulli distribution. In addition, we will only consider Gaussian slab distribution, since VI with uniform slab lacks of practical implementation and only possesses theoretical significance.

More importantly, with carefully chosen hyperparameter values, especially the prior probability that each edge is active, we establish the variational posterior consistency, and the corresponding convergence rate strikes the balance of statistical estimation error, variational error and the approximation error.

The theoretical results are validated by various simulations and real applications. Empirically we also demonstrate that the proposed method possesses good performance of variable selection and uncertainty quantification. While (Feng et al. 2017; Liang et al. 2018; Ye et al. 2018) only considered the neural network with single hidden layer for variable selection, we observe correct support recovery for neural networks with multi-layer networks.

4.2 Alternative Spike-and-slab Prior

As in Chapter 3, we aim to approximate f_0 in the generative model (3.1) by a sparse neural network. Specifically, given a network structure, i.e. the depth L and the width $\boldsymbol{p}$, f_0 is approximated by DNN models f_θ with sparse parameter vector $\theta \in \Theta = \mathbb{R}^T$. From a Bayesian perspective, we impose the following spike-and-slab prior (George et al. 1993; Ishwaran et al. 2005) on θ:

$$\theta_{\mathrm{i}} | \gamma_{\mathrm{i}} \sim \gamma_{\mathrm{i}} \mathcal{N}(0, \sigma_0^2) + (1 - \gamma_{\mathrm{i}}) \delta_0, \quad \gamma_{\mathrm{i}} \sim \mathrm{Bern}(\lambda), \tag{4.1}$$

for $\mathrm{i} = 1, \ldots, T$, where λ and σ_0^2 are hyperparameters representing the prior inclusion probability and the prior Gaussian variance, respectively. The choice of σ_0^2 and λ play an important role in sparse Bayesian learning, and in Section 4.3, we will establish theoretical guarantees for the variational inference procedure under proper deterministic choices of σ_0^2 and λ. Alternatively, hyperparameters may be chosen via an Empirical Bayesian (EB) procedure, but it is beyond the scope of this work. We assume $\mathcal{Q}$ is in the same family of spike-and-slab laws:

$$\theta_{\mathrm{i}}|\gamma_{\mathrm{i}} \sim \gamma_{\mathrm{i}}\mathcal{N}(\mu_{\mathrm{i}}, \sigma_{\mathrm{i}}^2) + (1-\gamma_{\mathrm{i}})\delta_0, \quad \gamma_{\mathrm{i}} \sim \mathrm{Bern}(\phi_{\mathrm{i}}) \tag{4.2}$$

for $\mathrm{i} = 1, \ldots, T$, where $0 \leqslant \phi_{\mathrm{i}} \leqslant 1$.

Comparing to pruning approaches e.g. Molchanov et al. 2017; Frankle et al. 2018; Zhu et al. 2018 that don't pursue sparsity among bias parameter b_{i}'s, the Bayesian modeling induces posterior sparsity for both weight and bias parameters.

Polson et al. 2018; Chérief-Abdellatif 2020 as well as Chapter 3 imposed sparsity specification as follows $\Theta(L, \boldsymbol{p}, s) = \{\theta \text{ as in model (1.5)} : ||\theta||_0 \leqslant s\}$ that not only posts great computational challenges, but also requires tuning for optimal sparsity level s. For example, it has been showed in Chapter 3 that given s, two error terms occur in the variation DNN inference: 1) the variational error $r_n(L, \boldsymbol{p}, s)$ caused by the variational Bayes approximation to the true posterior distribution and 2) the approximation error $\xi_n(L, \boldsymbol{p}, s)$ between f_0 and the best bounded-weight s-sparsity DNN approximation of f_0. Both error terms r_n and ξ_n depend on s (and their specific forms are given in next section). Generally speaking, as the model capacity (i.e., s) increases, r_n will increase and ξ_n will decrease. Hence the optimal choice s^* that strikes the balance between these two is

$$s^* = \operatorname*{argmin}_s \{r_n(L, \boldsymbol{p}, s) + \xi_n(L, \boldsymbol{p}, s)\}.$$

Therefore, one needs to develop a selection criteria for $\widehat{s}$ such that $\widehat{s} \approx s^*$. In contrast, our modeling in this chapter directly works on the whole sparsity regime without pre-specifying s, and is shown later to be capable of automatically attaining the same rate of convergence as if the optimal s^* were known.

4.3 Theoretical Results

In this section, we will establish the contraction rate of the variational sparse DNN procedure, without knowing s^*. For simplicity, we only consider equal-width neural network similar as in Chapter 3.

The following assumptions are imposed:

Condition 4.3.1 $p_i \equiv N \in \mathbb{Z}^+$ *that can depend on* n, *and* $\lim T = \infty$.

Condition 4.3.2 $\sigma(x)$ *is 1-Lipschitz continuous.*

Condition 4.3.3 *The hyperparameter* σ_0^2 *is set to be some constant, and* λ *satisfies* $\log(1/\lambda) = O\{(L+1)\log N + \log(p\sqrt{n/s^*})\}$ *and* $\log(1/(1-\lambda)) = O((s^*/T)\{(L+1)\log N + \log(p\sqrt{n/s^*})\})$.

Condition 4.3.2 is very mild, and includes ReLU, sigmoid and tanh. Note that Condition 4.3.3 gives a wide range choice of λ, even including the choice of λ independent of s^* (See Theorem 4.3.1 below).

We first state a lemma on an upper bound for the negative ELBO. Denote the log-likelihood ratio between p_0 and p_θ as $l_n(P_0, P_\theta) = \log(p_0(D)/p_\theta(D)) = \sum_{i=1}^n \log(p_0(D_i)/p_\theta(D_i))$. Given some constant $B > 0$, we define

$$\begin{aligned} r_n^* &:= r_n(L, N, s^*) = ((L+1)s^*/n)\log N + (s^*/n)\log(p\sqrt{n/s^*}), \\ \xi_n^* &:= \xi_n(L, N, s^*) = \inf_{\theta\in\Theta(L,\boldsymbol{p},s^*), \|\theta\|_\infty \leqslant B} ||f_\theta - f_0||_\infty^2. \end{aligned}$$

Recall that $r_n(L, N, s)$ and $\xi_n(L, N, s)$ denote the variational error and the approximation error.

Lemma 4.3.1 *Under Condition 4.3.1-4.3.3, then with dominating probability,*

$$\inf_{q(\theta)\in\mathcal{Q}} \left\{ KL(q(\theta)||\pi(\theta|\lambda)) + \int_\Theta l_n(P_0, P_\theta)q(\theta)d\theta \right\} \leqslant Cn(r_n^* + \xi_n^*) \tag{4.3}$$

where C *is either some positive constant if* $\lim n(r_n^* + \xi_n^*) = \infty$, *or any diverging sequence if* $\limsup n(r_n^* + \xi_n^*) \neq \infty$.

Noting that $\mathrm{KL}(q(\theta)||\boldsymbol{\pi}(\theta|\lambda))+\int_\Theta l_n(P_0,P_\theta)q(\theta)(d\theta)$ is the negative ELBO up to a constant, we therefore show the optimal loss function of the proposed variational inference is bounded.

The next lemma investigates the convergence of the variational distribution under the Hellinger distance, which is defined as

$$d^2(P_\theta,P_0)=\mathbb{E}_X\Big(1-\exp\{-[f_\theta(X)-f_0(X)]^2/(8\sigma_\epsilon^2)\}\Big).$$

In addition, let $s_n=s^*\log^{2\delta-1}(n)$ for any $\delta>1$. An assumption on s^* is required to strike the balance between r_n^* and ξ^*:

Condition 4.3.4 $\max\{s^*\log(p\sqrt{n/s^*},(L+1)s^*\log N\}=o(n)$ *and* $r_n^*\asymp\xi_n^*$.

Lemma 4.3.2 *Under Conditions 4.3.1-4.3.4, if* σ_0^2 *is set to be constant and* $\lambda\leqslant T^{-1}\exp\{-Mnr_n^*/s_n\}$ *for any positive diverging sequence* $M\to\infty$*, then with dominating probability, we have*

$$\int_\Theta d^2(P_\theta,P_0)\widehat{q}(\theta)d\theta\leqslant C\varepsilon_n^{*2}+\frac{3}{n}\inf_{q(\theta)\in\mathcal{Q}}\Big\{KL(q(\theta)||\boldsymbol{\pi}(\theta|\lambda))+\int_\Theta l_n(P_0,P_\theta)q(\theta)d\theta\Big\},\qquad(4.4)$$

where C *is some constant, and*

$$\varepsilon_n^*:=\varepsilon_n(L,N,s^*)=\sqrt{r_n(L,N,s^*)}\log^\delta(n),\ \textit{for any}\ \delta>1.$$

Remark The result (4.4) is of exactly the same form as in the existing literature (Pati et al. 2018), but it is not trivial in the following sense. The existing literature require the existence of a global testing function that separates P_0 and $\{P_\theta : d(P_\theta,P_0)\geqslant\varepsilon_n^*\}$ with exponentially decay rate of Type I and Type II errors. If such a testing function exists only over a subset $\Theta_n\subset\Theta$ (which is the case for our DNN modeling), then the existing result (Yang, Pati, et al. 2020) can only characterize the VB posterior contraction behavior within Θ_n, but not over the whole parameter space Θ. Therefore our result, which characterizes the convergence behavior for the overall VB posterior, represents a significant improvement beyond those works.

The above two lemmas together imply the following guarantee for VB posterior:

Theorem 4.3.1 *Let* σ_0^2 *be a constant and* $-\log\lambda = \log(T) + \delta[(L+1)\log N + \log\sqrt{n}p]$ *for any constant* $\delta > 0$*. Under Conditions 4.3.1-4.3.2, 4.3.4, we have with high probability*

$$\int_{\Theta} d^2(P_\theta, P_0)\widehat{q}(\theta)d\theta \leqslant C\varepsilon_n^{*2} + C'(r_n^* + \xi_n^*),$$

where C *is some positive constant and* C' *is any diverging sequence.*

The ε_n^{*2} denotes the estimation error from the statistical estimator for P_0. The variational Bayes convergence rate consists of estimating error, i.e., ε_n^{*2}, variational error, i.e., r_n^*, and approximation error, i.e., ξ_n^*. Given that the former two errors have only logarithmic difference, our convergence rate actually strikes the balance among all three error terms. The derived convergence rate has an explicit expression in terms of the network structure based on the forms of ε_n^*, r_n^* and ξ_n^*, in contrast with general convergence results in Pati et al. 2018; Zhang and Gao 2019; Yang, Pati, et al. 2020.

Remark Theorem 4.3.1 provides a specific choice for λ, which can be relaxed to the general conditions on λ in Lemma 4.3.2. In contrast to the heuristic choices such as $\lambda = \exp(-2\log n)$ BIC; Hubin et al. 2019, this theoretically justified choice incorporates knowledge of input dimension, network structure and sample size. Such an λ will be used in our numerical experiments in Section 4.5, but readers shall be aware of that its theoretical validity is only justified in an asymptotic sense.

Remark The convergence rate is derived under Hellinger metric, which is of less practical relevance than L_2 norm representing the common prediction error. One may obtain a convergence result under L_2 norm via a VB truncation (refer to Section 4.7.3, Theorem 4.7.1).

Remark If f_0 is an α-Hölder-smooth function with fixed input dimension p, then by choosing some $L \asymp \log n$, $N \asymp n/\log n$, combining with the approximation result Schmidt-Hieber 2017, Theorem 3, our theorem ensures rate-minimax convergence up to a logarithmic term.

4.4 Implementation

As in Chapter 3, to conduct optimization of negative ELBO via stochastic gradient optimization, we need to utilize the Gumbel-softmax approximation. Rewrite the loss function Ω as

$$-\mathbb{E}_{q(\theta|\gamma)q(\gamma)}[\log p_\theta(D)] + \sum_{\mathrm{i}=1}^{T} \mathrm{KL}(q(\gamma_\mathrm{i})||\boldsymbol{\pi}(\gamma_\mathrm{i})) + \sum_{\mathrm{i}=1}^{T} q(\gamma_\mathrm{i}=1)\mathrm{KL}(\mathcal{N}(\mu_\mathrm{i},\sigma_\mathrm{i}^2)||\mathcal{N}(0,\sigma_0^2)). \quad (4.5)$$

Apply Gumbel-softmax approximation (Jang et al. 2017; Maddison et al. 2017) to $\gamma_\mathrm{i} \sim \mathrm{Bern}(\phi_\mathrm{i})$, that is

$$\widetilde{\gamma}_\mathrm{i} = g_\tau(\phi_\mathrm{i}; u_\mathrm{i}) = \frac{1}{1+\exp(-(\log\frac{\phi_\mathrm{i}}{1-\phi_\mathrm{i}} + \log\frac{u_\mathrm{i}}{1-u_\mathrm{i}})/\tau)}, \quad u_\mathrm{i} \sim \mathcal{U}(0,1),$$

where τ is called the temperature and is chosen as 0.5 in the experiment. Besides, the normal variable $\mathcal{N}(\mu_\mathrm{i},\sigma_\mathrm{i}^2)$ is reparameterized by $\mu_\mathrm{i}+\sigma_\mathrm{i}\epsilon_\mathrm{i}$ for $\epsilon_\mathrm{i} \sim \mathcal{N}(0,1)$.

Recall that $\mathcal{Q}$ is reparameterized as $q_\omega \overset{d}{=} g(\omega,\nu)$ for some differentiable function g and random variable ν, then the stochastic estimator of the negative ELBO $\Omega(\omega)$ and its gradient are

$$\begin{aligned}
\widetilde{\Omega}^m(\omega) &= -\frac{n}{m}\frac{1}{K}\sum_{\mathrm{i}=1}^{m}\sum_{k=1}^{K}\log p_{g(\omega,\nu_k)}(D_\mathrm{i}) + \mathrm{KL}(q_\omega(\theta)||\boldsymbol{\pi}(\theta)),\\
\nabla_\omega\widetilde{\Omega}^m(\omega) &= -\frac{n}{m}\frac{1}{K}\sum_{\mathrm{i}=1}^{m}\sum_{k=1}^{K}\nabla_\omega\log p_{g(\omega,\nu_k)}(D_\mathrm{i}) + \nabla_\omega\mathrm{KL}(q_\omega(\theta)||\boldsymbol{\pi}(\theta)),
\end{aligned} \quad (4.6)$$

where D_i's are randomly sampled data points and ν_k's are iid copies of ν. Here, m and K are minibatch size and Monte Carlo sample size, respectively.

The complete variational inference procedure with Gumbel-softmax approximation is stated below.

Algorithm 3 Variational inference for sparse BNN with normal slab distribution.

```
parameters: $\omega = (\mu, \sigma', \phi')$ ,
where $\sigma_\mathrm{i} = \log(1 + \exp(\sigma'_\mathrm{i}))$, $\phi_\mathrm{i} = (1 + \exp(\phi'_\mathrm{i}))^{-1}$, for $\mathrm{i} = 1, \ldots, T$
repeat
    $D^m \leftarrow$ Randomly draw a minibatch of size $m$ from $D$
    $\epsilon_\mathrm{i}, u_\mathrm{i} \leftarrow$ Randomly draw $K$ samples from $\mathcal{N}(0, 1)$ and $\mathcal{U}(0, 1)$
    $\widetilde{\Omega}^m(\omega) \leftarrow$ Use (4.6) with ($D^m$, $\omega$, $\epsilon$, $u$); Use $\gamma$ in the forward pass
    $\nabla_\omega \widetilde{\Omega}^m(\omega) \leftarrow$ Use (4.6) with ($D^m$, $\omega$, $\epsilon$, $u$); Use $\widetilde{\gamma}$ in the backward pass
    $\omega \leftarrow$ Update with $\nabla_\omega \widetilde{\Omega}^m(\omega)$ using gradient descent algorithms (e.g. SGD or Adam)
until convergence of $\widetilde{\Omega}^m(\omega)$
return $\omega$
```

4.5 Experiments

We evaluate the empirical performance of the proposed variational inference through simulation study and MNIST data application. For the simulation study, we consider a teacher-student framework and a nonlinear regression function, by which we justify the consistency of the proposed method and validate the proposed choice of hyperparameters. As a byproduct, the performance of uncertainty quantification and the effectiveness of variable selection will be examined as well.

For all the numerical studies, we let $\sigma_0^2 = 2$, the choice of λ follows Theorem 4.3.1 (denoted by λ_{opt}): $\log(\lambda_{opt}^{-1}) = \log(T) + 0.1[(L+1)\log N + \log\sqrt{n}p]$. The remaining details of implementation (such as initialization, choices of K, m and learning rate) are provided in the Section. We will use VB posterior mean estimator $\widehat{f}_H = \sum_{h=1}^{H} f_{\theta_h}/H$ to assess the prediction accuracy, where $\theta_h \sim \widehat{q}(\theta)$ are samples drawn from the VB posterior and $H = 30$. The posterior network sparsity is measured by $\widehat{s} = \sum_{\mathrm{i}=1}^{T} \phi_\mathrm{i}/T$. Input nodes who have connection with $\phi_\mathrm{i} > 0.5$ to the second layer is selected as relevant input variables, and we report the corresponding false positive rate (FPR) and false negative rate (FNR) to evaluate the variable selection performance of our method.

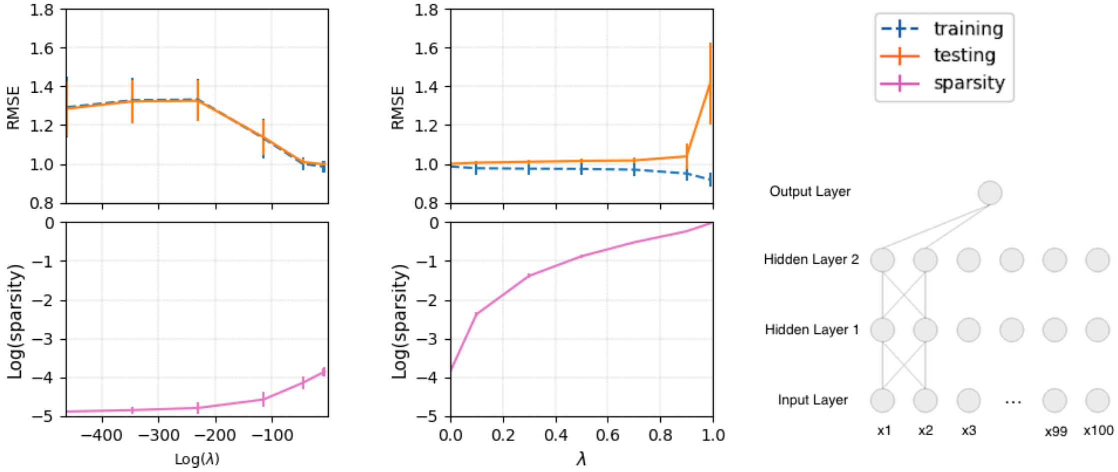
Our method will be compared with the dense variational BNN (VBNN) (Blundell et al. 2015) with independent centered normal prior and independent normal variational distribution, the AGP pruner (Zhu et al. 2018), the Lottery Ticket Hypothesis (LOT) (Frankle et al. 2018), the variational dropout (VD) (Molchanov et al. 2017) and the Horseshoe BNN (HS-BNN) (Ghosh, Yao, et al. 2018). In particular, VBNN can be regarded as a baseline method without any sparsification or compression. All reported simulation results are based on 30 replications (except that we use 60 replications for interval estimation coverages). Note that the sparsity level in methods AGP and LOT are user-specified. Hence, in simulation studies, we try a grid search for AGP and LOT, and only report the ones that yield highest testing accuracy. Furthermore, note that FPR and FNR are not calculated for HS-BNN since it only sparsifies the hidden layers nodewisely.

4.5.1 Simulation I: Teacher-student networks setup

We consider two teacher network settings for f_0: (A) densely connected with a structure of 20-6-6-1, $p = 20$, $n = 3000$, $\sigma(x) = \text{sigmoid}(x)$, $X \sim \mathcal{U}([-1,1]^{20})$, $\epsilon \sim \mathcal{N}(0,1)$ and network parameter $\theta_{\rm i}$ is randomly sampled from $\mathcal{U}(0,1)$; (B) sparsely connected as shown in Figure 4.1 (c), $p = 100$, $n = 500$, $\sigma(x) = \tanh(x)$, $X \sim \mathcal{U}([-1,1]^{100})$ and $\epsilon \sim \mathcal{N}(0,1)$, the network parameter $\theta_{\rm i}$'s are fixed (refer to Section 4.8 for details).

Table 4.1. Simulation results for Simulation I. SVBNN represents our sparse variational BNN method. The sparsity levels specified for AGP are 30% and 5%, and for LOT are 10% and 5%, respectively for the two cases.

		RMSE		Input variable selection			
	Method	**Train**	**Test**	**FPR(%)**	**FNR(%)**	**95% Coverage (%)**	**Sparsity(%)**
Dense	SVBNN	1.01 ± 0.02	1.01 ± 0.00	-	-	97.5 ± 1.71	6.45 ± 0.83
	VBNN	1.00 ± 0.02	1.00 ± 0.00	-	-	91.4 ± 3.89	100 ± 0.00
	VD	0.99 ± 0.02	1.01 ± 0.00	-	-	76.4 ± 4.75	28.6 ± 2.81
	HS-BNN	0.98 ± 0.02	1.02 ± 0.01	-	-	83.5 ± 0.78	64.9 ± 24.9
	AGP	0.99 ± 0.02	1.01 ± 0.00	-	-	-	30.0 ± 0.00
	LOT	1.04 ± 0.01	1.02 ± 0.00	-	-	-	10.0 ± 0.00
Sparse	SVBNN	0.99 ± 0.03	1.00 ± 0.01	0.00 ± 0.00	0.00 ± 0.00	96.4 ± 4.73	2.15 ± 0.25
	VBNN	0.92 ± 0.05	1.53 ± 0.17	100 ± 0.00	0.00 ± 0.00	90.7 ± 8.15	100 ± 0.00
	VD	0.86 ± 0.04	1.07 ± 0.03	72.9 ± 6.99	0.00 ± 0.00	75.5 ± 7.81	20.8 ± 3.08
	HS-BNN	0.90 ± 0.04	1.29 ± 0.04	-	-	67.0 ± 8.54	32.1 ± 20.1
	AGP	1.01 ± 0.03	1.02 ± 0.00	16.9 ± 1.81	0.00 ± 0.00	-	5.00 ± 0.00
	LOT	0.96 ± 0.01	1.04 ± 0.01	19.5 ± 2.57	0.00 ± 0.00	-	5.00 ± 0.00



(a) $\lambda \leqslant \lambda_{opt}$.　(b) $\lambda \geqslant \lambda_{opt}$.　(c) Sparse teacher network.

Figure 4.1. (a) $\lambda = \{10^{-200}, 10^{-150}, 10^{-100}, 10^{-50}, 10^{-20}, 10^{-5}, \lambda_{opt}\}$. (b) $\lambda = \{\lambda_{opt}, 0.1, 0.3, 0.5, 0.7, 0.9, 0.99\}$. (c) The structure of the target sparse teacher network. Please note that the x axes of figures (a) and (b) are in different scales.

First, we examine the impact of different choices of λ on our VB sparse DNN modeling. A set of different λ values are used, and for each λ, we compute the training square root MSE (RMSE) and testing RMSE based on $\widehat{f}_H$. Results for the simulation setting (B) are plotted in Figure 4.1 along with error bars (Refer to Section 4.8 for the plot under the simulation setting (A)). The figure shows that as λ increases, the resultant network becomes denser and the training RMSE monotonically decreases, while testing RMSE curve is roughly U-shaped. In other words, an overly small λ leads to over-sparsified DNNs with insufficient expressive power, and an overly large λ leads to overfitting DNNs. The suggested λ_{opt} successfully locates in the valley of U-shaped testing curve, which empirically justifies our theoretical choice of λ_{opt}.

We next compare the performance of our method (with λ_{opt}) to the benchmark methods, and present results in Table 4.1. For the dense teacher network (A), our method leads to the most sparse structure with comparable prediction error; For the sparse teacher network (B), our method not only achieves the best prediction accuracy, but also always selects the correct set of relevant input variables. Besides, we also explore uncertainty quantification of our methods, by studying the coverage of 95% Bayesian predictive intervals (refer to Section

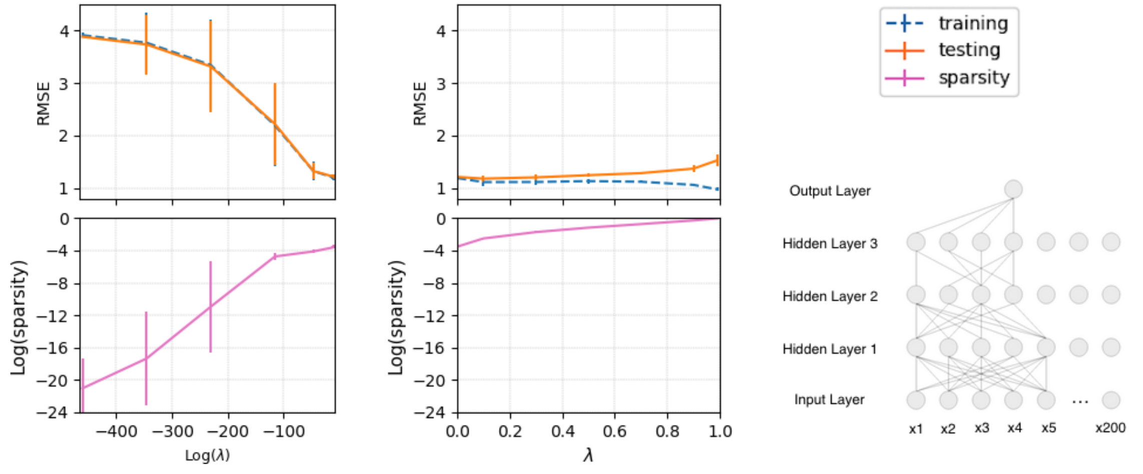
4.8 for details). Table 4.1 shows that our method obtains coverage rates slightly higher than the nominal levels while other (Bayesian) methods suffer from undercoverage problems.

4.5.2 Simulation II: Sparse nonlinear function

Consider the following sparse function f_0:

$$f_0(x_1,\dots,x_{200}) = \frac{7x_2}{1+x_1^2} + 5\sin(x_3x_4) + 2x_5, \quad \epsilon \sim \mathcal{N}(0,1), \tag{4.7}$$

all covariates are iid $\mathcal{N}(0,1)$ and data set contains $n = 3000$ observations. A ReLU network with $L = 3$ and $N = 7$ is used. Similar to the simulation I, we study the impact of λ, and results in Figure 4.2 justify that λ_{opt} is a reasonable choice. Table 4.2 compares the performances of our method (under λ_{opt}) to the competitive methods. Our method exhibits the best prediction power with minimal connectivity, among all the methods. In addition, our method achieves smallest FPR and acceptable FNR for input variable selection. In comparison, other methods select huge number of false input variables. Figure 4.2 (c) shows the selected network (edges with $\phi_i > 0.5$) in one replication that correctly identifies the input variables.



(a) $\lambda \leqslant \lambda_{opt}$.
(b) $\lambda \geqslant \lambda_{opt}$.
(c) Selected network structure.

Figure 4.2. (a) $\lambda = \{10^{-200}, 10^{-150}, 10^{-100}, 10^{-50}, 10^{-20}, 10^{-5}, \lambda_{opt}\}$. (b) $\lambda = \{\lambda_{opt}, 0.1, 0.3, 0.5, 0.7, 0.9, 0.99\}$. (c) A selected network structure for (4.7).

Table 4.2. Results for Simulation II. The sparsity levels selected for AGP and LOT are both 30%.

Method	Train RMSE	Test RMSE	FPR(%)	FNR(%)	Sparsity(%)
SVBNN	1.19 ± 0.05	1.21 ± 0.05	0.00 ± 0.21	16.0 ± 8.14	2.97 ± 0.48
VBNN	0.96 ± 0.06	1.99 ± 0.49	100 ± 0.00	0.00 ± 0.00	100 ± 0.00
VD	1.02 ± 0.05	1.43 ± 0.19	98.6 ± 1.22	0.67 ± 3.65	46.9 ± 4.72
HS-BNN	1.17 ± 0.52	1.66 ± 0.43	-	-	41.1 ± 36.5
AGP	1.06 ± 0.08	1.58 ± 0.11	82.7 ± 3.09	1.33 ± 5.07	30.0 ± 0.00
LOT	1.08 ± 0.09	1.44 ± 0.14	83.6 ± 2.94	0.00 ± 0.00	30.0 ± 0.00

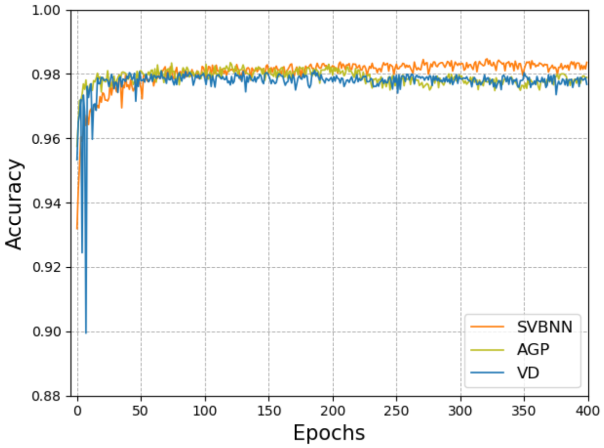


Figure 4.3. Testing accuracy for MNIST

MNIST application.

We evaluate the performance of our method on MNIST data for classification tasks, by comparing with benchmark methods. A 2-hidden layer DNN with 512 neurons in each layer is used. We compare the testing accuracy of our method (with λ_{opt}) to the benchmark methods at different epochs using the same batch size (refer to Section 4.8 for details). Figure 4.3 shows our method achieves best accuracy as epoch increases, and the final sparsity level for SVBNN, AGP and VD are 5.06%, 5.00% and 2.28%.

In addition, an illustration of our method's capability for uncertainty quantification on MNIST can be found in Section 4.8, where additional experimental results on UCI regression datasets can also be found.

4.6 Conclusion and Discussion

We proposed a variational inference method for deep neural networks under spike-and-slab priors with theoretical guarantees. Future direction could be investigating the theory behind choosing hyperparamters via the EB estimation instead of deterministic choices.

Furthermore, extending the current results to more complicated networks (convolutional neural network, residual network, etc.) is not trivial. Conceptually, it requires the design of structured sparsity (e.g., group sparsity in Neklyudov et al. 2017) to fulfill the goal of faster prediction. Theoretically, it requires deeper understanding of the expressive ability (i.e. approximation error) and capacity (i.e., packing or covering number) of the network model space. For illustration purpose, we include an example of Fashion-MNIST task using convolutional neural network in Section 4.8.5, and it demonstrates the usage of our method on more complex networks in practice.

4.7 Main Proofs

In this section, the detailed proofs for the theoretical results are provided.

4.7.1 Proof of Lemma 4.1

As a technical tool for the proof, we first restate the Lemma 6.1 in Chérief-Abdellatif and Alquier 2018 as follows.

Lemma 4.7.1 *For any $K > 0$, the KL divergence between any two mixture densities $\sum_{k=1}^K w_k g_k$ and $\sum_{k=1}^K \tilde{w}_k \tilde{g}_k$ is bounded as*

$$KL(\sum_{k=1}^K w_k g_k || \sum_{k=1}^K \tilde{w}_k \tilde{g}_k) \leqslant KL(\boldsymbol{w}||\tilde{\boldsymbol{w}}) + \sum_{k=1}^K w_k KL(g_k||\tilde{g}_k),$$

where $KL(\boldsymbol{w}||\tilde{\boldsymbol{w}}) = \sum_{k=1}^K w_k \log \frac{w_k}{\tilde{w}_k}$.

We begin the proof of Lemma 4.1

Proof 13 It suffices to construct some $q^*(\theta) \in \mathcal{Q}$, such that w.h.p,

$$\begin{aligned}&\mathrm{KL}(q^*(\theta)||\boldsymbol{\pi}(\theta|\lambda)) + \int_\Theta l_n(P_0, P_\theta) q^*(\theta)(d\theta)\\ \leqslant & C_1 n r_n^* + C_1' n \inf_\theta ||f_\theta - f_0||_\infty^2 + C_1' n r_n^*,\end{aligned}$$

where C_1, C_1' are some positive constants if $\lim n(r_n^* + \xi_n^*) = \infty$, or any diverging sequences if $\limsup n(r_n^* + \xi_n^*) \neq \infty$.

Recall $\theta^* = \arg\min_{\theta \in \Theta(L, \boldsymbol{p}, s^*, B)} ||f_\theta - f_0||_\infty^2$, then $q^*(\theta) \in \mathcal{Q}$ can be constructed as

$$\mathrm{KL}(q^*(\theta)||\boldsymbol{\pi}(\theta|\lambda)) \leqslant C_1 n r_n^*, \tag{4.8}$$

$$\int_\Theta ||f_\theta - f_{\theta*}||_\infty^2 q^*(\theta)(d\theta) \leqslant r_n^*. \tag{4.9}$$

We define $q^*(\theta)$ as follows, for i $= 1, \ldots, T$:

$$\begin{aligned}&\theta_{\mathrm{i}}|\gamma_{\mathrm{i}}^* \sim \gamma_{\mathrm{i}}^* \mathcal{N}(\theta_{\mathrm{i}}^*, \sigma_n^2) + (1 - \gamma_{\mathrm{i}}^*)\delta_0,\\ &\gamma_{\mathrm{i}}^* \sim \mathrm{Bern}(\phi_{\mathrm{i}}^*),\\ &\phi_{\mathrm{i}}^* = 1(\theta_{\mathrm{i}}^* \neq 0),\end{aligned} \tag{4.10}$$

To prove (4.8), denote Γ^T as the set of all possible binary inclusion vectors with length T, then $q^*(\theta)$ and $\boldsymbol{\pi}(\theta|\lambda)$ could be written as mixtures

$$q^*(\theta) = \sum_{\gamma \in \Gamma^T} 1(\gamma = \gamma^*) \prod_{\mathrm{i}=1}^T \gamma_{\mathrm{i}} \mathcal{N}(\theta_{\mathrm{i}}^*, \sigma_n^2) + (1 - \gamma_{\mathrm{i}})\delta_0,$$

and

$$\boldsymbol{\pi}(\theta|\lambda) = \sum_{\gamma \in \Gamma^T} \boldsymbol{\pi}(\gamma) \prod_{\mathrm{i}=1}^T \gamma_{\mathrm{i}} \mathcal{N}(0, \sigma_0^2) + (1 - \gamma_{\mathrm{i}})\delta_0,$$

where $\boldsymbol{\pi}(\gamma)$ is the probability for vector γ under prior distribution $\boldsymbol{\pi}$. Then,

$$
\begin{aligned}
&\mathrm{KL}(q^*(\theta)||\boldsymbol{\pi}(\theta|\lambda))\\
\leqslant&\log\frac{1}{\boldsymbol{\pi}(\gamma^*)}+\sum_{\gamma\in\Gamma^T}1(\gamma=\gamma^*)\mathrm{KL}\Big\{\prod_{\mathrm{i}=1}^{T}\gamma_{\mathrm{i}}\mathcal{N}(\theta_{\mathrm{i}}^*,\sigma_n^2)+(1-\gamma_{\mathrm{i}})\delta_0\Big|\Big|\prod_{\mathrm{i}=1}^{T}\gamma_{\mathrm{i}}\mathcal{N}(0,\sigma_0^2))+(1-\gamma_{\mathrm{i}})\delta_0\Big\}\\
=&\log\frac{1}{\lambda^{s^*}(1-\lambda)^{T-s^*}}+\sum_{\mathrm{i}=1}^{T}\mathrm{KL}\Big\{\gamma_{\mathrm{i}}^*\mathcal{N}(\theta_{\mathrm{i}}^*,\sigma_n^2)+(1-\gamma_{\mathrm{i}}^*)\delta_0||\gamma_{\mathrm{i}}^*\mathcal{N}(0,\sigma_0^2))+(1-\gamma_{\mathrm{i}}^*)\delta_0\Big\}\\
=&s^*\log(\frac{1}{\lambda})+(T-s^*)\log(\frac{1}{1-\lambda})+\sum_{\mathrm{i}=1}^{T}\gamma_{\mathrm{i}}^*\Big\{\frac{1}{2}\log\Big(\frac{\sigma_0^2}{\sigma_n^2}\Big)+\frac{\sigma_n^2+\theta_{\mathrm{i}}^{*2}}{2}-\frac{1}{2}\Big\}\\
\leqslant&C_0nr_n^*+\frac{s^*}{2}\sigma_n^2+\frac{s^*}{2}(B^2-1)+\frac{s^*}{2}\log\Big(\frac{\sigma_0^2}{\sigma_n^2}\Big)\\
\leqslant&(C_0+1)nr_n^*+\frac{s^*}{2}B^2+\frac{s^*}{2}\log\Big(\frac{8n}{s^*}\log(3pN)(2BN)^{2L+2}\Big\{(p+1+\frac{1}{BN-1})^2\\
&+\frac{1}{(2BN)^2-1}+\frac{2}{(2BN-1)^2}\Big\}\Big)\\
\leqslant&(C_0+2)nr_n^*+\frac{B^2}{2}s^*+(L+1)s^*\log(2BN)+\frac{s^*}{2}\log\log(3BN)+\frac{s^*}{2}\log\Big(\frac{n}{s^*}p^2\Big)\\
\leqslant&(C_0+3)nr_n^*+(L+1)s^*\log N+s^*\log\Big(p\sqrt{\frac{n}{s^*}}\Big)\\
\leqslant&C_1nr_n^*,\ \text{for sufficiently large n,}
\end{aligned}
$$

where C_0 and C_1 are some fixed constants. The first inequality is due to Lemma 4.7.1 and the second inequality is due to Condition 4.4.

Furthermore, by Appendix G of Chérief-Abdellatif 2020, it can be shown

$$
\begin{aligned}
&\int_{\Theta}||f_\theta-f_{\theta*}||_\infty^2q^*(\theta)(d\theta)\\
\leqslant&8a_n^2\log(3BN)(2BN)^{2L+2}\Big\{(p+1+\frac{1}{BN-1})^2+\frac{1}{(2BN)^2-1}+\frac{2}{(2BN-1)^2}\Big\}\\
\leqslant&\frac{s^*}{n}\leqslant r_n^*.
\end{aligned}
$$

Noting that

$$
\begin{aligned}
l_n(P_0, P_\theta) &= \frac{1}{2\sigma_\epsilon^2}(||Y - f_\theta(X)||_2^2 - ||Y - f_0(X)||_2^2) \\
&= \frac{1}{2\sigma_\epsilon^2}(||Y - f_0(X) + f_0(X) - f_\theta(X))||_2^2 - ||Y - f_0(X)||_2^2) \\
&= \frac{1}{2\sigma_\epsilon^2}(||f_\theta(X) - f_0(X)||_2^2 + 2\langle Y - f_0(X), f_0(X) - f_\theta(X)\rangle),
\end{aligned}
$$

Denote

$$
\begin{aligned}
\mathcal{R}_1 &= \int_\Theta ||f_\theta(X) - f_0(X)||_2^2 q^*(\theta)(d\theta), \\
\mathcal{R}_2 &= \int_\Theta \langle Y - f_0(X), f_0(X) - f_\theta(X)\rangle q^*(\theta)(d\theta).
\end{aligned}
$$

Since $||f_\theta(X) - f_0(X)||_2^2 \leqslant n||f_\theta - f_0||_\infty^2 \leqslant n||f_\theta - f_{\theta*}||_\infty^2 + n||f_{\theta*} - f_0||_\infty^2$,

$$
\mathcal{R}_1 \leqslant nr_n^* + n||f_{\theta*} - f_0||_\infty^2.
$$

Noting that $Y - f_0(X) = \epsilon \sim \mathcal{N}(0, \sigma_\epsilon^2 I)$, then

$$
\mathcal{R}_2 = \int_\Theta \epsilon^T (f_0(X) - f_\theta(X)) q^*(\theta)(d\theta) = \epsilon^T \int_\Theta (f_0(X) - f_\theta(X)) q^*(\theta)(d\theta) \sim \mathcal{N}(0, c_f \sigma_\epsilon^2),
$$

where $c_f = ||\int_\Theta (f_0(X) - f_\theta(X)) q^*(\theta)(d\theta)||_2^2 \leqslant \mathcal{R}_1$ due to Cauchy-Schwarz inequality. Therefore, $\mathcal{R}_2 = O_p(\sqrt{\mathcal{R}_1})$, and w.h.p., $\mathcal{R}_2 \leqslant C_0' \mathcal{R}_1$, where C_0' is some positive constant if $\lim n(r_n^* + \xi_n^*) = \infty$ or C_0' is any diverging sequence if $\limsup n(r_n^* + \xi_n^*) \neq \infty$. Therefore,

$$
\begin{aligned}
\int_\Theta l_n(P_0, P_\theta) q^*(\theta)(d\theta) = \mathcal{R}_1/2\sigma_\epsilon^2 + \mathcal{R}_2/\sigma_\epsilon^2 \leqslant &(2C_0' + 1) n(r_n^* + ||f_{\theta*} - f_0||_\infty^2)/2\sigma_\varepsilon^2 \\
\leqslant &C_1'(nr_n^* + ||f_{\theta*} - f_0||_\infty^2)), \text{ w.h.p.,}
\end{aligned}
$$

which concludes this lemma together with (4.8). ■

4.7.2 Proof of Lemma 4.2

Under Condition 4.1 - 4.2, we have the following lemma that shows the existence of testing functions over $\Theta_n = \Theta(L, \boldsymbol{p}, s_n)$, where $\Theta(L, \boldsymbol{p}, s_n)$ denotes the set of parameter whose L_0 norm is bounded by s_n.

Lemma 4.7.2 *Let* $\varepsilon_n^* = Mn^{-1/2}\sqrt{(L+1)s^* \log N + s^* \log(p\sqrt{n/s^*})}\log^{\delta}(n)$ *for any* $\delta > 1$ *and some large constant* M*. Let* $s_n = s^* \log^{2\delta-1} n$*. Then there exists some testing function* $\phi \in [0,1]$ *and* $C_1 > 0$*,* $C_2 > 1/3$*, such that*

$$\mathbb{E}_{P_0}(\phi) \leqslant \exp\{-C_1 n \varepsilon_n^{*2}\},$$

$$\sup_{\substack{P_\theta \in \mathcal{F}(L,\boldsymbol{p},s_n) \\ d(P_\theta,P_0) > \varepsilon_n^*}} \mathbb{E}_{P_\theta}(1-\phi) \leqslant \exp\{-C_2 n d^2(P_0, P_\theta)\}.$$

Proof 14 Due to the well-known result (e.g., Le Cam 1986, page 491 or Ghosal and Van Der Vaart 2007, Lemma 2), there always exists a function $\psi \in [0,1]$, such that

$$\begin{aligned} \mathbb{E}_{P_0}(\psi) &\leqslant \exp\{-nd^2(P_{\theta_1}, P_0)/2\}, \\ \mathbb{E}_{P_\theta}(1-\psi) &\leqslant \exp\{-nd^2(P_{\theta_1}, P_0)/2\}, \end{aligned}$$

for all $P_\theta \in \mathcal{F}(L, \boldsymbol{p}, s_n)$ satisfying that $d(P_\theta, P_{\theta_1}) \leqslant d(P_0, P_{\theta_1})/18$.

Let $K = N(\varepsilon_n^*/19, \mathcal{F}(L, \boldsymbol{p}, s_n), d(\cdot,\cdot))$ denote the covering number of set $\mathcal{F}(L, \boldsymbol{p}, s_n)$, i.e., there exists K Hellinger-balls with radius $\varepsilon_n^*/19$, that completely cover $\mathcal{F}(L, \boldsymbol{p}, s_n)$. For any $\theta \in \mathcal{F}(L, \boldsymbol{p}, s_n)$ (W.O.L.G, we assume P_θ belongs to the kth Hellinger ball centered at P_{θ_k}), if $d(P_\theta, P_0) > \varepsilon_n^*$, then we must have that $d(P_0, P_{\theta_k}) > (18/19)\varepsilon_n^*$ and there exists a testing function ψ_k, such that

$$\begin{aligned} \mathbb{E}_{P_0}(\psi_k) &\leqslant \exp\{-nd^2(P_{\theta_k}, P_0)/2\} \\ &\leqslant \exp\{-(18^2/19^2/2)n\varepsilon_n^{*2}\}, \\ \mathbb{E}_{P_\theta}(1-\psi_k) &\leqslant \exp\{-nd^2(P_{\theta_k}, P_0)/2\} \\ &\leqslant \exp\{-n(d(P_0, P_\theta) - \varepsilon_n^*/19)^2/2\} \\ &\leqslant \exp\{-(18^2/19^2/2)nd^2(P_0, P_\theta)\}. \end{aligned}$$

Now we define $\phi = \max_{k=1,\dots,K} \psi$. Thus we must have

$$\begin{aligned}\mathbb{E}_{P_0}(\phi) &\leqslant \sum_k \mathbb{E}_{P_0}(\psi_k) \leqslant K \exp\{-(18^2/19^2/2)n\varepsilon_n^{*2}\} \\ &\leqslant \exp\{-(18^2/19^2/2)n\varepsilon_n^{*2} - \log K\}.\end{aligned}$$

Note that

$$\begin{aligned}\log K &= \log N(\varepsilon_n^*/19, \mathcal{F}(L, \boldsymbol{p}, s_n), d(\cdot,\cdot)) \\ &\leqslant \log N(\sqrt{8}\sigma_\varepsilon \varepsilon_n^*/19, \mathcal{F}(L, \boldsymbol{p}, s_n), \|\cdot\|_\infty) \\ &\leqslant (s_n+1)\log(\frac{38}{\sqrt{8}\sigma_\varepsilon\varepsilon_n^*}(L+1)(N+1)^{2(L+1)}) \\ &\leqslant C_0(s_n \log\frac{1}{\varepsilon_n^*} + s_n\log(L+1) + s_n(L+1)\log N) \\ &\leqslant s_n(L+1)\log n \log N \leqslant s^*(L+1)\log N \log^{2\delta} n \\ &\leqslant n\varepsilon_n^{*2}/4, \text{ for sufficiently large n,} \end{aligned} \tag{4.11}$$

where C_0 is some positive constant, the first inequality is due to the fact

$$d^2(P_\theta, P_0) \leqslant 1 - \exp\{-\frac{1}{8\sigma_\epsilon^2}||f_0 - f_\theta||_\infty^2\}$$

and $\varepsilon_n^* = o(1)$, the second inequality is due to Lemma 10 of Schmidt-Hieber 2017[1], and the last inequality is due to $s_n \log(1/\varepsilon_n^*) \asymp s_n \log n$. Therefore,

$$\mathbb{E}_{P_0}(\phi) \leqslant \sum_k P_0(\psi_k) \leqslant \exp\{-C_1 n\varepsilon_n^{*2}\},$$

for some $C_1 = 18^2/19^2/2 - 1/4$. On the other hand, for any θ, such that $d(P_\theta, P_0) \geqslant \varepsilon_n^*$, say P_θ belongs to the kth Hellinger ball, then we have

$$\mathbb{E}_{P_\theta}(1-\phi) \leqslant \mathbb{E}_{P_\theta}(1-\psi_k) \leqslant \exp\{-C_2 n d^2(P_0, P_\theta)\},$$

---

[1] Although Schmidt-Hieber 2017 only focuses on ReLU network, its Lemma 10 could apply to any 1-Lipchitz continuous activation function.

where $C_2 = 18^2/19^2/2$. Hence we conclude the proof. ■

Lemma 4.7.3 restates the Donsker and Varadhan's representation for the KL divergence, whose proof can be found in Boucheron et al. 2013.

Lemma 4.7.3 *For any probability measure μ and any measurable function h with $\mathrm{e}^h \in L_1(\mu)$,*

$$\log \int \mathrm{e}^{h(\eta)}\mu(d\eta) = \sup_{\rho}\left[\int h(\eta)\rho(d\eta) - KL(\rho||\mu)\right].$$

We are now ready to prove Lemma 4.2

Proof 15 Denote Θ_n as the truncated parameter space $\{\theta : \sum_{\mathrm{i}=1}^{T} 1(\theta_{\mathrm{i}} \neq 0) \leqslant s_n\}$, where s_n is defined in Lemma 4.7.2. Noting that

$$\int_{\theta\in\Theta} d^2(P_\theta, P_0)\widehat{q}(\theta)d\theta = \int_{\theta\in\Theta_n} d^2(P_\theta, P_0)\widehat{q}(\theta)d\theta + \int_{\theta\in\Theta_n^c} d^2(P_\theta, P_0)\widehat{q}(\theta)d\theta, \tag{4.12}$$

it suffices to find upper bounds of the two components in RHS of (4.12).

We start with the first component. Denote $\widetilde{\boldsymbol{\pi}}(\theta)$ to be the truncated prior $\boldsymbol{\pi}(\theta)$ on set Θ_n, i.e., $\widetilde{\boldsymbol{\pi}}(\theta) = \boldsymbol{\pi}(\theta)1(\theta \in \Theta_n)/\boldsymbol{\pi}(\Theta_n)$, then by Lemma 4.7.2 and the same argument used in Theorem 3.1 of Pati et al. 2018, it could be shown

$$\int_{\Theta_n} \eta(P_\theta, P_0)\widetilde{\boldsymbol{\pi}}(\theta)d\theta \leqslant \mathrm{e}^{C_0 n\varepsilon_n^{*2}}, \text{w.h.p.} \tag{4.13}$$

for some $C_0 > 0$, where $\log \eta(P_\theta, P_0) = l_n(P_\theta, P_0) + \frac{n}{3}d^2(P_\theta, P_0)$. We further denote the $\widehat{q}(\theta)$ restricted on Θ_n as $\breve{q}(\theta)$, i.e., $\breve{q}(\theta) = \widehat{q}(\theta)1(\theta \in \Theta_n)/\widehat{q}(\Theta_n)$, then by Lemma 4.7.3 and (4.13), w.h.p.,

$$\begin{aligned}\frac{n}{3\widehat{q}(\Theta_n)}\int_{\Theta_n} d^2(P_\theta, P_0)\widehat{q}(\theta)d\theta &= \frac{n}{3}\int_{\Theta_n} d^2(P_\theta, P_0)\breve{q}(\theta)d\theta \\ \leqslant & Cn\varepsilon_n^{*2} + \mathrm{KL}(\breve{q}(\theta)||\widetilde{\boldsymbol{\pi}}(\theta)) - \int_{\Theta_n} l_n(P_\theta, P_0)\breve{q}(\theta)d\theta.\end{aligned} \tag{4.14}$$

Furthermore,

$$\begin{aligned}\mathrm{KL}(\breve{q}(\theta)||\widetilde{\pi}(\theta)) &= \frac{1}{\widehat{q}(\Theta_n)}\int_{\theta\in\Theta_n}\log\frac{\widehat{q}(\theta)}{\pi(\theta)}\widehat{q}(\theta)d\theta+\log\frac{\pi(\Theta_n)}{\widehat{q}(\Theta_n)}\\ &= \frac{1}{\widehat{q}(\Theta_n)}\mathrm{KL}(\widehat{q}(\theta)||\pi(\theta))-\frac{1}{\widehat{q}(\Theta_n)}\int_{\theta\in\Theta_n^c}\log\frac{\widehat{q}(\theta)}{\pi(\theta)}\widehat{q}(\theta)d\theta+\log\frac{\pi(\Theta_n)}{\widehat{q}(\Theta_n)},\end{aligned}$$

and similarly,

$$\int_{\Theta_n} l_n(P_\theta,P_0)\breve{q}(\theta)d\theta=\frac{1}{\widehat{q}(\Theta_n)}\int_{\Theta} l_n(P_\theta,P_0)\widehat{q}(\theta)d\theta-\frac{1}{\widehat{q}(\Theta_n)}\int_{\Theta_n^c} l_n(P_\theta,P_0)\widehat{q}(\theta)d\theta.$$

Combining the above two equations together, we have

$$\begin{aligned}&\frac{n}{3\widehat{q}(\Theta_n)}\int_{\Theta_n} d^2(P_\theta,P_0)\widehat{q}(\theta)d\theta\leqslant Cn\varepsilon_n^{*2}+\mathrm{KL}(\breve{q}(\theta)||\widetilde{\pi}(\theta))-\int_{\Theta_n} l_n(P_\theta,P_0)\breve{q}(\theta)d\theta\\ =&Cn\varepsilon_n^{*2}+\frac{1}{\widehat{q}(\Theta_n)}\left(\mathrm{KL}(\widehat{q}(\theta)||\pi(\theta))-\int_{\Theta} l_n(P_\theta,P_0)\widehat{q}(\theta)d\theta\right)\\ &-\frac{1}{\widehat{q}(\Theta_n)}\left(\int_{\Theta_n^c}\log\frac{\widehat{q}(\theta)}{\pi(\theta)}\widehat{q}(\theta)d\theta-\int_{\Theta_n^c} l_n(P_\theta,P_0)\widehat{q}(\theta)d\theta\right)+\log\frac{\pi(\Theta_n)}{\widehat{q}(\Theta_n)}.\end{aligned}\tag{4.15}$$

The second component of (4.12) trivially satisfies that $\int_{\theta\in\Theta_n^c} d^2(P_\theta,P_0)\widehat{q}(\theta)d\theta\leqslant\int_{\theta\in\Theta_n^c}\widehat{q}(\theta)d\theta$ $=\widehat{q}(\Theta_n^c)$. Thus, together with (4.15), we have that w.h.p.,

$$\begin{aligned}&\int d^2(P_\theta,P_0)\widehat{q}(\theta)d\theta\leqslant 3\widehat{q}(\Theta_n)C\varepsilon_n^{*2}+\frac{3}{n}\left(\mathrm{KL}(\widehat{q}(\theta)||\pi(\theta))-\int_{\Theta} l_n(P_\theta,P_0)\widehat{q}(\theta)d\theta\right)\\ &+\frac{3}{n}\int_{\Theta_n^c} l_n(P_\theta,P_0)\widehat{q}(\theta)d\theta+\frac{3}{n}\int_{\Theta_n^c}\log\frac{\pi(\theta)}{\widehat{q}(\theta)}\widehat{q}(\theta)d\theta+\frac{3\widehat{q}(\Theta_n)}{n}\log\frac{\pi(\Theta_n)}{\widehat{q}(\Theta_n)}+\widehat{q}(\Theta_n^c).\end{aligned}\tag{4.16}$$

The second term in the RHS of (4.16) is bounded by $C'(r_n^*+\xi_n^*)$ w.h.p., due to Lemma 4.1, where C' is either positive constant or diverging sequence depending on whether $n(r_n^*+\xi_n^*)$ diverges.

The third term in the RHS of (4.16) is bounded by

$$\begin{aligned}
&\frac{3}{n}\int_{\Theta_n^c} l_n(P_\theta, P_0)\widehat{q}(\theta)d\theta \\
=&\frac{3}{2n\sigma_\epsilon^2}\int_{\Theta_n^c}\left[\sum_{i=1}^n \epsilon_i^2 - \sum_{i=1}^n(\epsilon_i + f_0(X_i) - f_\theta(X_i))^2\right]\widehat{q}(\theta)d\theta \\
=&\frac{3}{2n\sigma_\epsilon^2}\int_{\Theta_n^c}\left[-2\sum_{i=1}^n(\epsilon_i \times (f_0(X_i) - f_\theta(X_i)) - \sum_{i=1}^n(f_0(X_i) - f_\theta(X_i))^2\right]\widehat{q}(\theta)d\theta \\
=&\frac{3}{2n\sigma_\epsilon^2}\left\{-2\sum_{i=1}^n \epsilon_i \int_{\Theta_n^c}(f_0(X_i) - f_\theta(X_i))\widehat{q}(\theta)d\theta - \int_{\Theta_n^c}\sum_{i=1}^n(f_0(X_i) - f_\theta(X_i))^2\widehat{q}(\theta)d\theta\right\}.
\end{aligned}$$

Conditional on X_i's, $-2\sum_{i=1}^n \epsilon_i \int_{\Theta_n^c}(f_0(X_i) - f_\theta(X_i))\widehat{q}(\theta)d\theta$ follows a normal distribution $\mathcal{N}(0, V^2)$, where $V^2 = 4\sigma_\epsilon^2\sum_{i=1}^n(\int_{\Theta_n^c}(f_0(X_i)-f_\theta(X_i))\widehat{q}(\theta)d\theta)^2 \leqslant 4\sigma_\epsilon^2\int_{\Theta_n^c}\sum_{i=1}^n(f_0(X_i)-f_\theta(X_i))^2$ $\widehat{q}(\theta)d\theta$. Thus conditional on X_i's, the third term in the RHS of (4.16) is bounded by

$$\frac{3}{2n\sigma_\epsilon^2}\left[\mathcal{N}(0, V^2) - \frac{V^2}{4\sigma_\epsilon^2}\right]. \tag{4.17}$$

Noting that $\mathcal{N}(0, V^2) = O_p(M_n V)$ for any diverging sequence M_n, (4.17) is further bounded, w.h.p., by

$$\frac{3}{2n\sigma_\epsilon^2}(M_n V - \frac{V^2}{4\sigma_\epsilon^2}) \leqslant \frac{3}{2n\sigma_\epsilon^2}\sigma_\epsilon^2 M_n^2.$$

Therefore, the third term in the RHS of (4.16) can be bounded by ε_n^{*2} w.h.p. (by choosing $M_n^2 = n\varepsilon_n^{*2}$).

The fourth term in the RHS of (4.16) is bounded by

$$\frac{3}{n}\int_{\Theta_n^c}\log\frac{\pi(\theta)}{\widehat{q}(\theta)}\widehat{q}(\theta)d\theta \leqslant \frac{3}{n}\widehat{q}(\Theta_n^c)\log\frac{\pi(\Theta_n^c)}{\widehat{q}(\Theta_n^c)} \leqslant \frac{3}{n}\sup_{x\in(0,1)}[x\log(1/x)] = O(1/n).$$

Similarly, the fifth term in the RHS of (4.16) is bounded by $O(1/n)$.

For the last term in the RHS of (4.16), by Lemma 4.7.5 in below, w.h.p., $\widehat{q}(\Theta_n^c) \leqslant \varepsilon_n^{*2}$.

Combine all the above result together, w.h.p.,

$$\int d^2(P_\theta, P_0)\widehat{q}(\theta)d\theta \leqslant C\varepsilon_n^{*2} + \frac{3}{n}\left(\mathrm{KL}(\widehat{q}(\theta)||\pi(\theta)) - \int_\Theta l_n(P_\theta, P_0)\widehat{q}(\theta)d\theta\right) + O(1/n),$$

where C is some constant. ■

Lemma 4.7.4 (Chernoff bound for Poisson tail) *Let $X \sim poi(\lambda)$ be a Poisson random variable. For any $x > \lambda$,*

$$P(X \geqslant x) \leqslant \frac{(\mathrm{e}\lambda)^x \mathrm{e}^{-\lambda}}{x^x}.$$

Lemma 4.7.5 *If $\lambda \leqslant T^{-1}\exp\{-Mnr_n^*/s_n\}$ for any positive diverging sequence $M \to \infty$, then w.h.p., $\widehat{q}(\Theta_n^c) = O(\varepsilon_n^{*2})$.*

Proof 16 By Lemma 4.1, we have that w.h.p.,

$$\mathrm{KL}(\widehat{q}(\theta)||\boldsymbol{\pi}(\theta|\lambda)) + \int_\Theta l_n(P_0, P_\theta)\widehat{q}(\theta)d\theta = \inf_{q_\theta \in \mathcal{Q}}\Big\{\mathrm{KL}(q(\theta)||\boldsymbol{\pi}(\theta|\lambda)) + \int_\Theta l_n(P_0, P_\theta)q(\theta)(d\theta)\Big\}$$
$$\leqslant Cnr_n^* \quad (\text{Note that } r_n^* \asymp \xi_n^*)$$

where C is either a constant or any diverging sequence, depending on whether nr_n^* diverges. By the similar argument used in the proof of Lemma 4.1,

$$\int_\Theta l_n(P_0, P_\theta)\widehat{q}(\theta)d\theta \leqslant \frac{1}{2\sigma_\epsilon^2}\left(\int_\Theta ||f_\theta(X) - f_0(X)||_2^2\widehat{q}(\theta)(d\theta) + Z\right)$$

where Z is a normal distributed $\mathcal{N}(0, \sigma_\epsilon^2 c_0)$, where $c_0 \leqslant c_0 = \int_\Theta ||f_\theta(X) - f_0(X)||_2^2\widehat{q}(\theta)(d\theta)$. Therefore, $-\int_\Theta l_n(P_0, P_\theta)\widehat{q}(\theta)d\theta = (1/2\sigma_\epsilon^2)[-c_0 + O_p(\sqrt{c_0})]$, and $\mathrm{KL}(\widehat{q}(\theta)||\boldsymbol{\pi}(\theta|\lambda)) \leqslant Cnr_n^* + (1/2\sigma_\epsilon^2)[-c_0 + O_p(\sqrt{c_0})]$. Since $Cnr_n^* \to \infty$, we must have w.h.p., $\mathrm{KL}(\widehat{q}(\theta)||\boldsymbol{\pi}(\theta|\lambda)) \leqslant Cnr_n^*/2$. On the other hand,

$$\begin{aligned}\mathrm{KL}(\widehat{q}(\theta)||\boldsymbol{\pi}(\theta|\lambda)) &= \sum_{\mathrm{i}=1}^{T}\mathrm{KL}(\widehat{q}(\theta_\mathrm{i})||\boldsymbol{\pi}(\theta_\mathrm{i}|\lambda)) \geqslant \sum_{\mathrm{i}=1}^{T}\mathrm{KL}(\widehat{q}(\gamma_\mathrm{i})||\boldsymbol{\pi}(\gamma_\mathrm{i}|\lambda)) \\ &= \sum_{\mathrm{i}=1}^{T}\left[\widehat{q}(\gamma_\mathrm{i} = 1)\log\frac{\widehat{q}(\gamma_\mathrm{i} = 1)}{\lambda} + \widehat{q}(\gamma_\mathrm{i} = 0)\log\frac{\widehat{q}(\gamma_\mathrm{i} = 0)}{1 - \lambda}\right].\end{aligned} \tag{4.18}$$

Let us choose $\lambda_0 = 1/T$, and $A = \{\mathrm{i} : \widehat{q}(\gamma_\mathrm{i} = 1) \geqslant \lambda_0\}$, then the above inequality (4.18) implies that $\sum_{\mathrm{i}\in A}\widehat{q}(\gamma_\mathrm{i} = 1)\log(\lambda_0/\lambda) \leqslant Cnr_n^*/2$. Noting that $\lambda \leqslant T^{-1}\exp\{-Mnr_n^*/s_n\}$, it further implies $\sum_{\mathrm{i}\in A}\widehat{q}(\gamma_\mathrm{i} = 1) \leqslant s_n/M \prec s_n$.

Under distribution $\widehat{q}$, by Bernstein inequality,

$$Pr(\sum_{i\in A}\gamma_i \geqslant 2s_n/3) \leqslant Pr(\sum_{i\in A}\gamma_i \geqslant s_n/2 + \sum_{i\in A}\mathbb{E}(\gamma_i)) \leqslant \exp\left(-\frac{s_n^2/8}{\sum_{i\in A}\mathbb{E}[\gamma_i^2]+s_n/6}\right)$$
$$= \exp\left(-\frac{s_n^2/8}{\sum_{i\in A}\widehat{q}(\gamma_i=1)+s_n/6}\right) \leqslant \exp(-cs_n) = O(\varepsilon_n^{*2})$$

for some constant $c>0$, where the last inequality holds since $\log(1/\varepsilon_n^{*2}) = O(\log n) \prec s_n$.

Under distribution $\widehat{q}$, $\sum_{i\notin A}\gamma_i$ is stochastically smaller than $Bin(T,\lambda_0)$. Since $T\to\infty$, then by Lemma 4.7.4,

$$Pr(\sum_{i\notin A}\gamma_i \geqslant s_n/3) \leqslant Pr(Bin(T,\lambda_0)\geqslant s_n/3) \to Pr(\text{poi}(1)\geqslant s_n/3)$$
$$=O(\exp\{-C's_n\}) = O(\varepsilon_n^{*2})$$

for some $C'>0$. Trivially, it implies that w.h.p, $Pr(\sum_i \gamma_i \geqslant s_n) = O(\varepsilon_n^{*2})$ for VB posterior $\widehat{q}$.

■

4.7.3 Main theorem

Theorem 4.7.1 *Under Conditions 4.1-4.2, 4.4 and set* $-\log\lambda = \log(T)+\delta[(L+1)\log N + \log\sqrt{np}]$ *for any constant* $\delta>0$*, we then have that w.h.p.,*

$$\int_\Theta d^2(P_\theta,P_0)\widehat{q}(\theta)d\theta \leqslant C\varepsilon_n^{*2} + C'(r_n^*+\xi_n^*),$$

where C *is some positive constant and* C' *is any diverging sequence. If* $\|f_0\|_\infty < F$*, and we truncated the VB posterior on* $\Theta_F = \{\theta : \|f_\theta\|_\infty \leqslant F\}$*, i.e.,* $\widehat{q}_F \propto \widehat{q}1(\theta\in\Theta_F)$*, then, w.h.p.,*

$$\int_{\Theta_F}\mathbb{E}_X|f_\theta(X)-f_0(X)|^2\widehat{q}_F(\theta)d\theta \leqslant \frac{C\varepsilon_n^{*2}+C'(r_n^*+\xi_n^*)}{C_F\widehat{q}(\Theta_F)}$$

where $C_F = [1-\exp(-4F^2/8\sigma_\epsilon^2)]/4F^2$*, and* $\widehat{q}(\Theta_F)$ *is the VB posterior mass of* Θ_F*.*

Proof 17 The convergence under squared Hellinger distance is directly result of Lemma 4.1 and 4.2, by simply checking the choice of λ satisfies required conditions. The convergence under L_2 distance relies on inequality $d^2(P_\theta, P_0) \geqslant C_F \mathbb{E}_X|f_\theta(X) - f_0(X)|^2$ for $C_F = [1 - \exp(-4F^2/8\sigma_\epsilon^2)]/4F^2$ when both f_θ and f_0 are bounded by F. Then, w.h.p,

$$\int_{\Theta_F} \mathbb{E}_X|f_\theta(X) - f_0(X)|^2 \widehat{q}_F(\theta)d\theta \leqslant C_F^{-1} \int_{\Theta_F} d^2(P_\theta, P_0)\widehat{q}_F(\theta)d\theta$$

$$\leqslant \frac{1}{C_F \widehat{q}(\Theta_F)} \int_\Theta d^2(P_\theta, P_0)\widehat{q}(\theta)d\theta \leqslant \frac{C\varepsilon_n^{*2} + C'(r_n^* + \xi_n^*)}{C_F \widehat{q}(\Theta_F)}.$$

■

4.8 Additional experimental results

4.8.1 Algorithm implementation details for the numerical experiments

Initialization As mentioned by Sønderby et al. 2016 and Molchanov et al. 2017, training sparse BNN with random initialization may lead to bad performance, since many of the weights could be pruned too early. In our case, we assign each of the weights and biases a inclusion variable, which could reduce to zero quickly in the early optimization stage if we randomly initialize them. As a consequence, we deliberately initialize ϕ_i to be close to 1 in our experiments. This initialization strategy ensures the training starts from a fully connected neural network, which is similar to start training from a pre-trained fully connected network as mentioned in ibid. The other two parameters μ_i and σ_i are initialized randomly.

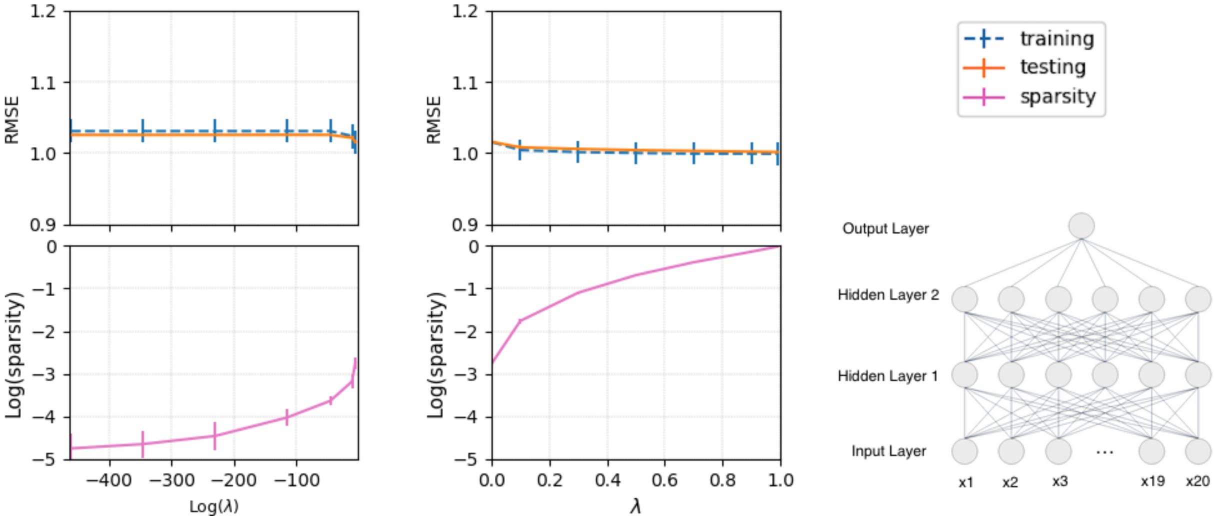
Other implementation details in simulation studies We set $K = 1$ and learning rate $= 5 \times 10^{-3}$ during training. For Simulation I, we choose batch size $m = 1024$ and $m = 128$ for (A) and (B) respectively, and run 10000 epochs for both cases. For simulation II, we use $m = 512$ and run 7000 epochs. Although it is common to set up an annealing schedule for temperature parameter τ, we don't observe any significant performance improvement compared to setting τ as a constant, therefore we choose $\tau = 0.5$ in all of our experiments. The optimization method used is Adam.

The implementation details for UCI datasets and MNIST can be found in Section 4.8.3 and 4.8.4 respectively.

4.8.2 Teacher student networks

The network parameter θ for the sparse teacher network setting (B) is set as following: $W = \{W_{1,11} = W_{1,12} = W_{2,11} = W_{2,12} = 2.5, W_{1,21} = W_{1,22} = W_{2,21} = W_{2,22} = 1.5, W_{3,11} = 3$ and $W_{3,21} = 2\}$; $b = \{b_{1,1} = b_{2,1} = b_{3,1} = 1$ and $b_{1,2} = b_{2,2} = -1\}$.

Figure 4.4 displays the simulation result for simulation I under dense teacher network (A) setting. Unlike the result under sparse teacher network (B), the testing accuracy seems monotonically increases as λ increases (i.e., posterior network gets denser). However, as shown, the increasing of testing performance is rather slow, which indicates that introducing sparsity has few negative impact to the testing accuracy.



(a) $\lambda \leqslant \lambda_{opt}$. (b) $\lambda \geqslant \lambda_{opt}$. (c) Dense teacher network.

Figure 4.4. (a) $\lambda = \{10^{-200}, 10^{-150}, 10^{-100}, 10^{-50}, 10^{-20}, 10^{-5}, \lambda_{opt}\}$. (b) $\lambda = \{\lambda_{opt}, 0.1, 0.3, 0.5, 0.7, 0.9, 0.99\}$. (c) The structure of the target dense teacher network.

Coverage rate In this paragraph, we explain the details of how we compute the coverage rate values of Bayesian intervals reported in the main text. A fixed point $(x_1^{(*)}, \ldots, x_p^{(*)})$ is prespecified, and let $x^{(1)}, \ldots, x^{(1000)}$ be 1000 equidistant points from -1 to 1. In each run, we compute the Bayesian credible intervals of response means (estimated by 600 Monte Carlo samples) for 1000 different input x's: $(x^{(1)}, x_2^{(*)}, \ldots, x_p^{(*)}), \ldots, (x^{(1000)}, x_2^{(*)}, \ldots, x_p^{(*)})$. It is repeated by 60 times and the average coverage rate (over all different x's and 60 runs) is reported. Similarly, we replace $x_2^{(*)}$ (or $x_3^{(*)}$) by $x^{(\mathrm{i})}$ (i $= 1, \ldots, 1000$), and compute their average coverage rate. The complete coverage rate results are shown in Table 4.3. Note that Table 1 in the main text shows 95% coverage of x_3 for (A) and 95% coverage of x_1 for (B).

Table 4.3. Coverage rates for teacher networks.

		90 % coverage (%)			**95% coverage (%)**		
	Method	$\boldsymbol{x_1}$	$\boldsymbol{x_2}$	$\boldsymbol{x_3}$	$\boldsymbol{x_1}$	$\boldsymbol{x_2}$	$\boldsymbol{x_3}$
Dense	SVBNN	93.8 ± 2.84	93.1 ± 4.93	93.1 ± 2.96	97.9 ± 1.01	97.9 ± 1.69	97.5 ± 1.71
	VBNN	85.8 ± 2.51	82.4 ± 2.62	86.3 ± 1.88	92.7 ± 2.83	91.3 ± 2.61	91.4 ± 2.43
	VD	61.3 ± 2.40	60.0 ± 2.79	64.9 ± 6.17	74.9 ± 1.79	71.8 ± 2.33	76.4 ± 4.75
	HS-BNN	83.1 ± 1.67	80.0 ± 1.21	76.9 ± 1.70	88.1 ± 1.13	84.1 ± 1.48	83.5 ± 0.78
Sparse	SVBNN	92.3 ± 8.61	94.6 ± 5.37	98.3 ± 0.00	96.4 ± 4.73	97.7 ± 3.71	100 ± 0.00
	VBNN	86.7 ± 10.9	87.0 ± 11.3	93.3 ± 0.00	90.7 ± 8.15	91.9 ± 9.21	96.7 ± 0.00
	VD	65.2 ± 0.08	63.7 ± 6.58	65.9 ± 0.83	75.5 ± 7.81	74.6 ± 7.79	76.6 ± 0.40
	HS-BNN	59.0 ± 8.52	59.4 ± 4.38	56.6 ± 2.06	67.0 ± 8.54	68.2 ± 3.62	66.5 ± 1.86

4.8.3 Real data regression experiment: UCI datasets

We follow the experimental protocols of Hernández-Lobato et al. 2015, and choose five datasets for the experiment. For the small datasets ”Kin8nm”, ”Naval”, ”Power Plant” and ”wine”, we choose a single-hidden-layer ReLU network with 50 hidden units. We randomly select 90% and 10% for training and testing respectively, and this random split process is repeated for 20 times (to obtain standard deviations for our results). We choose minibatch size $m = 128$, learning rate $= 10^{-3}$ and run 500 epochs for ”Naval”, ”Power Plant” and ”Wine”, 800 epochs for ”Kin8nm”. For the large dataset ”Year”, we use a single-hidden-layer ReLU network with 100 hidden units, and the evaluation is conducted on a single split. We choose $m = 256$, learning rate $= 10^{-3}$ and run 100 epochs. For all the five datasets, λ is chosen as λ_{opt}: $\log(\lambda_{opt}^{-1}) = \log(T) + 0.1[(L+1)\log N + \log\sqrt{np}]$, which is the same as other numerical studies. We let $\sigma_0^2 = 2$ and use grid search to find σ_ϵ that yields the best prediction accuracy. Adam is used for all the datasets in the experiment.

We report the testing squared root MSE (RMSE) based on $\widehat{f}_H$ (defined in the main text) with $H = 30$, and also report the posterior network sparsity $\widehat{s} = \sum_{\mathrm{i}=1}^{T} \phi_{\mathrm{i}}/T$. For the purpose of comparison, we list the results by Horseshoe BNN (HS-BNN) (Ghosh and Doshi-Velez 2017) and probalistic backpropagation (PBP) (Hernández-Lobato et al. 2015). Table 4.4

demonstrates that our method achieves best prediction accuracy for all the datasets with a sparse structure.

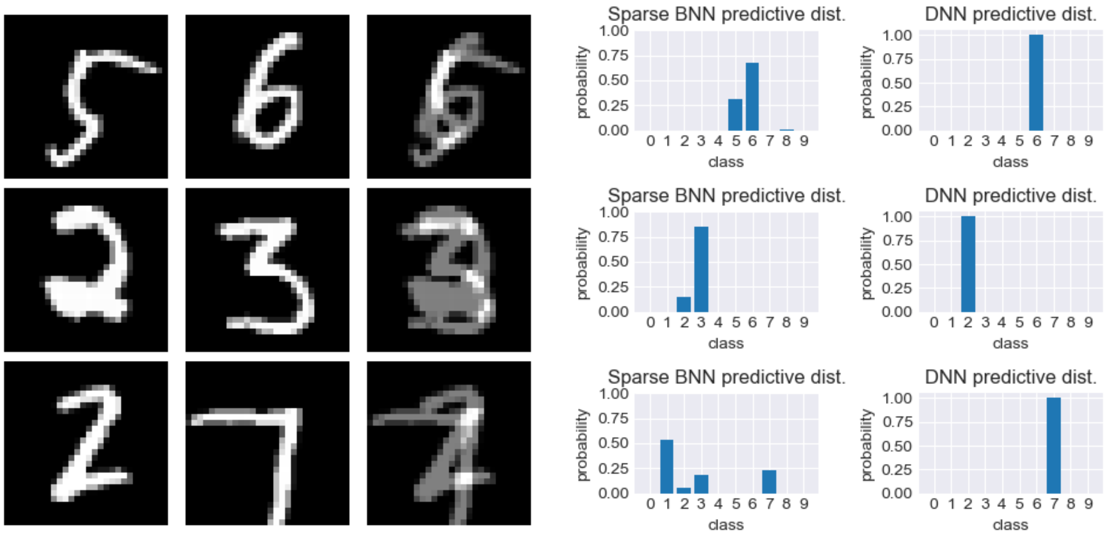
Table 4.4. Results on UCI regression datasets.

		Test RMSE			Posterior sparsity(%)
Dataset	$n(p)$	SVBNN	HS-BNN	PBP	SVBNN
Kin8nm	8192 (8)	0.08 ± 0.00	0.08 ± 0.00	0.10 ± 0.00	64.5 ± 1.85
Naval	11934 (16)	0.00 ± 0.00	0.00 ± 0.00	0.01 ± 0.00	82.9 ± 1.31
Power Plant	9568 (4)	4.01 ± 0.18	4.03 ± 0.15	4.12 ± 0.03	56.6 ± 3.13
Wine	1599 (11)	0.62 ± 0.04	0.63 ± 0.04	0.64 ± 0.01	59.9 ± 4.92
Year	515345 (90)	8.87 ± NA	9.26 ± NA	8.88 ± NA	20.8 ± NA

4.8.4 Real data classification experiment: MNIST dataset

The MNIST data is normalized by mean equaling 0.1306 and standard deviation equaling 0.3081. For all methods, we choose the same minibatch size $m = 256$, learning rate $= 5 \times 10^{-3}$ for our method and 3×10^{-3} for the others, total number of epochs is 400 and the optimization algorithm is RMSprop. AGP is pre-specified at 5% sparsity level.

Besides the testing accuracy reported in the main text, we also examine our method's ability of uncertainty quantification for MNIST classification task. We first create ambiguous images by overlaying two examples from the testing set as shown in Figure 4.5 (a). To perform uncertainty quantification using our method, for each of the overlaid images, we generate θ_h from the VB posterior $\widehat{q}(\theta)$ for $h = 1, \dots, 100$, and calculate the associated predictive probability vector $f_{\theta_h}(x) \in \mathbb{R}^{10}$ where x is the overlaid image input, and then use the estimated posterior mean $\widehat{f}(x) = \sum_{h=1}^{100} f_{\theta_h}(x)/100$ as the Bayesian predictive probability vector. As a comparison, we also calculate the predictive probability vector for each overlaid image using AGP as a frequentist benchmark. Figure 4.5 (b) shows frequentist method gives almost a deterministic answer (i.e., predictive probability is almost 1 for certain digit) that is obviously unsatisfactory for this task, while our VB method is capable of providing knowledge of certainty on these out-of-domain inputs, which demonstrates the advantage of Bayesian method in uncertainty quantification on the classification task.



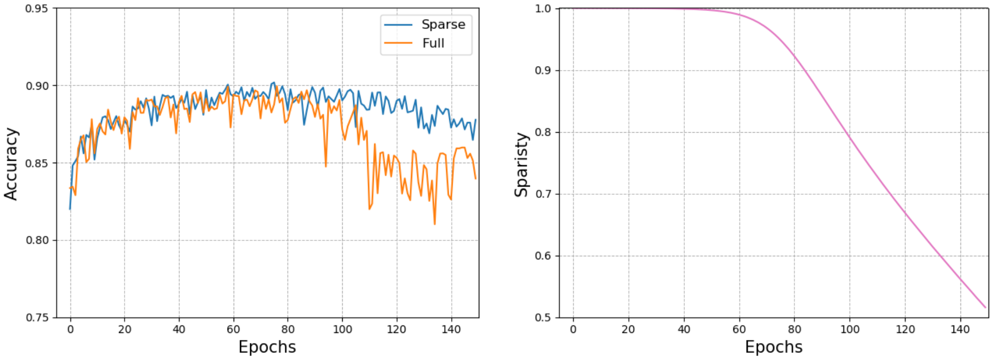
(a) Overlaid images (on the last column) (b) Predictive distribution for overlaid images

Figure 4.5. Top row of (b) exhibits the predictive distribution for the top overlaid image, which is made by 5 and 6; Middle row of (b) exhibits the predictive distribution for the middle overlaid image, which is made by 2 and 3; Bottom row of (b) exhibits the predictive distribution for the bottom overlaid image, which is made by 2 and 7.

4.8.5 Illustration of CNN: Fashion-MNIST dataset

In this section, we perform an experiment on a more complex task, the Fashion-MNIST dataset. To illustrate the usage of our method beyond feedforward networks, we consider using a 2-Conv-2-FC network: The feature maps for the convolutional layers are set to be 32 and 64, and the filter size are 5×5 and 3×3 respectively. The paddings are 2 for both layers and the it has a 2×2 max pooling for each of the layers; The fully-connected layers have $64 \times 8 \times 8$ neurons. The activation functions are all ReLUs. The dataset is prepocessed by random horizontal flip. The batchsize is 1024, learning rate is 0.001, and Adam is used for optimization. We run the experiment for 150 epochs.

We use both SVBNN and VBNN for this task. In particular, the VBNN, which uses normal prior and variational distributions, is the full Bayesian method without compressing, and can be regarded as the baseline for our method. Figure 4.6 exhibits our method attains



(a) Accuracy. (b) Sparsity.

Figure 4.6. Fashion-MNIST experiment.

higher accuracy as epoch increases and then decreases as the sparisty goes down. Meanwhile, the baseline method - full BNN suffers from overfitting after 80 epochs.

5. SUMMARY

We apply variance inference as a computational efficient alternative to high dimensional linear regression and sparse deep learning, respectively. For both problems, we are able to provide the theoretical guarantees as well as efficient algorithms.

Possible future directions are (1) to explore efficient implementation other than black-box variational inference for other heavy tail shrinkage priors besides the Student-t under the high dimensional regression setting, and attempt to provide theoretical guarantee for variable selection; (2) to extend the current results on Bayesian sparse deep learning to more complicated networks (convolutional neural network, residual network, etc.) both theoretically and computationally as mentioned in Section 4.6. Furthermore, proposing a theoretical framework regarding uncertainty quantification for sparse Bayesian neural network is also a challenging but promising topic.

REFERENCES

Alquier, P. and Ridgway, J. (2017). "Concentration of tempered posteriors and of their variational approximations". In: *arXiv:1706.09293*.

Armagan, A., Dunson, D. B., Lee, J., Bajwa, W. U., and Strawn, N. (2013). "Posterior consistency in linear models under shrinkage priors". In: *Biometrika* 100, pp. 1011–1018.

Bai, J., Song, Q., and Cheng, G. (2019). "Adaptive Variational Bayesian Inference for Sparse Deep Neural Network". In: *arXiv preprint arXiv:1910.04355*.

Bai, J., Song, Q., and Cheng, G. (2020a). "Efficient variational inference for sparse deep learning with theoretical guarantee". In: *34th Conference on Neural Information Processing Systems (NeurIPS 2020)*. Vancouver, Canada.

Bai, J., Song, Q., and Cheng, G. (2020b). "Nearly optimal variational inference for high dimensional regression with shrinkage priors". In: *arXiv preprint arXiv:2010.12887*.

Bauler, B. and Kohler, M. (2019). "On deep learning as a remedy for the curse of dimensionality in nonparametric regression". In: *The Annals of Statistics* 47.4, pp. 2261–2285.

Bhadra, A., Datta, J., Polson, N. G., and Willard, B. T. (2019). "The horseshoe-like regularization for feature subset selection". In: *Sankhya B*.

Bickel, P., Choi, D., Chang, X., and Zhang, H. (2013). "Asymptotic normality of maximum likelihood and its variational approximation for stochastic blockmodels". In: *The Annals of Statistics* 41.4, pp. 1922–1943.

Bishop, C. (2006). *Pattern Recognition and Machine Learning*. New York: Springer.

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

Blei, D. M., Jordan, M. I., and Paisley, J. (2012). "Variational Bayesian inference with stochastic search". In: *Proceedings of the 29th International Conference on Machine Learning (ICML 2012)*, pp. 1367–1374.

Blundell, C., Cornebise, J., and Kavukcuoglu, K. (2015). "Weight uncertainty in neural networks". In: *Proceedings of the 32nd International Conference on International Conference on Machine Learning (ICML 15)*. Lille, France, pp. 1613–1622.

Bölcskei, H., Grohs, P., Kutyniok, G., and Petersen, P. (2019). "Optimal approximation with sparsely connected deep neural networks". In: *CoRR* abs/1705.01714.

Boucheron, S., Lugosi, G., and Massart, P. (2013). *Concentration inequalities: A nonasymptotic theory of independence.* Oxford University press.

Carbonetto, P. and Stephens, M. (2012). “Scalable variational inference for Bayesian variable selection in regression, and its accuracy in genetic association studies”. In: *Bayesian Analysis* 7 (1), pp. 73–107.

Carvalho, C. M., Polson, N. G., and Scott, J. G. (2010). “The horseshoe estimator for sparse signals”. In: *Biometrika* 97, pp. 465–480.

Carvalho, C. M., Polson, N. G., and Scott, J. G. (2009). “Handling sparsity via the horseshoe”. In: *Artificial Intelligence and Statistics*, pp. 73–80.

Castillo, I., Schmidt-Hieber, J., and Van der Vaart, A. W. (2015). “Bayesian linear regression with sparse priors”. In: *The Annals of Statistics*, pp. 1986–2018.

Celisse, A., Daudin, J.-J., and Pierre, L. (2012). “Consistency of maximum-likelihood and variational estimators in the stochastic block model”. In: *Electronic Journal of Statistics* 6, pp. 1847–1899.

Cheang, G. H. (2010). “Approximation with neural networks activated by ramp sigmoids”. In: *Journal of Approximation Theory* 162.8, pp. 1450–1465.

Cheang, G. H. and Barron, A. R. (2000). “A better approximation for balls”. In: *Journal of Approximation Theory* 104.2, pp. 183–203.

Cheng, Y., Wang, D., Zhou, P., et al. (2018). “Model compression and acceleration for deep neural networks: The principles, progress, and challenges”. In: *IEEE Signal Processing Magazine* 35.1, pp. 126–136.

Chérief-Abdellatif, B.-E. (2020). “Convergence rates of variational inference in sparse deep learning”. In: *Proceedings of the 37th International Conference on Machine Learning (ICML 2020).* Vienna, Austria.

Chérief-Abdellatif, B.-E. and Alquier, P. (2018). “Consistency of variational Bayes inference for estimation and model selection in mixtures”. In: *Electronic Journal of Statistics* 12.2, pp. 2995–3035.

Cybenko, G. (1989). “Approximation by superpositions of a sigmoidal function”. In: *Mathematics of Control, Signals and Systems.*

Deng, W., Zhang, X., Liang, F., and Lin, G. (2019). “An adaptive empirical Bayesian method for sparse deep learning”. In: *33rd Conference on Neural Information Processing Systems (NeurIPS 2019).* Vancouver, Canada.

Dieng, A. B., Tran, D., Ranganath, R., Paisley, J., and Blei, D. M. (2017). “Variational Inference via χ^2-Upper Bound Minimization”. In: *31st Conference on Neural Information Processing Systems (NIPS 2017)*. Long Beach, CA.

Feng, J. and Simon, N. (2017). “Sparse input neural networks for high-dimensional nonparametric regression and classification”. In: *arXiv preprint arXiv:1711.07592*.

Frankle, J. and Carbin, M. (2018). “The lottery ticket hypothesis: finding sparse, trainable neural networks”. In: *arXiv preprint arXiv:1803.03635*.

Gale, T., Elsen, E., and Hooker, S. (2019). “The state of sparsity in deep neural networks”. In: *arXiv preprint arXiv:1902.09574*.

Gao, C., Vaart, A. W. van der, and Zhou, H. H. (2020). “A general framework for Bayes structured linear models”. In: *Annals of Statistics* 48, pp. 2848–2878.

George, E. and McCulloch, R. (1993). “Variable selection via Gibbs sampling”. In: *Journal of the American Statistical Association* 88, pp. 881–889.

Ghosal, S. (1999). “Asymptotic normality of posterior distributions in high- dimensional linear models”. In: *Bernoulli* 5, pp. 315–331.

Ghosal, S. and Van Der Vaart, A. (2007). “Convergence rates of posterior distributions for noniid observations”. In: *The Annals of Statistics* 35.1, pp. 192–223.

Ghosh, P. and Chakrabarti, A. (2015). “Posterior concentration properties of a general class of shrinkage estimators around nearly black vectors”. In: *arXiv preprint arXiv:1412.8161*.

Ghosh, S. and Doshi-Velez, F. (2017). “Model selection in Bayesian neural networks via horseshoe priors”. In: *arXiv preprint arXiv:1705.10388*.

Ghosh, S., Yao, J., and Doshi-Velez, F. (2018). “structured variational learning of Bayesian neural networks with horseshoe priors”. In: *Proceedings of the 35th International Conference on Machine Learning (ICML 2018)*. Stockholm, Sweden.

Giordano, R., Broderick, T., and Jordan, M. (2018). “Covariances, robustness, and variational Bayes”. In: *Journal of Machine Learning Research* 19, pp. 1–49.

Giordano, R., Broderick, T., and Jordan, M. (2015). “Linear response methods for accurate covariance estimates from mean field variational Bayes”. In: *Advances in Neural Information Processing Systems*, pp. 1441–1449.

Glorot, X., Bordes, A., and Bengio, Y. (2011). “Deep sparse rectifier neural networks”. In: *Proceedings of the 14th International Conference on Artificial Intelligence and Statistics (AISTATS) 2011*. Fort Lauderdale, FL.

Goldt, S., Advani, M. S., Saxe, A. M., Krzakala, F., and ZdeborovÃ¡, L. (2019). “Dynamics of stochastic gradient descent for two-layer neural networks in the teacher-student setup”. In: *33rd Conference on Neural Information Processing Systems (NeurIPS 2019).* Vancouver, Canada.

Goodfellow, I., Bengio, Y., and Courville, A. (2016). *Deep Learning.* MIT Press.

Graves, A. (2011). “Practical variational inference for neural networks”. In: *Advances in Neural Information Processing Systems 24 (NIPS 2011)*, pp. 2348–2356.

Griffin, J. E. and Brown, P. J. (2012). “Structuring shrinkage: some correlated priors for regression”. In: *Biometrika* 99, pp. 481–487.

Guo, Y., Zhang, C., Zhang, C., and Chen, Y. (2018). “Sparse DNNs with improved adversarial robustness”. In: *32nd Conference on Neural Information Processing Systems (NeurIPS 2018).* Montréal, Canada., pp. 240–249.

Hall, P., Ormerod, J., and Wand, M. (2011). “Theory of Gaussian variational approximation for a Poisson mixed model”. In: *Statistica Sinica* 21, pp. 369–389.

Hall, P., Pham, T., Wand, M., and Wang, S. (2011). “Asymptotic normality and valid inference for Gaussian variational approximation”. In: *The Annals of Statistic* 39.5, pp. 2502–2532.

Han, S., Mao, H., and Dally, W. (2016). “Deep compression: compressing deep neural networks with pruning, trained quantization and huffman coding”. In: *International Conference on Learning Representations (ICLR).*

Hans, C. (2009). “Bayesian lasso regression”. In: *Biometrika* 96, pp. 835–845.

Hastie, T., Tibshirani, R., and Friedman, J. H. (2001). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction.* New York: Springer.

Hernández-Lobato, J. and Adams, R. (2015). “Probabilistic backpropagation for scalable learning of bayesian neural networks”. In: *Proceedings of the 32nd International Conference on Machine Learning (ICML 2015).* Lille, France.

Hoffman, M. D., Blei, D. M., Wang, C., and Paisley, J. (2013). “Stochastic variational inference”. In: *The Journal of Machine Learning Research* 14.1, pp. 1303–1347.

Huang, X., Wang, J., and Liang, F. (2016). “A Variational Algorithm for Bayesian Variable Selection”. In: *arXiv preprint arXiv:1602.07640.*

Hubin, A. and Storvik, G. (2019). “Combining model and parameter uncertainty in Bayesian neural networks”. In: *arXiv:1903.07594.*

Ishwaran, H. and Rao, S. (2005). “Spike and slab variable selection: Frequentist and Bayesian strategies”. In: *Annals of Statistics* 33.2, pp. 730–773.

Ismailov, V. (2017). “Approximation by sums of ridge functions with fixed directions”. In: *St. Petersburg Mathematical Journal* 28.6, pp. 741–772.

Jaiswal, P., Rao, V. A., and Honnapppa, H. (2019). “Asymptotic Consistency of $\alpha-$Rényi-Approximate Posteriors”. In: *arXiv preprint arXiv:1902.01902*.

Jang, E., Gu, S., and Poole, B. (2017). “Categorical reparameterization with gumbel-softmax”. In: *International Conference on Learning Representations (ICLR 2017)*.

Johnson, V. E. and Rossel, D. (2012). “Bayesian Model Selection in High-dimensional Settings”. In: *Journal of the American Statistical Association* 107, pp. 649–660.

Jordan, M., Ghahramani, Z., Jaakkola, T., et al. (1999). “An introduction to variational methods for graphical models”. In: *Machine Learning*.

Kingma, D. and Ba, J. L. (2015). “ADAM: A method for stochastic optimization”. In: *International Conference on Learning Representations (ICLR 2015)*.

Kingma, D. and Welling, M. (2014). “Auto-Encoding Variational Bayes”. In: *arXiv:1312.6114*.

Le Cam, L. (1986). *Asymptotic methods in statistical decision theory*. New York: Springer Science & Business Media.

Li, Y. and Turner, R. E. (2016). “Rényi Divergence variational inference”. In: *30th Conference on Neural Information Processing Systems (NIPS 2016)*. Barcelona, Spain.

Liang, F., Li, Q., and Zhou, L. (2018). “Bayesian neural networks for selection of drug sensitive genes”. In: *Journal of the American Statistical Association* 113 (523), pp. 955–972.

Louizos, C., Welling, M., and Kingma, D. P. (2018). “Learning sparse neural networks through l0 regularization”. In: *ICLR 2018*.

MacKay, D. (1992). “A practical bayesian framework for backpropagation networks”. In: *Nerual Computation*.

Maddison, C., Mnih, A., and Teh, Y. W. (2017). “The concrete distribution: a continuous relaxation of discrete random variables”. In: *International Conference on Learning Representations (ICLR 2017)*.

Mhasker, H., Liao, Q., and Poggio, T. (2017). “When and why are deep networks better than shallow ones?” In: *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence (AAAI-17)*, pp. 2343–2349.

Minka, T. P. (2001). “Expectation Propagation for Approximate Bayesian Inference”. In: *Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence (UAI2001).*

Mitchell, T. J. and Beauchamp, J. J. (1988). “Bayesian variable selection in linear regression”. In: *Journal of the American Statistical Association* 83 (404), pp. 1023–1032.

Mocanu, D., Mocanu, E., Stone, P., et al. (2018). “Scalable training of artificial neural networks with adaptive sparse connectivity inspired by network science”. In: *Nature Communications* 9, p. 2383.

Molchanov, D., Ashukha, A., and Vetrov, D. (2017). “Variational dropout sparsifies deep neural networks”. In: *Proceedings of the 34th International Conference on Machine Learning (ICML 2017)*, pp. 2498–2507.

Narisetty, N. N. and He, X. (2014). “Bayesian variable selection with shrinking and diffusing priors”. In: *The Annals of Statistics* 42.

Neal, R. (1992). “Bayesian learning via stochastic dynamics”. In: *Advances in Neural Information Processing Systems 5 (NIPS 1992)*, pp. 475–482.

Neklyudov, K., Molchanov, D., Ashukha, A., and Vetrov, D. (2017). “Structured Bayesian pruning via log-normal multiplicative noise”. In: *31st Conference on Neural Information Processing Systems (NIPS 2017).* Long Beach, CA.

Ormerod, J., You, C., and Muller, S. (2017). “A variational Bayes approach to variable selection”. In: *Electronic Journal of Statistics* 11.2, pp. 3549–3594.

Park, T. and Casella, G. (2008). “The Bayesian Lasso”. In: *Journal of the American Statistical Association* 103, pp. 681–686.

Pati, D., Bhattacharya, A., and Yang, Y. (2018). “On the Statistical optimality of variational Bayes”. In: *Proceedings of the 21st International Conference on Artificial Intelligence and Statistics (AISTATS) 2018.* Lanzarote, Spain.

Polson, N. and Ročková, V. (2018). “Posterior concentration for sparse deep learning”. In: *32nd Conference on Neural Information Processing Systems (NeurIPS 2018).* MontrÃ©al, Canada, pp. 930–941.

Ranganath, R., Gerrish, S., and Blei, D. M. (2013). “Black box variational inference”. In: *arXiv preprint arXiv:1401.0118.*

Ranganath, R., Tran, D., and Blei, D. M. (2016). “Hierarchical variational models”. In: *Proceedings of the 33rd International Conference on Machine Learning (ICML 2016).* New York, NY, USA.

Ray, K. and Szabo, B. (2020). “Variational Bayes for high-dimensional linear regression with sparse priors”. In: *arXiv:1904.07150*.

Reid, S., TIbshirani, R., and Friedman, J. (2016). “A study of error variance estimation in lasso regression”. In: *Statistica Sinica* 26 (1), pp. 35–67.

Rezende, D., Mohamed, S., and Wierstra, D. (2014). “Stochastic backpropagation and approximate inference in deep generative models”. In: *Proceedings of the 31st International Conference on Machine Learning (ICML 14)*. Beijing, China, pp. 1278–1286.

Ročková, V. and George, E. (2014). “EMVS: the EM approach to Bayesian variable selection”. In: *Journal of the American Statistical Association* 109 (506), pp. 828–846.

Ročková, V. and George, E. (2018). “The Spike-and-Slab LASSO”. In: *Journal of the American Statistical Association* 113 (521), pp. 431–444.

Rolnick, D. and Tegmark, M. (2018). “The power of deeper networks for expressing natural functions”. In: *International Conference on Learning Representations (ICLR)*.

Schmidt-Hieber, J. (2017). “Nonparametric regression using deep neural networks with ReLU activation function”. In: *arXiv:1708.06633*.

Sønderby, C., Raiko, T., Maaløe, L., Sønderby, S., and Ole, W. (2016). “How to Train Deep Variational Autoencoders and Probabilistic Ladder Networks”. In: *Proceedings of the 33rd International Conference on International Conference on Machine Learning (ICML 16)*. New York, NY.

Song, Q. and Liang, F. (2014). “A split-and-merge Bayesian variable selection approach for ultra-high dimensional regression”. In: *Journal of the Royal Statistical Society, Series B* 77, pp. 947–972.

Song, Q. (2020). “Bayesian shrinkage towards sharp minimaxity”. In: *Electronic Journal of Statistics* 14, pp. 2714–2741.

Song, Q. and Liang, F. (2017). “Nearly optimal Bayesian shrinkage for high dimensional regression”. In: *arXiv:1712.08964*.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. (2014). “Dropout: a simple way to prevent neural networks from overfitting”. In: *Journal of Machine Learning Research* 15, pp. 1929–1958.

Tian, Y. (2018). “A theoretical framework for deep locally connected relu network”. In: *arXiv preprint arXiv:1809.10829*.

Tran, D., Blei, D. M., and Airoldi, E. M. (2015). “Copula variational inference.” In: *Advances in Neural Information Processing Systems 28 (NIPS 2015).*

Van Der Pas, S. L., Salomond, J. B., and Schmidt-Hieber, J. (2016). “Conditions for posterior contraction in the sparse normal means problem”. In: *Electronic journal of statistics* 10, pp. 976–1000.

Van Der Pas, S. L., Szabó, B., and Van Der Vaart, A. (2017). “Uncertainty quantification for the horseshoe (with discussion)”. In: *Bayesian Analysis* 12, pp. 1221–1274.

Van Der Pas, S. L., Kleijn, B. J., and Van Der Vaart, A. W. (2014). “The horseshoe estimator: Posterior concentration around nearly black vectors”. In: *Electronic Journal of Statistics* 8, pp. 2585–2618.

Vershynin, R. (2012). “Introduction to the non-asymptotic analysis of random matrices”. In: *Compressed Sensing: Theory and Applications*. Ed. by Y. Eldar and G. Kutyniok. Cambridge University Press, pp. 210–268.

Wang, B. and Titterington, M. (2004). “Inadequacy of interval estimates corresponding to variational Bayesian approximations”. In: *Workshop on Artificial Intelligence and Statistics*, pp. 373–380.

Wang, Y. and Blei, D. (2019). “Frequentist Consistency of Variational Bayes”. In: *Journal of the American Statistical Association* 114, pp. 1147–1161.

Westling, T. and McCormick, T. H. (2019). “Beyond prediction: A framework for inference with variational approximations in mixture models”. In: *Journal of Computational and Graphical Statistics* 28.4, pp. 778–789.

Yang, Y., Wainwright, M. J., and Jordan, M. I. (2016). “On the computational complexity of high-dimensional Bayesian variable selection”. In: *The Annals of Statistics* 44, pp. 2497–2532.

Yang, Y., Pati, D., and Bhattacharya, A. (2020). “α-variational inference with statistical guarantees”. In: *Annals of Statistics* 48.2, pp. 886–905.

Ye, M. and Sun, Y. (2018). “Variable selection via penalized neural network: a drop-out-one loss approach”. In: *Proceedings of the 35th International Conference on International Conference on Machine Learning (ICML 18)*. Stockholm, Sweden, pp. 5620–5629.

You, C., Ormerod, J., and Muller, S. (2014). “On variational Bayes estimation and variational information criteria for linear regression models”. In: *Australian and New Zealand Journal of Statistics* 56.1, pp. 73–87.

Zhang, C. H. (2010). “Nearly unbiased variable selection under minimax concave penalty”. In: *The Annals of Statistics* 38.

Zhang, F. and Gao, C. (2019). “Convergence rates of variational posterior distributions”. In: *arXiv preprint arXiv:1712.02519*.

Zhu, M. and Gupta, S. (2018). “To prune, or not to prune: Exploring the efficacy of pruning for model compression”. In: *International Conference on Learning Representations (ICLR)*.

A. APPENDIX TO CHAPTER 2

A.1 Comparison of Minimizing Joint KL and Marginal KL

Our presented theory investigates the asymptotics of the variational Bayes distribution that minimizes the marginal KL divergence of $\boldsymbol{\beta}$. In such case, the negative ELBO is

$$\widetilde{\Omega} = -\int \log p(\boldsymbol{Y}|\boldsymbol{\beta})q(\boldsymbol{\beta})d\boldsymbol{\beta} + \mathrm{KL}(q(\boldsymbol{\beta})\|\pi(\boldsymbol{\beta})), \tag{A.1}$$

where for $\mathrm{j} = 1, \dots, p_n$,

$$\pi(\beta_\mathrm{j}) = \frac{1}{\sqrt{\nu_0}s_0}\left(1 + \nu_0^{-1}\left(\frac{\beta_\mathrm{j}}{s_0}\right)^2\right)^{-\frac{\nu_0+1}{2}}, \text{ and } q(\beta_\mathrm{j}) = \frac{1}{\sqrt{\nu}s}\left(1 + (\nu)^{-1}\left(\frac{\beta_\mathrm{j} - \widetilde{\mu}_\mathrm{j}}{s_\mathrm{j}}\right)^2\right)^{-\frac{\nu+1}{2}},$$

with $s_0 = \sqrt{b_n/a_0}$, $\nu_0 = 2a_0$, $s = \sqrt{b_\mathrm{j}/a_\mathrm{j}}$ and $\nu = 2a_\mathrm{j}$. However $\mathrm{KL}(q(\boldsymbol{\beta})\|\pi(\boldsymbol{\beta}))$ has no analytical expression, and the optimization of (A.1) will then require Monte Carlo estimation and gradient descent type algorithms.

Therefore, for the simplicity of the computation, the ELBO optimization algorithm described in Section 2.4 targets to minimize the joint KL divergence of $(\boldsymbol{\beta}, \boldsymbol{\lambda})$ rather than the marginal KL divergence of $\boldsymbol{\beta}$. In other words, there is a gap between our computational algorithm and our theory.

To justify that our implemented procedure (i.e., minimizing the joint KL divergence of $(\boldsymbol{\beta}, \boldsymbol{\lambda})$) is a close approximation of the variational procedure studied by our theory (i.e., minimizing the marginal KL divergence of $(\boldsymbol{\beta})$), We compare the two procedures via a toy example. Specifically, we would like to compare variational posterior means $\boldsymbol{\mu}$ (by minimizing (2.4)) and $\widetilde{\boldsymbol{\mu}}$ (by minimizing (A.1)). Consider a linear model with $n = 100$, $p_n = 100$ and $\boldsymbol{\beta}^0 = (10, 10, 10, 10, 10, 0, \dots, 0)^T$. Suppose σ is known and equals 1. For both two procedures, we choose $a_0 = 2$ and $b_n/a_0 = \log(p_n)/[np_n^{2+1/a_0}p_n^{6/a_0}]$. We use Lasso estimator for both the initial value of $\boldsymbol{\mu}$ and $\widetilde{\boldsymbol{\mu}}$, and Adam (Kingma and Ba 2015) is used for minimizing (A.1) with the learning rate being 0.001.

We run the experiment for 100 times, and the means and standard deviations of the mean squared error (MSE) $(\sum_{k=1}^K(\mu_k - \widetilde{\mu}_k)^2/K)$ for both the nonzero entries $\boldsymbol{\beta}_{\xi^0}$ and zero entries $\boldsymbol{\beta}_{(\xi^0)^c}$ are reported: For $\boldsymbol{\beta}_{\xi^0}$, the MSE is 0.0108 ± 0.0038; For $\boldsymbol{\beta}_{(\xi^0)^c}$, the MSE is 0.0006 ± 0.0008.

This toy example shows there is little estimation difference in minimizing (2.4) or (A.1), and thus in practice the Algorithm 1 in the main text is preferred due to its simple form of coordinate descent update.

B. APPENDIX TO CHAPTER 3

B.1 Toy Example: Linear Regression

In this section, we aim to demonstrate that there is little difference between the results using inverse-CDF reparameterization and Gumbel-softmax approximation via a toy example.

Consider a linear regression model:

$$Y_\mathrm{i} = X_\mathrm{i}^T\beta + \epsilon_\mathrm{i}, \quad \epsilon_\mathrm{i} \sim \mathcal{N}(0,1), \quad \mathrm{i} = 1, \dots, n,$$

We simulate a dataset with 1000 observations and 200 predictors, where $\beta_{50} = \beta_{100} = \beta_{150} = 10$, $\beta_{75} = \beta_{125} = -10$ and $\beta_\mathrm{j} = 0$ for all other j.

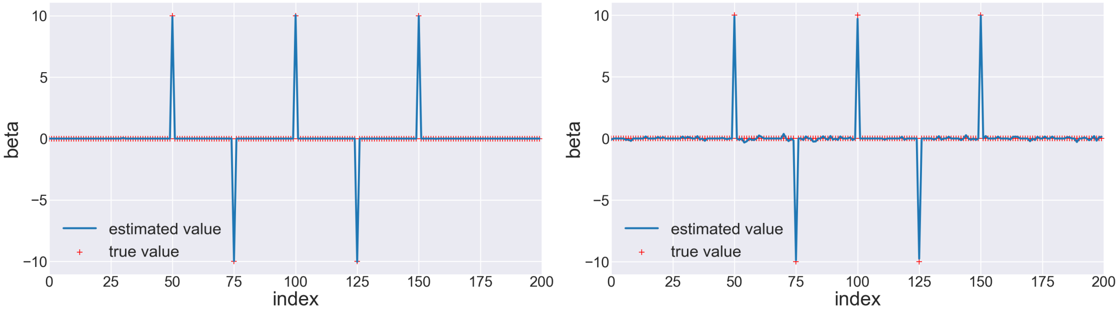
A spike-and-slab prior is imposed on β such that

$$\beta_\mathrm{j}|\gamma_\mathrm{j} \sim \gamma_\mathrm{j}\mathcal{N}(0,\sigma_0^2) + (1-\gamma_\mathrm{j})\delta_0, \quad \gamma_\mathrm{j} \sim \mathrm{Bern}(\lambda),$$

for $\mathrm{j} = 1, \dots, 200$, where $\sigma_0 = 5$ and $\lambda = 0.03$. The variational distribution $q(\beta)\mathcal{Q}$ is chosen as

$$\beta_\mathrm{j}|\gamma_\mathrm{j} \sim \gamma_\mathrm{j}\mathcal{N}(\mu_\mathrm{j},\sigma_\mathrm{j}^2) + (1-\gamma_\mathrm{j})\delta_0, \quad \gamma_\mathrm{j} \sim \mathrm{Bern}(\phi_\mathrm{j}).$$

We use both Gumbel-softmax approximation and inverse-CDF reparameterization for the stochastic optimization of ELBO, and plot posterior mean $\mathbb{E}_{\widehat{q}(\beta)}(\beta_\mathrm{j}|\gamma_\mathrm{j})$ (blue curve) against the true value (red curve). Figure *B*.1 shows that inverse-CDF reparameterization exhibits only slightly higher error in estimating zero coefficients than the Gumbel-softmax approximation, which indicates the two methods has little difference on this toy example.



(a) Gumbel-softmax reparametrization

(b) Inverse-CDF reparametrization

Figure B.1. Linear regression

VITA

Jincheng Bai was born and raised in Xi'an, Shaanxi Province, China. He obtained a bachelor's degree in information and computing science (computational maths) at the Xi'an Jiaotong University in 2011. Before joining Purdue, he received a master's degree in applied maths at the Bowling Green State University in 2013. Later on, he entered Purdue University to pursue a doctoral degree in statistics in 2014 and earned the degree in 2020. His research interests center around machine learning, deep learning and Bayesian statistics. He had internship experience at Google and J.P. Morgan, respectively. After graduation, he would like to work as a full-time Quantitative Researcher at J.P. Morgan.