# Efficient Variational Inference for Sparse Deep Learning with Theoretical Guarantee

Jincheng Bai, Qifan Song, Guang Cheng

## Introduction

**Motivation:**

- Compress deep learning model for hardware limited devices.
- Recover potential sparsity structure of the target function.
- Variable selection.

**Existing approaches:**

Pruning methods ( no theoretical guarantee on choosing the threshold):

- Frequentist: Zhu and Gupta 2018; Frankle and Carbin 2018, etc.
- Bayesian: Molchanov et al. 2017; Ghosh et al. 2018, etc.

Theoretical work (no efficient implementation):

- Polson and Rockova 2018; Cherief-Abdellatif 2020, etc.

**Our contribution:**

A complete package of both theory and computation for sparse DNN from a Bayesian perspective.

## Model setup

**Nonparametric regression**

Consider a nonparametric regression model with random covariates:

$$Y_i = f_0(X_i) + \epsilon_i, \quad i = 1, \dots, n,$$

where $X_i \sim \mathcal{U}([-1,1]^p)$ and $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$.

**Deep neural network**

The DNN $f_\theta(X)$ is used to approximate $f_0(X)$, where $\theta$ denotes all the coefficients in the neural network. The network configuration includes

- L: number of hidden layers
- N: width of the network (assume equal width)
- s: sparsity level ($\|\theta\|_0 \le s$)

**Optimal sparsity level:**

Define $s^*$ as

$$s^* = \text{argmin}_s\{r_n(L,N,s) + \xi_n(L,N,s)\},$$

such that $s^*$ strikes the balance (trade-off) between the variational error $r_n$ and the approximation error $\xi_n$. Note that $s^*$ is generally unknown, but our modeling is capable of automatically attaining the same rate of convergence as if $s^*$ is known.

## Variational Inference

**Sparse Bayesian deep learning with spike-and-slab prior**

- Prior distribution $\pi(\theta)$:

  $$\theta_i|\gamma_i \sim \gamma_i \mathcal{N}(0, \sigma_0^2) + (1-\gamma_i)\delta_0, \ \gamma_i \sim Bern(\lambda), i = 1, \dots, T,$$

  Note: the theoretical guarantees will be established under proper deterministic choices of hyperparameters $\lambda$ and $\sigma_0$.

- Variational distribution $q(\theta)$:

  $$\theta_i|\gamma_i \sim \gamma_i \mathcal{N}(\mu_i, \sigma_i^2) + (1-\gamma_i)\delta_0, \ \gamma_i \sim Bern(\phi_i), i = 1, \dots, T.$$

  The variational parameters are $(\mu_i, \sigma_i, \phi_i)$, the transformed variational parameters are denoted as $\omega$.

## Theoretical Results

**Conditions:**

The main theorems are under the following conditions.

- The activation function is 1-Lipschitz continuous.
- The hyperparameter $\sigma_0^2$ is set to be some constant, and $\lambda$ satisfies

  $\log(1/\lambda) = O\{(L+1)\log N + \log(p\sqrt{n/s^*})\}$ and

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

**Main theorems:**

- $P_0$ and $P_\theta$ are the probability measure corresponding to $f_0$ and $f_\theta$ ; $D$ is the dataset; $l_n(P_0, P_\theta) = \log(p_0(D)/p_\theta(D))$.
- Given some constant $B > 0$, we define

  $$r_n^* \coloneqq r_n(L,N,s^*) = ((L+1)s^*/n)\log N + (s^*/n)\log(p\sqrt{n/s^*})$$
  $$\xi_n^* \coloneqq \xi_n(L,N,s^*) = \inf_{\|\theta\|_\infty \le B} \|f_\theta - f_0\|_\infty^2$$
- The squared Hellinger distance is defined as

  $$d^2(P_\theta, P_0) = E_X(1 - \exp(-(f_\theta(X) - f_0(X))^2/8\sigma_\epsilon^2)$$
- In addition, let $s_n = s^* \log^{2\delta-1} n$ for any $\delta > 1$.

**Lemma 1**

With dominating probability,

$$\inf_{q(\theta)\in\mathcal{Q}} \left\{ KL(q(\theta)\|\pi(\theta|\lambda)) + \int_\Theta l_n(P_0, P_\theta) q(\theta) d\theta \right\} \le C' n (r_n^* + \xi_n^*),$$

where $C'$ is any diverging sequence.

**Lemma 2**

If $\sigma_0^2$ is set to be constant and $\lambda \le T^{-1}\exp\{-Mnr_n^*/s_n\}$ for any positive diverging sequence $M$, then with dominating probability,

$$\int_\Theta d^2(P_\theta, P_0)\hat{q}(\theta) \le C\varepsilon_n^{*2} + \frac{3}{n} \inf_{q(\theta)\in\mathcal{Q}} \left\{ KL(q(\theta)\|\pi(\theta|\lambda)) + \int_\Theta l_n(P_0, P_\theta) q(\theta) d\theta \right\},$$

where C is some constant and the estimation error is

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

Lemma 1 and Lemma 2 together imply the following guarantee for VB posterior:

**Theorem 1**

Let $\sigma_0^2$ be a constant and $-\log\lambda = \log T + \delta[(L+1)\log N + \log\sqrt{n}\,p]$ for any constant $\delta > 0$, we have with high probability

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

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

## Implementation

Since it is impossible to reparameterize the discrete variable $\gamma$ by a continuous system, we apply the Gumbel-softmax approximation (Maddison et al. 2017), and $\gamma_i \sim \text{Bern}(\phi_i)$ is approximated by $\widetilde{\gamma}_i \sim \text{Gumbel-softmax}(\phi_i, \tau)$, where

$$\widetilde{\gamma}_i = (1 + \exp(-\eta_i/\tau))^{-1}, \quad \eta_i = \log\frac{\phi_i}{1-\phi_i} + \log\frac{u_i}{1-u_i},$$
$$u_i \sim \mathcal{U}(0,1), \quad \tau > 0 \text{ is the temperature.}$$

**Algorithm 1** Variational inference for sparse BNN with normal slab distribution.

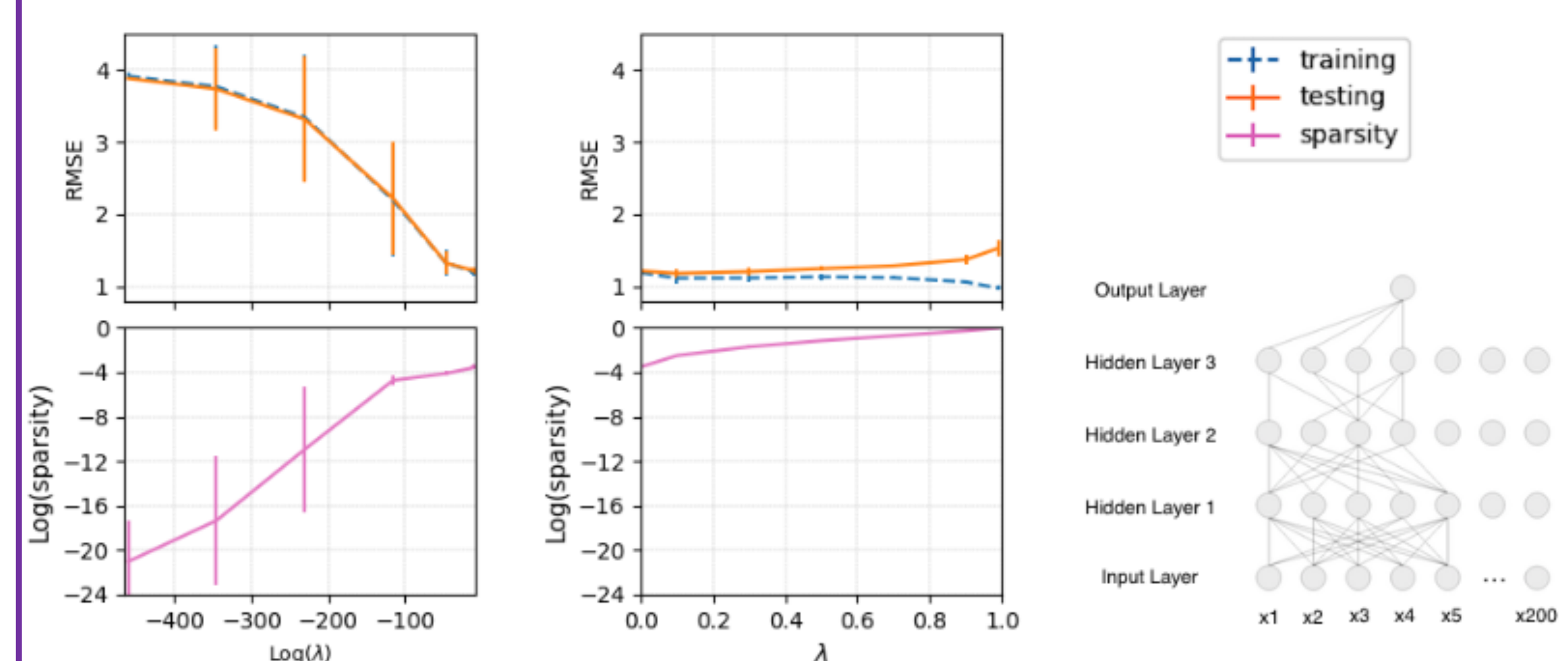
1: parameters: $\omega = (\mu, \sigma', \phi')$ , where $\sigma_i = \log(1+\exp(\sigma_i'))$, $\phi_i = (1+\exp(\phi_i'))^{-1}$
2: objective: negative ELBO $\Omega(\omega)$ ,
3: **repeat**
4: $D^m \leftarrow$ Randomly draw a minibatch of size $m$ from $D$
5: $\epsilon_i, u_i \leftarrow$ Randomly draw $K$ samples from $\mathcal{N}(0,1)$ and $\mathcal{U}(0,1)$
6: $\widetilde{\Omega}^m(\omega) \leftarrow$ Stochastic estimator with ($D^m$, $\omega$, $\epsilon$, $u$); Use $\gamma$ in the forward pass
7: $\nabla_\omega \widetilde{\Omega}^m(\omega) \leftarrow$ Stochastic estimator with ($D^m$, $\omega$, $\epsilon$, $u$); Use $\widetilde{\gamma}$ in the backward pass
8: $\omega \leftarrow$ Update with $\nabla_\omega \widetilde{\Omega}^m(\omega)$ using gradient descent algorithms (e.g. SGD, RMSprop
9: or Adam)
10: **until** convergence of $\widetilde{\Omega}^m(\omega)$
11: **return** $\omega$

## Experiment

Consider the following sparse target function $f_0$:

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

The $\lambda_{opt}$ is chosen according to Theorem 1.




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

Figure 1. Nonlinear regression example.

Figure 1(a) and 1(b) show $\lambda_{opt}$ is a reasonable choice and a possible network structure is provided in 1(c).

**Selected references**

Ghosh, S., Yao, J., and Doshi-Velez, F. (2018). Structured variational learning of Bayesian neuralnetworks with horseshoe priors. In ICML 2018.

Polson, N. and Rockova, V. (2018). Posterior concentration for sparse deep learning. In NeurIPS 2018.

Zhu, M. and Gupta, S. (2018). To prune, or not to prune: Exploring the efficacy of pruning for model compression. In ICLR 2018