Summary: Deterministic Variational Inference for Robust Bayesian Neural Networks

summarized by Ibrahim Akbar
Department of Electrical and Computer Engineering
University of California, San Diego

***Abstract*—This paper considers the problem of variational inference in bayesian neural networks. Specifically, they are concerned with improving the efficiency and robustness of learning priors over the weight space. To do this they introduce a deterministic approximation of the reconstruction term in the ELBO and a general robust method for prior selection from hierarchical priors for the KL Divergence.**

I. Introduction

Full bayesian inference for most neural network is intractable due to the model's nonlinearity causing the true posterior distribution to be highly complex. This has led to the use of variational methods which attempt to approximate the posterior. With regard to bayesian variational inference; this paper addresses the general issue of prior selection to reduce the sensitivity BNNs have toward such initializations and the high variance that arises in the gradients when employing Monte Carlo approximation techniques.

The authors list their contributions as:

- Development of a deterministic procedure for propagating uncertain activations through neural networks with uncertain weights and ReLU or Heaviside activation functions.
- Development of an EB method for principled tuning of weight priors during BNN training.
- Experimental results showing the accuracy and efficiency of our method and applicability to heteroscedastic and homoscedastic regression on real datasets.

II. Problem Formulation

Given a model $\mathcal{M}$ parameterized by $\boldsymbol{\omega}$ and a dataset $\mathcal{D} = \{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^{\mathbb{N}}$, the inference goal is to determine the posterior distribution $p(\boldsymbol{\omega} \mid \mathbf{X}, \mathbf{Y})$ through Bayes' rule. Since this posterior is generally intractable we consider the parameterized variational distribution $q(\boldsymbol{\omega}; \boldsymbol{\theta})$ that is optimal for $\boldsymbol{\theta}^*$ in terms of KL Divergence.

$$\boldsymbol{\theta}^* = \arg\min_{\boldsymbol{\theta} \in \Theta} KL\left[q(\boldsymbol{\omega}; \boldsymbol{\theta}) || p(\boldsymbol{\omega} \mid \mathbf{X}, \mathbf{Y})\right]$$

Given a prior, $p(\boldsymbol{\omega})$, we can can reformulate this as the evidence lower bound (ELBO):

$$\boldsymbol{\theta}^* = \arg\max_{\boldsymbol{\theta} \in \Theta} \left\{ \mathbb{E}_{\boldsymbol{\omega} \sim q}\big[\log p(\mathbf{Y} \mid \mathbf{X}, \boldsymbol{\omega})\big] - KL\big[q(\boldsymbol{\omega}; \boldsymbol{\theta}) || p(\boldsymbol{\omega})\big] \right\} \tag{1}$$

The goal of this paper is to derive an explicit deterministic approximation of the first term, known as reconstruction term, and choose priors $p(\boldsymbol{\omega})$ empirically to increase robustness to the choice of variance parameters.

III. Deterministic Variational Approach

There are two main parts for computing the reconstruction term: propagation of distributions through activations to compute $\tilde{q}(\mathbf{a}^L)$, and evaluation of unparameterized log-likelihood $\mathcal{L}$. In Fig 1 we can see the general architecture used to accomplish these tasks.

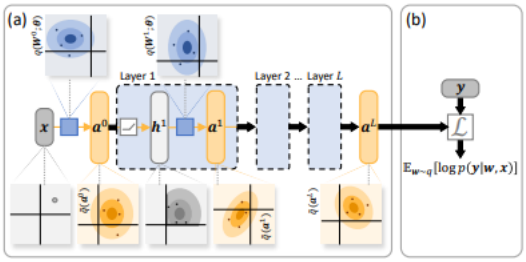

Fig. 1: Feed-forward architecture for reconstruction term computation.

Moment Propagation.
We can consider the model, $\mathcal{M}$, as a set of layers each containing an non-linear and affine transformation,

$$\mathcal{M} := \{(\boldsymbol{h}^l, \boldsymbol{a}^l) : \boldsymbol{h}^l = f(\boldsymbol{a}^{l-1}), \boldsymbol{a}^l = \boldsymbol{h}^l \boldsymbol{W}^l + \boldsymbol{b}^l\}_{l=1}^{\mathbb{N}}$$

where $\{\boldsymbol{W}, \boldsymbol{b}\} \subset \boldsymbol{\omega}$ are random variables representing the weights and are assumed independent per layer. $\boldsymbol{a}^l$ is argued to be Gaussian under the Central Limit Theorem given a sufficiently large latent space and finite 1^{st} and 2^{nd} moment since it is formulated as the linear combination of the elements of $\boldsymbol{h}^l$. Given that $\boldsymbol{a}^l$ is Gaussian we can appoximate the 1^{st} and 2^{nd} moment,

$$\langle a_i \rangle = \langle h_j \rangle \langle W_{ji} \rangle + \langle b_i \rangle \tag{2}$$

$$\mathrm{Cov}(a_i, a_k) =$$

$$\langle h_j h_l \rangle \mathrm{Cov}(W_{ji}, W_{lk}) + \langle W_{ji} \rangle \mathrm{Cov}(h_j, h_l) \langle W_{lk} \rangle + \mathrm{Cov}(b_i, b_k) \tag{3}$$

where $\langle a_i \rangle := \mathbb{E}_q[a_i]$ and $h_j W_{ji} = \sum_{j=1}^{n} h_j W_{ji}$ is called Einstein notation. To reduce approximation, gaussian distributions are considered for the mean and covariance of the weights so that all that is left determine are the moments $\langle h_j \rangle$ and $\langle h_j h_l \rangle$

$$\langle h_j \rangle \propto \int f(\alpha_j) \exp\left[-\frac{(\alpha_j - \langle a_j^{l-1} \rangle)^2}{2\Sigma_{jj}^{l-1}} \right] d\alpha_j \tag{4}$$

$$\langle h_j h_l \rangle \propto \int f(\alpha_j) f(\alpha_l) \exp\left[-\frac{1}{2} \zeta^T \Lambda^{-1} \zeta \right] d\alpha_j d\alpha_l \tag{5}$$

$$\zeta = \begin{pmatrix} \alpha_j - \langle a_j^{l-1} \rangle \\ \alpha_l - \langle a_l^{l-1} \rangle \end{pmatrix}$$

$$\Lambda = \begin{pmatrix} \Sigma_{jj}^{l-1} & \Sigma_{jl}^{l-1} \\ \Sigma_{lj}^{l-1} & \Sigma_{ll}^{l-1} \end{pmatrix}$$

Closed form solutions exist for (4) when considering Heaviside or ReLU non-linearity for f. For (5), we can approximate the moment through,

$$\langle h_j h_l \rangle = S_{jl}^{l-1} \left\{ A(\mu_j^{l-1}, \mu_l^{l-1}, \rho_{jl}^{l-1}) + \exp[-Q(\mu_j^{l-1}, \mu_l^{l-1}, \rho_{jl}^{l-1})] \right\} \tag{6}$$

where the key idea is that the asymptotes A of the non-linearities as well as the residuals Q in the form of a polynomial provide a good first order approximation of the moment. Fig 2 provides a visual representation of this process. Due to CLT these approximations provide sufficient information for us to explicitly determine $\tilde{q}(\boldsymbol{a}^L)$ through sequential distribution propagation.

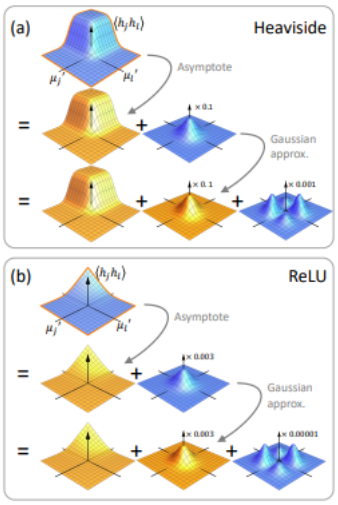

Fig. 2: Model activation function approximation.

Log-Likelihood Evaluation.
We can evaluate the expected log-likelihood $\mathbb{E}_{\boldsymbol{\omega}\sim q}\big[\log p(y \mid \boldsymbol{x}, \boldsymbol{\omega})\big]$ through directly evaluating $\mathbb{E}_{\boldsymbol{a}^L \sim q(\boldsymbol{a}^L)}\big[\log p(y \mid \boldsymbol{a}^L)\big]$ since $q(y \mid \boldsymbol{a}^L)$ is a parameter free transformation.

IV. Empirical Bayes for Variational BNNs

Considering a d-dimensional Gaussian prior, $p(\boldsymbol{\omega}) = \mathcal{N}(\mu_p, \Sigma_p)$, and variational distribution, $q = \mathcal{N}(\mu_q, \Sigma_q)$, the KL divergence has the form,

$$\frac{1}{2}\left[\log \frac{\det(\Sigma_p)}{\det(\Sigma_q)} - d + \text{tr}(\Sigma_p^{-1}\Sigma_q) + (\mu_p - \mu_q)^T \Sigma_p^{-1}(\mu_p - \mu_q)\right] \tag{7}$$

Rather than using this directly the authors propose conditioning the prior on a hyper-parameter $\boldsymbol{s}$ such that $\boldsymbol{\omega} \sim p(\boldsymbol{\omega} \mid \boldsymbol{s}); \boldsymbol{s} \sim p(\boldsymbol{s})$, where $\boldsymbol{s}$ is distributed according to a inverse gamma distribution and acts as a conjugate prior for the diagonal gaussian variance. Further through partioning the weights $\boldsymbol{\omega}$ into sets $\{\lambda\}$ such that an element s_λ of $\boldsymbol{s}$ can be assigned to each set,

$$s_\lambda \sim \text{Inv-Gamma}(\alpha, \beta), \quad w_i^\lambda \sim \mathcal{N}(0, s_\lambda)$$

we can consider solving the MAP optimization problem for the KL divergence,

$$s_\lambda^* = \arg\min_{s_\lambda} KL\Big[q(\boldsymbol{\omega}; \boldsymbol{\theta}) || p(\boldsymbol{\omega}^\lambda \mid s_\lambda) - \log p(s_\lambda)\Big]$$

This leads to the closed-form solution,

$$s_\lambda^* = \frac{\text{tr}(\Sigma_q^\lambda + \mu_q^\lambda (\mu_q^\lambda)^T) + 2\beta}{\Omega_\lambda + 2\alpha + 2}$$

where $\Omega_\lambda := |\lambda|$. We can then use s_λ^* to determine the diagonal entries of Σ_p and solve (1).

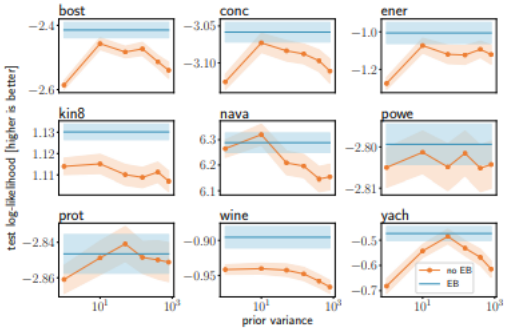

Fig. 3: Test Log-likelihood with tuned prior (orange) and EB (blue).

V. Experiments

The proposed method, DVI, is evaluated with small networks on UCI datasets in comparison with a large variety of other methods in the Appendix showing that it is capable of performing at the state of art in terms of log-likelihood. Furthermore, the algorithm is also tested extensively on it's own with a variety of configurations in the BNNs to see how the main assumptions hold of CLT, indepdendence, and trunctated polynomial hold. A further study that would be interesting would compare the variance estimates to observe the a rate of decay.

VI. Comments

- The deterministic approximation of the reconstruction term feels very heurstic-ish. Is there a better method to generalize this to a large set of non-linearities?
- Having a selective hyper-parameter throughout the network is an interesting idea. Was the inverse gamma the only prior considered?
- Appendix is extensive and a good reference.