ICML
International Conference
On Machine Learning

Reparameterized Importance Sampling for Robust Variational Bayesian Neural Networks

Yunfei Long
Harbin Engineering University
Harbin, Heilongjiang, China
2016064109@hrbeu.edu.cn

Zilin Tian
Harbin Engineering University
Harbin, Heilongjiang, China
tzl@hrbeu.edu.cn

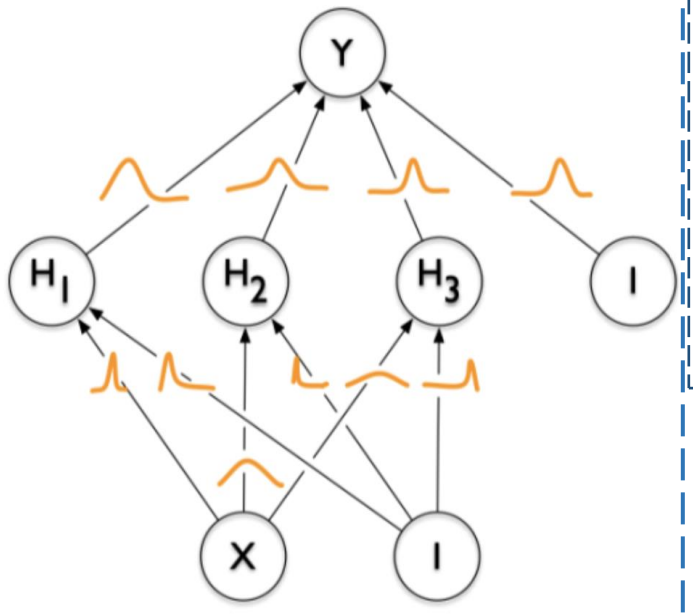
Liguo Zhang*
Harbin Engineering University
Harbin, Heilongjiang, China
zhangliguo@hrbeu.edu.cn

Huosheng Xu
Harbin Engineering University
Harbin, Heilongjiang, China
xuhuosheng@hrbeu.edu.cn

HARBIN ENGINEERING UNIVERSITY

Background

Bayesian Neural Network (BNN) is a probabilistic model that combines Bayesian inference with neural networks.



Bayesian neural network forward pass process

Data: $D = \{(x_n, y_n)|_{n=1}^{N}\}$

Weight : $\mathcal{W} = \{w_i\}$

Prior distribution : $p(\mathcal{W})$

Posterior distribution : $p(\mathcal{W}|D) \propto p(D|\mathcal{W})p(\mathcal{W})$

This characteristic empowers BNNs to estimate the level of uncertainty in their predictions, which is a crucial consideration, especially in safety-critical systems. However, due to the computational intractability of exact inference for posteriors, approximate methods are commonly employed in practice.

Background

ICML International Conference On Machine Learning

Variational Bayesian Neural Network (Variational BNN): Approximate neural network weights posterior based on variational inference method.

Mean-field variational inference (MFVI) stands out as a powerful paradigm for approximating the Bayesian posterior with flexible variational distributions.

variational distributions :

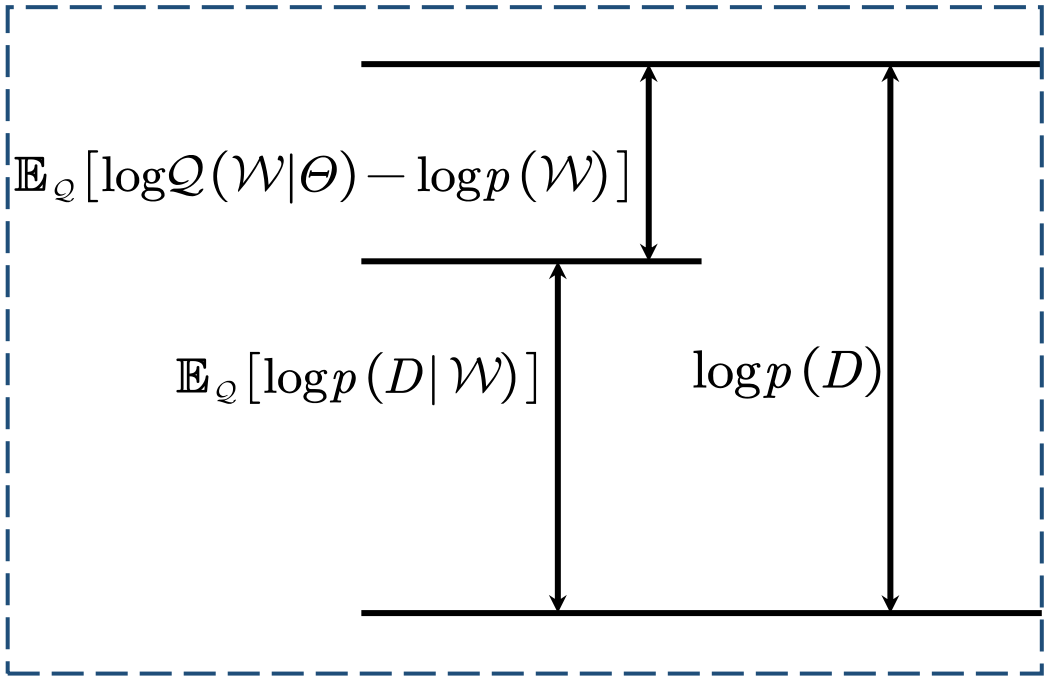
$$\mathcal{Q}(\mathcal{W}|\Theta)=\prod_l q(w_l|\theta_l),\theta_l=[\mu_l,\sigma_l^2]$$

Loss Function:

$$\mathcal{L}(D,\Theta)=\mathbb{E}_{\mathcal{Q}}[\log\mathcal{Q}(\mathcal{W}|\Theta)-\log p(\mathcal{W})]-\mathbb{E}_{\mathcal{Q}}[\log p(D|\mathcal{W})]$$

Monte Carlo Sampling:

$$\mathbb{E}_{\mathcal{Q}}[\log p(D|\mathcal{W})]\approx\frac{1}{M}\sum_{m=1}^{M}p(D|\mathcal{W}^{(m)})$$



HARBIN ENGINEERING UNIVERSITY

Background

The variance of the back-propagating will become higher as the number of layers increases.

The reparameterization trick transforms the sampling procedure that generates weights w_l from $q(w_l|\theta_l)$ as a differentiable mapping t .

$$w_l = t(\epsilon;\theta_l) = \mu_l + \sigma_l^2 \odot \epsilon, \epsilon \sim N(0, I)$$

For a single weight sampling, the gradient of variational parameters can be calculated by:

$$\nabla_{\theta_l} \triangleq -\underbrace{\nabla_{\theta_l}^{lg}[\log p(D|\mathcal{W})]}_{likelihood\ gradient} + \underbrace{\nabla_{\theta_l}^{rg}[\log(q(w_l|\theta_l) - p(w_l))]}_{regularizer\ gradient}$$

Regularizer gradient

$$\nabla_{\mu_l}^{rg}[\log q(w_l|\theta_l) - p(w_l)] = \epsilon$$

$$\nabla_{\sigma_l}^{rg}[\log q(w_l|\theta_l) - p(w_l)] \propto \epsilon^2$$

Likelihood gradient

$$\nabla_{\mu_l}^{lg}[\log p(D|\mathcal{W})] = \frac{\partial \log p(\mathcal{D}|\mathcal{W})}{\partial w_l}$$

$$\nabla_{\sigma_l}^{lg}[\log p(D|\mathcal{W})] \propto \frac{\partial \log p(\mathcal{D}|\mathcal{W})}{\partial w_l}\epsilon$$

Moment Propagation in Mean-field Variational Inference

ICML
International Conference
On Machine Learning

In a feed-forward BNN, the activation action of l-th layer can be expressed as follows:

$$z_l = \delta(w_l z_{l-1}), w_l \sim q_l(w_l|\theta_l)$$

Monte Carlo sampling estimation of first-order moments:

$$\tilde{z}_l \approx \frac{1}{M}\sum_{m=1}^{M} \delta(w_l^m \tilde{z}_{l-1}), w_l^m \sim q_l(w_l|\theta_l)$$

Importance sampling estimation of first-order moments:

$$\tilde{z}_l = \mathbb{E}_{r_i}\left[\frac{q(w_l|\theta_l)}{r_l(w_l)}\delta(w_l \tilde{z}_{l-1})\right]$$

$$\approx \frac{1}{M}\sum_{m=1}^{M}\frac{q(w_l^m|\theta_l)}{r_l(w_l^m)}\delta(w_l^m \tilde{z}_{l-1}), w_l^m \sim r_l(w_l)$$

Proposition 1. We can obtain an optional optimal proposal distribution for minimizing the variance of first-order moments

Proof. As: For $r_l = r_l^*$

$$\mathbb{E}_{r_l}[f(w_l)^2\gamma_l^2] \geqslant [\mathbb{E}_{r_l}|f(w_l)\gamma_l|]^2 = \left(\int |f(w_l)|q_l(w_l)\,dw_l\right)^2$$

$$\mathbb{E}_{r_l^*}[f(w_l)^2\gamma_l^2] = \int \frac{f(w_l)^2 q(w_l)^2}{|f(w_l)|q(w_l)} dw_l \int |f(w_l)|q(w_l)dw_l = \left(\int |f(w_l)|q(w_l)\,dw_l\right)^2$$

Proposition 1

$$r_l^*(w_l) \propto |f(w_l)|q_l(w_l)$$

Approximating the Optimal Proposal Distribution

ICML
International Conference
On Machine Learning

Referring to the definitions of mean and variance, we can derive:

$$\overset{*}{\mu} = \int q(w_l|\theta_l)\,|f(w_l)|\,w_l\,dw_l$$
$$= \mathbb{E}_{q_l}[|f(w_l)|w_l]$$

$$\overset{*}{\sigma}^2 = \mathbb{E}_{q_l}[(|f(w_l)|w_l)^2]$$
$$-\left[\mathbb{E}_{q_l}[|f(w_l)|w_l]\right]^2$$

To simplify notation, we define all the non-linear functions related to w_l as $h(w_l)$.

Taylor Expand:

$$\tilde{h}(w_l) = h(\mu_l) + h'(\mu_l)(w_l - \mu_l) + \frac{h''(\mu_l)}{2}(w_l - \mu_l)^2$$

Reparameterization Trick:

$$\tilde{h}_{\theta_l}(\epsilon) = h(\mu_l) + h'(\mu_l)(t(\epsilon;\theta_l) - \mu_l) + \frac{h''(\mu_l)}{2}(t(\epsilon;\theta_l) - \mu_l)^2$$
$$= h(\mu_l) + h'(\mu_l)\sigma_l^2 \odot \epsilon + \frac{h''(\mu_l)}{2}(\sigma_l^2 \odot \epsilon)^2$$

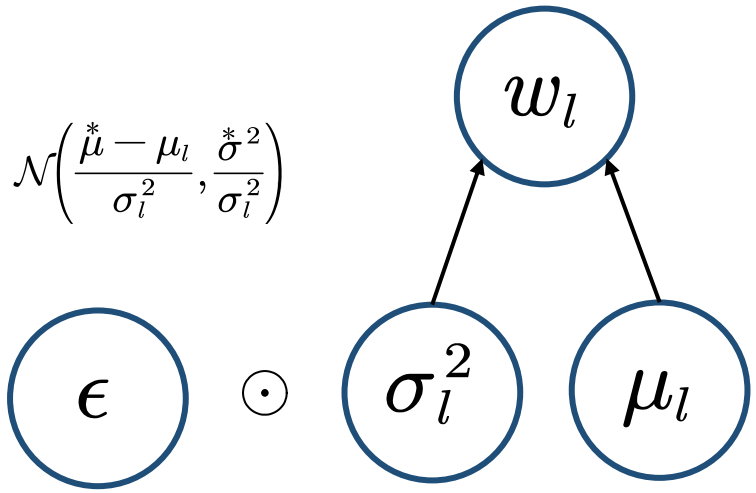
Approximate Expectations:

$$\mathbb{E}_{q_l}\left[\tilde{h}(w_l)\right] = h(\mu_l) + \frac{h''(\mu_l)}{2}(\sigma_l^2)^2$$

Defining Distribution on Exogenous Randomness

ICML
International Conference
On Machine Learning

$$w_l = \mu_l + \sigma_l^2 \odot \epsilon, \epsilon \sim \mathcal{N}\left(\frac{\overset{*}{\mu} - \mu_l}{\sigma_l^2}, \frac{\overset{*}{\sigma}^2}{\sigma_l^2}\right)$$



Proof:

$$\begin{aligned}\mathbb{E}\left[\mu_l + \sigma_l^2 \odot \epsilon\right] &= \mu_l + \sigma^2 \mathbb{E}\left(\epsilon\right) \\ &= \mu_l + \sigma^2 \frac{\overset{*}{\mu} - \mu}{\sigma^2} \\ &= \overset{*}{\mu} = \mathbb{E}_{r(w)}\left[w\right]\end{aligned}$$

$$\begin{aligned}var\left[\mu_l + \sigma_l^2 \odot \epsilon\right] &= var\left[\sigma_l^2 \odot \epsilon\right] \\ &= \mathbb{E}\left[\left(\sigma_l^2 \odot \epsilon\right)^2\right] - \left(\mathbb{E}\left[\sigma_l^2 \odot \epsilon\right]\right)^2 \\ &= \sigma_l^4 \left(\frac{\overset{*}{\mu} - \mu}{\sigma_l^2}\right)^2 + \sigma_l^4 \left(\frac{\overset{*}{\sigma}^2}{\sigma_l^2}\right)^2 - \left(\overset{*}{\mu} - \mu\right)^2 \\ &= \overset{*}{\sigma}^2 = var_{r(w)}\left[w\right]\end{aligned}$$

Algorithm 1 The first moment propagation in l-th via Reparameterized Importcance Sampling

1: Variational posterior $q_l(w_l)$ parameters $\theta_l = (\mu_l, \sigma_l^2)$, the first moment of last layer $\tilde{z}_{l-1}$.
2: Approximate the optimal proposal distribution $r_l(w_l|\overset{*}{\mu}_l, \overset{*}{\sigma}_l^2)$:
- Using Eq. (15) in Eq. (11) to calculate the mean $\overset{*}{\mu}_l$ by set $h(w_l) = |f(w_l)|w_l$
- UsingEq. (15) in Eq. (12) to calculate the variance $\overset{*}{\sigma}_l^2$ by set $h(w_l) = (|f(w_l)|w_l)^2$

3: **for** $m = 1$ to M **do**
4: Sample $\epsilon^m \sim \mathcal{N}\left(\frac{\overset{*}{\mu}_l - \mu_l}{\sigma_l^2}, \frac{\overset{*}{\sigma}_l^2}{\sigma_l^2}\right)$.
5: Let $w^m = \mu_l + \sigma_l^2 \odot \epsilon^m$
6: Calculate $\gamma_l^m = \frac{q(w_l^m|\theta_l)}{r_l(w_l^m)} \approx \frac{1}{|f(w_l)|}$
7: Calculate $\delta\left(w_l^m \tilde{z}_{l-1}\right)$
8: **end for**
9: Calculate $\tilde{z}_l \approx \frac{1}{M} \sum_{m=1}^{M} \gamma_l^m \delta\left(w_l^m \tilde{z}_{l-1}\right)$

Algorithm 1 entails a detailed description of how the first moment propagates across each layer.

HARBIN ENGINEERING UNIVERSITY

Results

ICML International Conference On Machine Learning

1. Comparison with Baseline MFVI

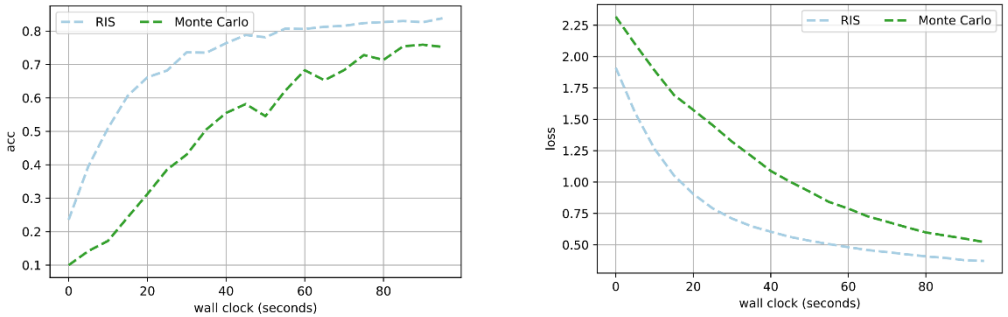


Figure 1. Optimization trace of applying MFVI to approximate the posterior of the neural network. We run the standard Monte Carlo sampling (green line) and the RIS (light blue line) with 10 samples

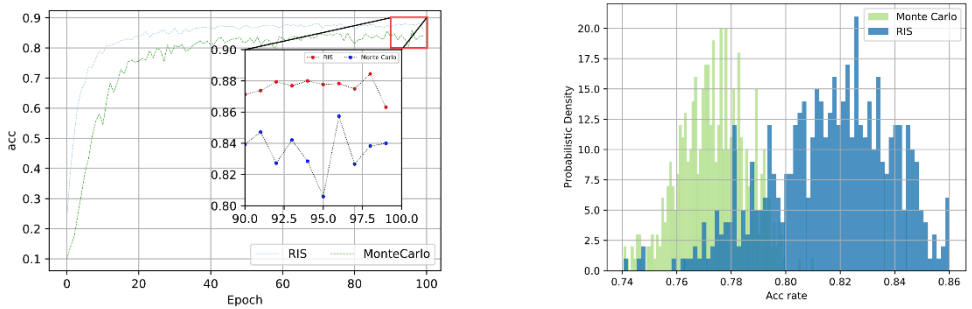


Figure 2. Accuracy rates of Bayesian ResNet20 trained with using RIS and with Monte Carlo samplings sampling during a 100-epoch training. Accuracy comparison of 1000 models sampled from Bayesian neural networks using RIS and Monte Carlo sampling.

2. Compare with SOTA Methods

Dataset	model	Method	Accuracy
CIFAR-10	ResNet-20	MFVI	22.46 ± 0.70
CIFAR-10	ResNet-20	MFVI(tempered))	83.56 ± 0.45
CIFAR-10	ResNet-20	SWAG	86.80 ± 0.10
CIFAR-10	ResNet-20	VOGN	85.36 ± 0.25
CIFAR-10	ResNet-20	GLM	84.35 ± 0.18
CIFAR-10	ResNet-20	Adversarial Sampling	86.33 ± 0.45
CIFAR-10	ResNet-20	RIS	**87.37 ± 0.26**
CIFAR-10	ResNet-20	HMC	**90.02 ± 0.26**

Table 1. Comparison of classification accuracies for ResNet-20 trained with the proposed method and other SOTA methods on the CIFAR-10 dataset.

HARBIN ENGINEERING UNIVERSITY

ICML
International Conference
On Machine Learning

Thanks

HARBIN
ENGINEERING
UNIVERSITY