Bayesian Neural Network via Stochastic Gradient Descent

Abhinav Sagar*
Vellore Institute of Technology
Vellore, Tamil Nadu, India
abhinavsagar4@gmail.com

Abstract

The goal of bayesian approach used in variational inference is to minimize the KL divergence between the variational distribution and unknown posterior distribution. This is done by maximizing the Evidence Lower Bound (ELBO). A neural network is used to parametrize these distributions using Stochastic Gradient Descent. This work derives the variational inference models using bayesian neural networks. We show how SGD can be applied on bayesian neural networks by gradient estimation techniques. For validation, we tested our model on 5 publicly available UCI datasets and the metrics chosen for evaluation are Root Mean Square Error (RMSE) error and negative log likelihood. Our work considerably beats the previous state of the art approaches for regression using SGD.

1 Introduction

Recently, there has been a lot of work done on inference using probabilistic models. In this approach, rather than considering the parameters of the neural network as point estimates, we sample them as continuous distributions. Using this approach, helps us infer the uncertainty involved while making the predictions. This is very important in sensitive domains where not only we want the predictions made by the model but also with how much certainty it is making those predictions.

The problem with this approach lies in the calculation of posterior distribution which is often intractable. Hence for the computation, it is necessary to convert the variational distribution into a tractable posterior distribution. Variational inference is used to convert the inference problem into an optimization problem with the objective of minimizing the KL-divergence between variational distribution and the true posterior. This is done by maximizing the ELBO.

In this paper, we present a new technique of training Bayesian Neural Network using stochastic gradient descent. We show how distributions can be parameterized by using variational inference techniques. We validate our work on UCI datasets and show our approach outperforms previous state of the art in this domain.

We summarize our main contributions as follows:

• An approach to train Bayesian Neural Network using stochastic gradient descent.

• A theoretical analysis of our approach which uses an alternative lower bound backed by variational inference techniques.

• Evaluation on the UCI dataset using test RMSE and log likelihood as the evaluation metrics shows we outperform all previous state-of-the-art methods on regression datasets.

*Website of author - https://abhinavsagar.github.io/

Preprint. Under review.

2 Background

2.1 Variational Inference and the ELBO

A probabilistic model is denoted using observations x, latent variables z and model parameters θ. The optimal θ value has to be found to maximize the marginal likelihood as defined in Equation 1 where we refer $p\theta(x|z)$ as the generative distribution and $p\theta(z)$ as the prior distribution.

$$p_\theta(\mathbf{x}) = \int p_\theta(\mathbf{x}|\mathbf{z})p_\theta(\mathbf{z})\mathrm{d}\mathbf{z} \tag{1}$$

Computing the posterior by doing inference is intractable as it requires doing an integration over z. Hence we maximize Evidence Lower Bound (ELBO) in variational inference which can be computed by approximating posterior on the latent variable as defined in Equation 2:

$$\begin{aligned}\ln p_\theta(\mathbf{x}) &= \ln \int p_\theta(\mathbf{x}|\mathbf{z})p_\theta(\mathbf{z})\mathrm{d}\mathbf{z} \\ &= \ln \int \frac{q_\phi(\mathbf{z}|\mathbf{x})}{q_\phi(\mathbf{z}|\mathbf{x})} p_\theta(\mathbf{x}|\mathbf{z})p_\theta(\mathbf{z})\mathrm{d}\mathbf{z} \\ &= \ln \int q_\phi(\mathbf{z}|\mathbf{x}) \frac{p_\theta(\mathbf{x}|\mathbf{z})p_\theta(\mathbf{z})}{q_\phi(\mathbf{z}|\mathbf{x})}\mathrm{d}\mathbf{z}\end{aligned} \tag{2}$$

2.2 Stochastic Gradient Descent

We can optimize the variational parameters ϕ by using stochastic gradient descent with the following update rule as defined in Equation 3. Here γ is learning rate and η is the size of randomly sampled mini batches of training data.

$$\phi_{t+1} = \phi_t - \gamma \sum_{i=1} \nabla_\phi \mathcal{L}\left(\mathbf{x}^{(i)}\right) \tag{3}$$

We define $ESBOS()$ as an alternative lower bound having its posterior distribution unknown. Hence we use joint distribution instead. The weights are normalized to tackle the intractable integral present in the denominator. The dataset has been divided into mini-batches and the operations are carried on them in successive iterations. The $ESBOS()$ is defined in Equation 4:

$$\hat{\nabla}_\lambda \mathcal{S} = \sum_i^M \hat{w}^{(i)} \left[\left(\log \omega\left(\theta^{(i)}\right) + 1\right) \nabla_\lambda \log p\left(\mathcal{D}, \theta^{(i)}\right) - \nabla_\lambda \log q\left(\theta^{(i)}; \lambda\right)\right] \tag{4}$$

Using the reparameterization trick, the gradient can be written as shown in Equation 5:

$$\hat{\nabla}_\lambda \mathcal{S} = \sum_i^M \hat{w}^{(i)} \left[\left(\log \omega\left(\theta^{(i)}\right) + 1\right) \nabla_\lambda \log p\left(\mathcal{D}, g_\lambda\left(\epsilon^{(i)}\right)\right) - \nabla_\lambda \log q\left(g_\lambda\left(\epsilon^{(i)}\right); \lambda\right)\right] \tag{5}$$

This operation is very helpful in reducing the complexity of variational inference models. Now Stochastic Gradient Descent (SGD) can be applied to minimize the $ESBOS()$. This operation is shown in Equation 6:

$$\hat{\mathcal{S}}^* = \sum_{i=1}^M \hat{w}^{(i)} \left[\sum_{n=1}^N \log p\left(x_n|\theta^{(i)}\right) + \log p\left(\theta^{(i)}\right) - \log q\left(\theta^{(i)}; \lambda^*\right)\right] \tag{6}$$

2.3 Algorithm

Next we present the algorithm used in this work:

Algorithm 1: Bayesian Neural Network via Stochastic Gradient Descent

---

Initialize hyperparameter λ_0 and the learning rate α_0 using arbitrary values
while *not converged* **do**

  generate M samples $\left\{\theta^{(i)}\right\}_{i=1}^{M} : \epsilon^{(i)} \sim \mathcal{N}(0,1), \theta^{(i)} = g_\lambda\left(\epsilon^{(i)} = \mu + \sigma\epsilon^{(i)}\right.$
  calculate the weight $\log w^{(i)} = \frac{N}{S}\sum_{n=1}^{S} \log p\left(x_n|\theta^{(i)}\right) + \log p\left(\theta^{(i)}\right) - \log q\left(\theta^{(i)}; \lambda_t\right)$
  evaluate the gradient $\hat{\nabla}_\lambda \mathcal{S} = -\sum \hat{w}^{(i)} \nabla_\lambda \log q\left(g_\lambda\left(\epsilon^{(i)}\right)\right.$
  $w^{(i)} = \exp\left(\log w^{(i)} + \min\left\{\log w^{(i)}\right\}\right)$
  update $\lambda_{t+1} = \lambda_t - \alpha_t * \nabla_\lambda \mathcal{S}$

end

---

3 Simulation Studies

For bayesian neural network regression, we used datasets from UCI repository: Boston, Concrete, Energy, Protein, Wine. We used a neural network with one hidden layer with 50 neurons in each case. We set (0, 1) as the prior distribution for the weight and bias of the neural network, ReLU is used as the activation function and batch size value of 32. The datasets are randomly partitioned into 90 percent with training data and 10 percent for testing, and the results are averaged over 50 random trials. The average RMSE loss and average log likelihood values obtained are shown in Table 1 and Table 2 respectively.

Table 1: Bayesian neural network regression: average test RMSE(lower is better)

Dataset	Ours	(Chen et al., 2018a)	(Li and Turner, 2016b)	(Hoffman et al., 2013)
Boston	2.58±0.13	2.71±0.29	2.89±0.17	2.977±0.093
Concrete	4.79±0.36	5.04±0.27	5.42±0.11	5.506±0.103
Energy	0.74±0.08	0.95±0.15	0.51±0.01	1.734±0.051
Protein	4.38±0.07	4.43±0.05	4.45±0.02	4.623±0.009
Wine	0.59±0.04	0.61±0.03	0.63±0.01	0.614±0.008

Table 2: Bayesian neural network regression: average negative test LL(lower is better)

Dataset	Ours	(Chen et al., 2018a)	(Li and Turner, 2016b)	(Hoffman et al., 2013)
Boston	2.36±0.17	2.40±0.09	2.52±0.03	2.579±0.042
Concrete	2.93±0.07	3.02±0.04	3.11±0.02	3.137±0.021
Energy	1.53±0.04	1.65±0.04	0.77±0.02	1.981±0.024
Protein	2.85±0.02	2.89±0.01	2.91±0.00	2.950±0.002
Wine	0.92±0.03	0.93±0.04	0.96±0.01	0.931±0.014

Our method achieves better results on both the metrics compared to the previous state of the art methods.

4 Conclusions

In this work, we showed how a Bayesian neural network can be trained using stochastic gradient descent. We outlined the problem in the context of variational inference and how it can be converted into a tractable one using ELBO. We presented our algorithm which used gradient estimation techniques for computing the inference. We evaluate our work on UCI datasets and our network outperforms previous state of the art on regression benchmarks.

Acknowledgments

We would like to thank Nvidia for providing the GPUs.

References

A. Alemi, B. Poole, I. Fischer, J. Dillon, R. A. Saurous, and K. Murphy. Fixing a broken elbo. In *International Conference on Machine Learning*, pages 159–168, 2018.

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

L. Bottou. Stochastic gradient descent tricks. In *Neural networks: Tricks of the trade*, pages 421–436. Springer, 2012.

P. Chaudhari and S. Soatto. Stochastic gradient descent performs variational inference, converges to limit cycles for deep networks. In *2018 Information Theory and Applications Workshop (ITA)*, pages 1–10. IEEE, 2018.

L. Chen, C. Tao, R. Zhang, R. Henao, and L. C. Duke. Variational inference and model selection with generalized evidence bounds. In *International conference on machine learning*, pages 893–902, 2018a.

X. Chen, S. Liu, R. Sun, and M. Hong. On the convergence of a class of adam-type algorithms for non-convex optimization. *arXiv preprint arXiv:1808.02941*, 2018b.

J. Domke and D. R. Sheldon. Importance weighting and variational inference. In *Advances in neural information processing systems*, pages 4470–4479, 2018.

H. Duan, L. Yang, J. Fang, and H. Li. Fast inverse-free sparse bayesian learning via relaxed evidence lower bound maximization. *IEEE Signal Processing Letters*, 24(6):774–778, 2017.

D. Flam-Shepherd, J. Requeima, and D. Duvenaud. Mapping gaussian process priors to bayesian neural networks. In *NIPS Bayesian deep learning workshop*, 2017.

M. Hardt, B. Recht, and Y. Singer. Train faster, generalize better: Stability of stochastic gradient descent. In *International Conference on Machine Learning*, pages 1225–1234, 2016.

J. M. Hernández-Lobato and R. Adams. Probabilistic backpropagation for scalable learning of bayesian neural networks. In *International Conference on Machine Learning*, pages 1861–1869, 2015.

M. D. Hoffman and M. J. Johnson. Elbo surgery: yet another way to carve up the variational evidence lower bound. In *Workshop in Advances in Approximate Bayesian Inference, NIPS*, volume 1, page 2, 2016.

M. D. Hoffman, D. M. Blei, C. Wang, and J. Paisley. Stochastic variational inference. *The Journal of Machine Learning Research*, 14(1):1303–1347, 2013.

D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

D. P. Kingma and M. Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

A. Kucukelbir, D. Tran, R. Ranganath, A. Gelman, and D. M. Blei. Automatic differentiation variational inference. *The Journal of Machine Learning Research*, 18(1):430–474, 2017.

Y. Li and R. E. Turner. Rényi divergence variational inference. In *Advances in Neural Information Processing Systems*, pages 1073–1081, 2016a.

Y. Li and R. E. Turner. Variational inference with rényi divergence. *stat*, 1050:6, 2016b.

Q. Liu and D. Wang. Stein variational gradient descent: A general purpose bayesian inference algorithm. In *Advances in neural information processing systems*, pages 2378–2386, 2016.

I. Loshchilov and F. Hutter. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*, 2016.

A. Malinin and M. Gales. Reverse kl-divergence training of prior networks: Improved uncertainty and adversarial robustness. In *Advances in Neural Information Processing Systems*, pages 14547–14558, 2019.

V. Mullachery, A. Khera, and A. Husain. Bayesian neural networks. *arXiv preprint arXiv:1801.07710*, 2018.

K. P. Murphy. *Machine learning: a probabilistic perspective*. MIT press, 2012.

G. Roeder, Y. Wu, and D. K. Duvenaud. Sticking the landing: Simple, lower-variance gradient estimators for variational inference. In *Advances in Neural Information Processing Systems*, pages 6925–6934, 2017.

J. Yao, W. Pan, S. Ghosh, and F. Doshi-Velez. Quality of uncertainty quantification for bayesian neural network inference. *arXiv preprint arXiv:1906.09686*, 2019.

Y. Yao, A. Vehtari, D. Simpson, and A. Gelman. Yes, but did it work?: Evaluating variational inference. *arXiv preprint arXiv:1802.02538*, 2018.

C. Zhang, J. Bütepage, H. Kjellström, and S. Mandt. Advances in variational inference. *IEEE transactions on pattern analysis and machine intelligence*, 41(8):2008–2026, 2018.