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 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 extends the work done by others by deriving the variational inference models. We show how SGD can be applied on bayesian neural networks by gradient estimation techniques. For validation, we have tested our model on 5 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 bayesian neural networks.

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 to find out the predictions made by the model but also with how much certainty it is making the 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 approach is used to convert the inference problem into an optimization problem with the objective of minimizing the KL-divergence between variational distribution and 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. Then we show how distributions can be parameterized by using variational inference techniques. We validated our work on UCI datasets and show our approach is better than the 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 given in Equation 1 where we refer to $p\theta(x|z)$ as the generative distribution and to $p\theta(z)$ 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 shown 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 ascent with the following update rule as given 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 ESBO S()having its posterior unknown. Hence we use joint distribution instead. The weights are normalized to remove the intractable integration present in the denominator. The dataset has been divided into mini-batches and the operations are carried on them in successive iterations. The ESBO 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 Fig 4. This operation is very helpful in reducing the complexity of variational inference models as shown in Equation 5.

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

Now Stochastic Gradient Descent (SGD) can be applied to minimize the ESBO. 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 $\lambda_0$ and the learning rate $\alpha_0$ 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 have used datasets from UCI repository: Boston, Concrete, Energy, Protein, Wine. We have 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 as the activation function and batch size as 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 are given in Table 1 and Table 2 respectively. Our method archives much better results on both the above metrics compared to the previous state of the art.

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

Dataset	Ours	Rényi-VI[6]	CLBO-VI [6]	ELBO-VI[12]	BPB[4]
Boston	2.58±0.13	2.86±0.40	2.71±0.29	2.89±0.17	2.977±0.093
Concrete	4.79±0.36	5.15±0.25	5.04±0.27	5.42±0.11	5.506±0.103
Energy	0.74±0.08	1.00±0.18	0.95±0.15	0.51±0.01	1.734±0.051
Protein	4.38±0.07	4.65±0.07	4.43±0.05	4.45±0.02	4.623±0.009
Wine	0.59±0.04	0.62±0.03	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	Rényi-VI[6]	CLBO-VI [6]	ELBO-VI[12]	BPB[4]
Boston	2.36±0.17	2.46±0.16	2.40±0.09	2.52±0.03	2.579±0.042
Concrete	2.93±0.07	3.04±0.07	3.02±0.04	3.11±0.02	3.137±0.021
Energy	1.53±0.04	1.67±0.05	1.65±0.04	0.77±0.02	1.981±0.024
Protein	2.85±0.02	2.93±0.00	2.89±0.01	2.91±0.00	2.950±0.002
Wine	0.92±0.03	0.94±0.04	0.93±0.04	0.96±0.01	0.931±0.014

4 Conclusions

In this work, we showed how a Bayesian neural network can be trained using stochastic gradient descent. We started with presenting the problem in variational inference approaches and how to convert the problem into a tractable one using ELBO. Next we presented our algorithm which used gradient estimation techniques for doing the inference. We evaluated our work on UCI datasets and this method produces better results than the previous state of the art.

Acknowledgments

We would like to thank Nvidia for providing the GPUs.

References

1. Adji B. Dieng, Dustin Tran, Rajesh Ranganath, John Paisley, and David M. Blei. Variational inference via upper bound minimization. In Proceedings of the Neural Information Processing Systems, 2017.

2. Justin Domke and Daniel Sheldon. Importance weighting and variational inference. In Proceedings of the Neural Information Processing Systems, 2018.

3. José Miguel Hernández-Lobato and Ryan P. Adams. Probabilistic backpropagation for scalable learning of Bayesian neural networks. In Proceedings of the International Conference on Machine Learning, 2015.

4. Matthew D. Hoffman, David M. Blei, Chong Wang, and John William Paisley. Stochastic variational inference. Journal of Machine Learning Research.

5. Cheng Zhang, Judith Butepage, Hedvig Kjellstrom, and Stephan Mandt. Advances in variational inference. ArXiv preprint, arXiv:1711.05597, 2018.

6. Chenyang Tao, Liqun Chen, Ruiyi Zhang, Ricardo Henao, and Lawrence Carin. Variational inference and model selection with generalized evidence bounds. In Proceedings of the International Conference on Machine Learning, 2018.

7. K. P. Murphy. Machine Learning: A Probabilistic Perspective. MIT press, 2012.

8. Y. Li and R. E. Turner. Variational inference with Rényi divergence. In Proceedings of the Neural Information Processing Systems, 2016.

9. D. Kingma and M. Welling. Auto-encoding variational Bayes. In Proceedings of the 2nd International Conference on Learning Representations (ICLR), 2014.

10. Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. arXiv preprint, arXiv:1412.698, 2014.

11. José Miguel Hernández-Lobato and Ryan P. Adams. Probabilistic backpropagation for scalable learning of Bayesian neural networks. In Proceedings of the International Conference on Machine Learning, 2015.

12. Y. Li and R. E. Turner. Variational inference with Rényi divergence. In Proceedings of the Neural Information Processing Systems, 2016.

13. Chenyang Tao, Liqun Chen, Ruiyi Zhang, Ricardo Henao, and Lawrence Carin. Variational inference and model selection with generalized evidence bounds. In Proceedings of the International Conference on Machine Learning, 2018.

14. Flam-Shepherd, Daniel, James Requeima, and David Duvenaud. "Mapping Gaussian process priors to Bayesian neural networks." NIPS Bayesian deep learning workshop. 2017.

15. Yao, Jiayu, et al. "Quality of uncertainty quantification for Bayesian neural network inference." arXiv preprint arXiv:1906.09686 (2019).

16. Mullachery, Vikram, Aniruddh Khera, and Amir Husain. "Bayesian neural networks." arXiv preprint arXiv:1801.07710 (2018).

17. Hoffman, Matthew D., and Matthew J. Johnson. "Elbo surgery: yet another way to carve up the variational evidence lower bound." Workshop in Advances in Approximate Bayesian Inference, NIPS. Vol. 1. 2016.

18. Duan, Huiping, et al. "Fast inverse-free sparse bayesian learning via relaxed evidence lower bound maximization." IEEE Signal Processing Letters 24.6 (2017): 774-778.

19. Alemi, Alexander A., et al. "Fixing a broken ELBO." arXiv preprint arXiv:1711.00464 (2017).

20. Blei, David M., Alp Kucukelbir, and Jon D. McAuliffe. "Variational inference: A review for statisticians." Journal of the American statistical Association 112.518 (2017): 859-877.

21. Kucukelbir, Alp, et al. "Automatic differentiation variational inference." The Journal of Machine Learning Research 18.1 (2017): 430-474.

22. Zhang, Cheng, et al. "Advances in variational inference." IEEE transactions on pattern analysis and machine intelligence 41.8 (2018): 2008-2026.

23. Li, Yingzhen, and Richard E. Turner. "Rényi divergence variational inference." Advances in Neural Information Processing Systems. 2016.

24. Liu, Qiang, and Dilin Wang. "Stein variational gradient descent: A general purpose bayesian inference algorithm." Advances in neural information processing systems. 2016.

25. Chaudhari, Pratik, and Stefano Soatto. "Stochastic gradient descent performs variational inference, converges to limit cycles for deep networks." 2018 Information Theory and Applications Workshop (ITA). IEEE, 2018.

26. Roeder, Geoffrey, Yuhuai Wu, and David K. Duvenaud. "Sticking the landing: Simple, lower-variance gradient estimators for variational inference." Advances in Neural Information Processing Systems. 2017.

27. Yao, Yuling, et al. "Yes, but did it work?: Evaluating variational inference." arXiv preprint arXiv:1802.02538 (2018).

28. Malinin, Andrey, and Mark Gales. "Reverse kl-divergence training of prior networks: Improved uncertainty and adversarial robustness." Advances in Neural Information Processing Systems. 2019.

29. Chen, Xiangyi, et al. "On the convergence of a class of adam-type algorithms for non-convex optimization." arXiv preprint arXiv:1808.02941 (2018).

30. Kingma, Diederik P., and Jimmy Ba. "Adam: A method for stochastic optimization." arXiv preprint arXiv:1412.6980 (2014).

31. Bottou, Léon. "Stochastic gradient descent tricks." Neural networks: Tricks of the trade. Springer, Berlin, Heidelberg, 2012. 421-436.

32. Loshchilov, Ilya, and Frank Hutter. "Sgdr: Stochastic gradient descent with warm restarts." arXiv preprint arXiv:1608.03983 (2016).

33. Hardt, Moritz, Benjamin Recht, and Yoram Singer. "Train faster, generalize better: Stability of stochastic gradient descent." arXiv preprint arXiv:1509.01240 (2015).