Studying Uncertainity in Bayesian Neural Networks

Sharan Yalburgi, Julyan Arbel

July, 2019

1 Abstract

Availability of uncertainity measure is the one of the major reasons why Bayesian Neural Networks are topics of interest. It is therefore essential to understand the behavior of these uncertainity measures with varying prior conditions, amount of data and modelling tasks.

2 Introduction

Contributions of this paper is as follows:

3 Previous Work

3.1 Study of Bayesian Neural Networks with Non-Smooth Functions Imaizumi and Fukumizu, 2018

3.2 Consistency of Posterior Distributions for Neural Networks Lee, 2000

3.3 Classification kwith imperfect training labels Cannings, Fan, and Samworth, 2018

- Study effect of imperfect training labels on performance of classification methods.
- Bound the excess risk of an arbitrary classifier trained with imperfect labels in terms of its excess risk for predicting a noisy labe

3.4 A Theoretical Analysis of Deep Neural Networks for Texture Classification Basu et al., 2016

- Derives the upper bounds on the VC dimension of Convolutional Neural Network as well as Dropout and Dropconnect networks and the relation between excess error rate of Dropout and Dropconnect networks.
- Introduces concept of Intrinsic Dimension to show that texture datasets have a higher dimensionality than color/shape based data.

3.5 Understanding Priors in Bayesian Neural Networks at the Unit Level - Vladimirova et al., 2019

- Studies the "induced" prior distributions at a unit level.
  - I layer units are Gaussian
  - II layer units are sub-exponential
  - Units in deeper layers are characterized by sub-Weibull distributions.
- This paper is devoted to the investigation of hidden units prior distributions in Bayesian neural networks under the assumption of independent Gaussian weights.

3.6 Consistency of posterior distributions for neural networks - Lee, 2000

Considers only feedforward neural networks with a single hidden layer of units with logistic activation functions(not done for popularly used ReLU activations) and linear output unit(no Sigmoid at the end). Shows that posterior probability of feedforward neural networks is **"asymptotically consistent"**. $\hat{g_n}$ is asymptotically consistent for true regression function g_0 if

$$\int (\hat{g_n}(x) - g_0(x))^2 f_0(x) dx \xrightarrow{p} 0 \tag{1}$$

p here means that the probability,
$P(|\int (\hat{g_n}(x) - g_0(x))^2 f_0(x) dx| > \epsilon) \to 0$ as $n \to \infty$.
Asymptotically consistence in terms of Hellinger neighbourhood: If (X_i, Y_i) f_0, the posterior is asymptotically consistence for f_0(true joint distribution over X & Y) over Hellinger neighbourhood if for every $epsilon > 0$

$$P(f : D_H(f, f_0) <= \epsilon | (X_i, Y_i) \forall i) \xrightarrow{p} 1 \tag{2}$$

where Hellinger distance

$$D_H$$

is defined as

$$D_H(f, f_0) = \sqrt{\int \int (\sqrt{f(x,y)} - \sqrt{f_0(x,y)})^2 dx dy}$$

Another approach to consistency: Posterior probability of every neighbourhood of true function tends to 1, i.e, $P(A_\epsilon | Y) \longrightarrow 1$ as $n \to 1$.

Sieve Approach: The number of hidden nodes grows with the sample size so that asymptotically they use arbitrarily large number of nodes.

Parameter Approach: Number of hidden nodes taken as a parameter(not sure if as a hyperparameter) and show that joint posterior is also consistent. Extends earlier results on "universal approximation properties of neural networks to the Bayesian setting".

Shows mathematically that using a neural network to estimate and **continuous or square integrable** function

will be consistent with probability tending to one given enough data. This is when the number of nodes grew with amount of data in a controlled way but $k(n) \to \infty$ as $n \to \infty$ where k is the number of nodes in the network or when the number of nodes is a parameter which can be estimated from the data.

4 Theory

4.1 L2 regularization is equivalent to Gaussian prior

From Figueiredo, 2003, assuming

$$y_n = \beta x_n + \epsilon, \tag{3}$$

where ϵ is Gaussian noise with mean 0 and variance σ^2. This gives rise to a Gaussian likelihood:

$$\prod_{n=1}^{N} \mathcal{N}(y_n|\beta x_n, \sigma^2). \tag{4}$$

Let us regularise parameter β by imposing the Gaussian prior $N(\beta|0, \lambda^1)$, where λ is a strictly positive scalar. Hence, combining the likelihood and the prior we simply have:

$$\prod_{n=1}^{N} \mathcal{N}(y_n|\beta x_n, \sigma^2)\mathcal{N}(\beta|0, \lambda^{-1}). \tag{5}$$

Let us take the logarithm of the above expression. Dropping some constants we get:

$$\sum_{n=1}^{N} -\frac{1}{\sigma^2}(y_n - \beta x_n)^2 - \lambda\beta^2 + \text{const.} \tag{6}$$

Therefore, here λ is sessentially the hyperparameter controlling the extent of regularisation.

We wan to prove

Taking inspiration from Rockova et al., 2018

$$\mathcal{F}(L, p, \lambda) = (f_B^{DL} : argmin_B \|B\|_2) \tag{7}$$

Where

$$B = \{(W_1, a_1), (W_2, a_2), \ldots, (W_L, a_L)\}, \tag{8}$$

and F is set of deep nets with gaussian priors on the parameters.

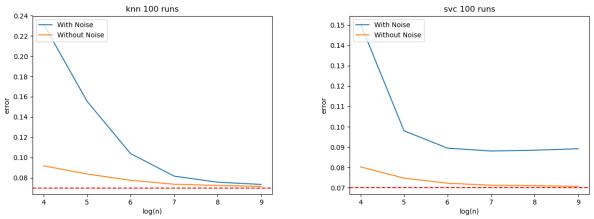
We want to prove

$$\Pi(A|Y^{(n),x_i}{}_{i=1}^{n}(9)$$

5 Experiments

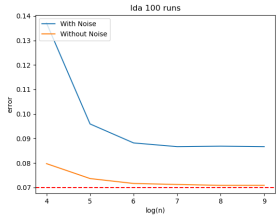
5.1 Results from Cannings, Fan, and Samworth, 2018

- SVMs and Knns are robust/consistent to corrupted/imperfect data.
- Whereas, LDAs are not unless prior probabilities of each class are equal.



(a)

(b)

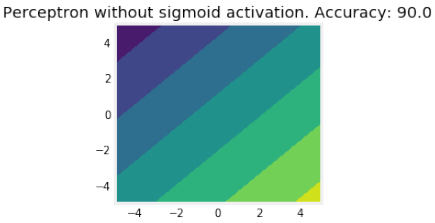


(c)

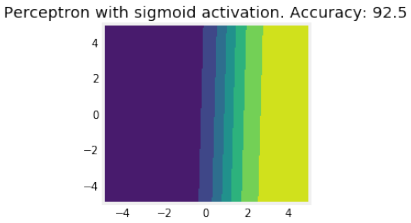
eval5eval5

(d)

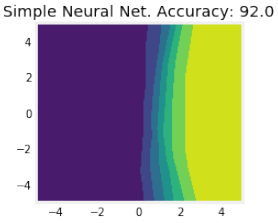
Table 1: Reproduced results from Cannings, Fan, and Samworth, 2018 with data generated from multivariate Gaussian & Perceptron Findings



(a) Prediction contour of Perceptron without Sigmoid activation

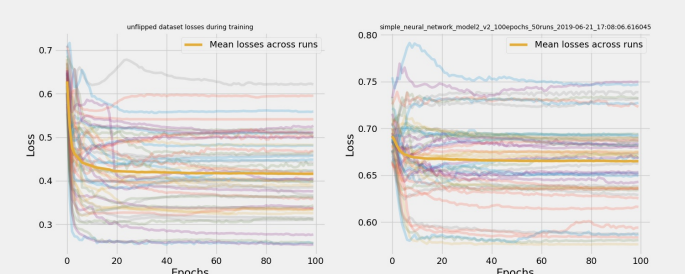


(b) Prediction contour of Perceptron without Sigmoid activation

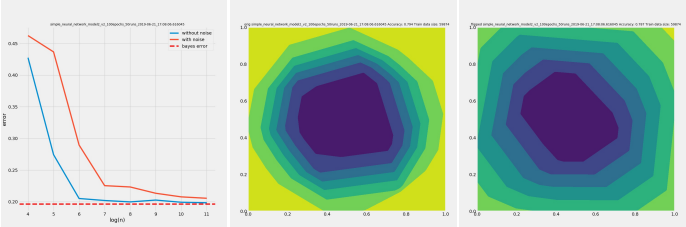


(c) Prediction contour of Single Hidden Layer Neural Network without Sigmoid activation

Figure 1: Analysis of Consistency of Perceptron with noisy data



(a) Training losses of the different instances of the simple neural network



(b) Average error rates of the samples trained neural networks.

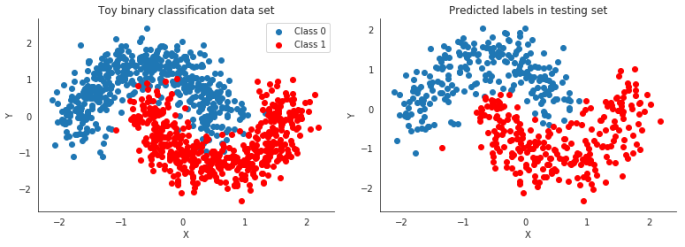
(c) Prediction contour of trained single hidden layer network with original data.(without noise)

(d) Prediction contour of trained neural network with original data.(without noise)

Figure 2: Analysis of Consistency of Neural Networks with noisy data

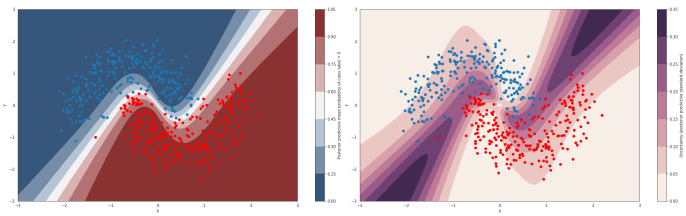
5.2 Consistency in Perceptron and Deep Neural Networks

5.3 Analysis of binary two-moons dataset classification using Bayesian Neural Networks(PyMC3)



(a) Toy Two Moon Dataset

(b) Prediction of Test Set by the Bayesian-NN



(c) Prediction Contour

(d) Uncertainty Contour

Figure 3: Bayesian Neural Network trained to classify Two-Moon Dataset

5.4 Analysis of Bayesian Regression

Analysis of sensitivity of continuous regression using Bayesian Neural Networks with changing σ(numpyro)

We start of our regression analysis my trying fit simple smooth

The results can be found in Figure: 4

Analysis of sensitivity of discontinuous regression using Bayesian Neural Networks with changing σ(numpyro)

Previous work on performance Neural Networks on non-smooth functions has been done in Imaizumi and Fukumizu, 2018. We extend this work by studying the behaviour of Neural Networks in the bayesian paradigm for non somooth regression. One of the aspects which interests us is the quantified uncertainity produced by the BNN. We also investigated its behaviour with changing prior σ (applied to each weight).

Key Takeaways

σ does have a considerable affect on the performance of the BNN. This could be because of excessive constraints on priors with smaller σ when compared to a more relaxed prior for a large σ. This may have lead to a more "regularized" training leading to a "simpler" learned model.

NOTE: Regression training of BNN regardless of type of functions require considerable amount of warmup steps when training using MCMC based algorithms like Hamiltonian Monte Carlo(HMC).

Asymtotic analysis of uncertainity bounds of discontinuous regression using Bayesian Neural Networks(numpyro)

In Fig: 7 we analyse the behaviour of uncertainity bounds with exponentially increasing amount of data. We conjecture that the regression confidence bound of a bayesian neural networks asymtotically (w.r.t. data) converges to true confidence bound.

6 Conclusion

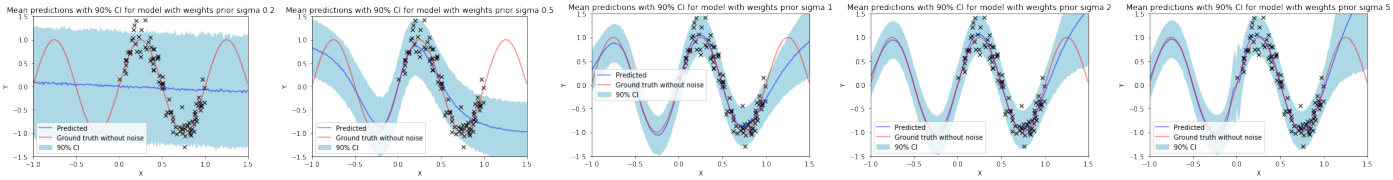
7 Future Work

References

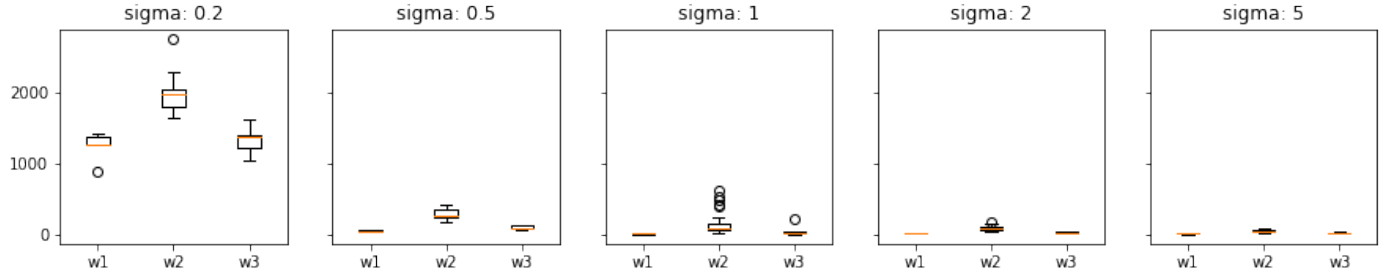
Basu, Saikat et al. (2016). "A Theoretical Analysis of Deep Neural Networks for Texture Classification". In: *CoRR* abs/1605.02699. arXiv: 1605.02699. URL: http://arxiv.org/abs/1605.02699 (cit. on p. 1).

Cannings, Timothy I, Yingying Fan, and Richard J Samworth (2018). "Classification with imperfect training labels". In: *arXiv preprint arXiv:1805.11505* (cit. on pp. 1, 2).

Figueiredo, M. A. T. (Sept. 2003). "Adaptive sparseness for supervised learning". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 25.9, pp. 1150–1159. ISSN: 0162-8828. DOI: 10.1109/TPAMI.2003.1227989 (cit. on p. 2).

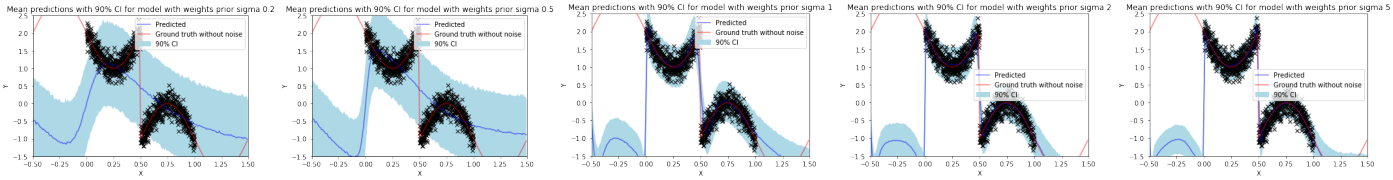


(a) sigma: 0.2 (b) sigma: 0.5 (c) sigma: 1.0 (d) sigma: 2.0 (e) sigma: 5.0

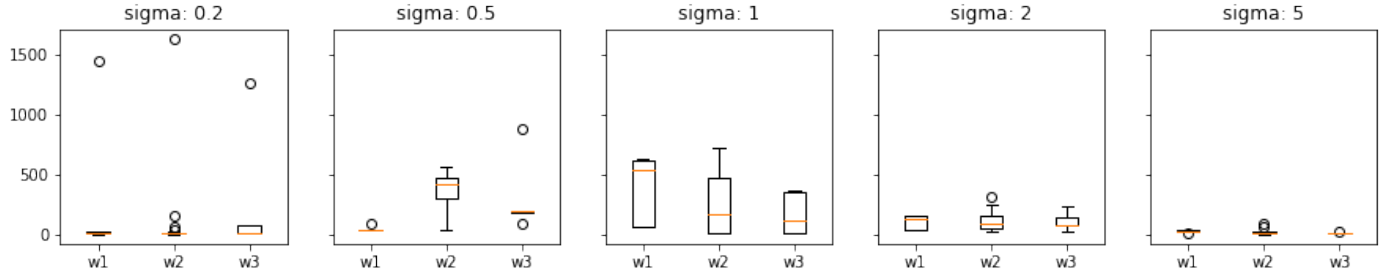


(f) Box plot of effective sample size of weights of each layer

Figure 4: Continuous regression using Bayesian Neural Networks

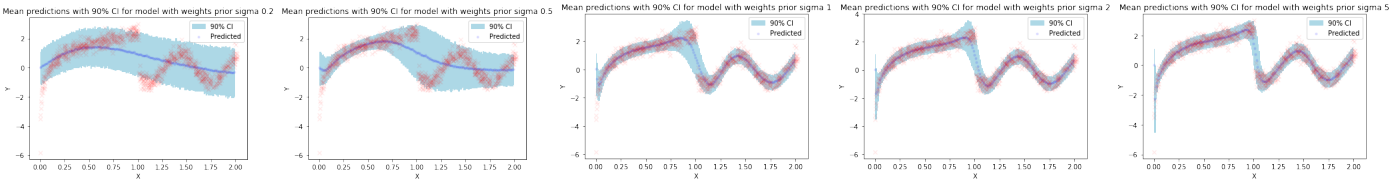


(a) sigma: 0.2 (b) sigma: 0.5 (c) sigma: 1.0 (d) sigma: 2.0 (e) sigma: 5.0

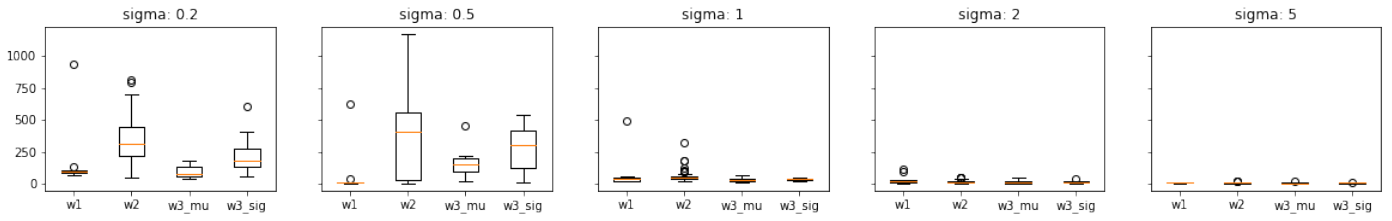


(f) Box plot of effective sample size of weights of each layer

Figure 5: Simple Discontinuous regression using Bayesian Neural Networks

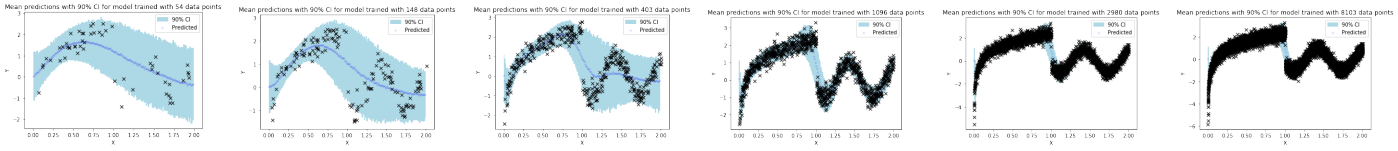


(a) sigma: 0.2 (b) sigma: 0.5 (c) sigma: 1.0 (d) sigma: 2.0 (e) sigma: 5.0



(f) Box plot of effective sample size of weights of each layer

Figure 6: Discontinuous regression using Bayesian Neural Networks



(a) #data : 54 (b) #data : 148 (c) #data : 403 (d) #data : 1096 (e) #data : 2980 (f) #data : 8103

Figure 7: Asymtotic analysis of discontinuous regression using Bayesian Neural Networks

Imaizumi, Masaaki and Kenji Fukumizu (Feb. 2018). "Deep Neural Networks Learn Non-Smooth Functions Effectively". In: *arXiv e-prints*, arXiv:1802.04474, arXiv:1802.04474. arXiv: `1802.04474 [stat.ML]` (cit. on pp. 1, 3).

Lee, Herbert KH (2000). "Consistency of posterior distributions for neural networks". In: *Neural Networks* 13.6, pp. 629–642 (cit. on p. 1).

Rockova, Veronika et al. (2018). "Posterior concentration for sparse deep learning". In: *Advances in Neural Information Processing Systems*, pp. 930–941 (cit. on p. 2).

Vladimirova, Mariia et al. (2019). "Understanding priors in Bayesian neural networks at the unit level". In: *International Conference on Machine Learning*, pp. 6458–6467 (cit. on p. 1).