Jeffreys' prior for layered neural networks

Yoichi Motomura

Electrotechnical Laboratory, AIST, MITI

Information Science Division, Mathematical Informatics Section
Tsukuba, Ibaraki, 305, JAPAN.

E-Mail:motomura@etl.go.jp

ABSTRACT

In this paper, Jeffreys' prior for a neural network is discussed in the framework of the Bayesian statistics. For a good performance of generalization, the regularization methods which reduce both cost function and regularization term are commonly used. In the Bayesian statistics, the regularization term can be naturally derived from prior distribution of parameters. Jeffreys' prior is known as a typical non-informative objective prior. In the case of neural networks, however, it is not easy to express Jeffreys' prior as a simple function of parameters. In this paper, numerical analysis of Jeffreys' prior for neural networks is given. The approximation of Jeffreys' prior is given from a parameter transformation getting to make Jeffreys' prior as a simple function. Some learning techniques are also discussed as applications of these results.

keywords: Jeffreys' prior, bayes, regularization, neural network, generalization, weight decay, initial parameter

1 INTRODUCTION

With finite number of examples, it is very important to get good generalization in the field of training neural networks. Sometimes, since a system has too much flexibility against the number of training examples, it comes a poor generalization of the result of learning. To get better performance of generalization, the flexibility must be reduced during the learning process. In the Bayesian statistics, the measure of flexibility of the system can be considered as probability distribution of the parameter.

The Bayesian approach has been widely applied to neural networks. For example, Tishby, Solla and Levin studied theoretically learning curves by the Bayesian formulations.[1] Mackay applied Bayesian inference to an interpolation problem as probabilistic data fitting from noisy data, and proposed a model comparison method how to determine a weight decay rate.[2]

From Bayesian point of view, the prior distribution restricts the flexibility in the parameter space appropriately, so it could improve generalization ability. What is important here is to give the prior correctly. It is not easy to get a correct prior before learning and whether a prior is appropriate or not depends on the target of learning. Therefore various priors have been proposed for neural networks.[3] If "subjective prior" which depends on each

particular problem is adopted into learning, we should select the best fitted prior to target.[2] On the other hand, "objective prior" is helpful to aim at a universal prior that could be applicable to any problem. In the traditional Bayesian statistics, "non-informative prior" is discussed as "objective prior" and applied when only little knowledge is available at first. Jeffreys' prior[4] is a typical one.

In this paper, Jeffreys' prior for neural networks is discussed and the case of a single parameter is calculated. A simple relation between Jeffreys' prior and the parameter is presented by numerical analysis and approximation. It is turned out that weight parameters of hidden units are more important than other parameters in a sense of the non-informative prior. As applications of this studies, a weight decay term and a method to set the initial condition of weights are also proposed.

2 STATISTICAL VIEW

Layered neural network with one hidden layer is discussed. For the sake of simplicity, we consider only neural networks with one dimensional single input and output unit, and n hidden units. We denote the input value as x and the output value as y.

Thus, the neural network is defined as:

$$f(x;\theta) = \sum_i v_i \frac{1}{1+\exp(-(w_i x + b_i))} + a, \tag{1}$$

where v_i is a weight of the connection from the i-th hidden unit to the output unit, w_i is a weight of the connection from the i-th input unit to the hidden unit, a is a threshold parameter of the output unit, b_i is a threshold parameter of the i-th hidden unit, and θ in the eq.(1) is used to represent all of the variable parameters simply.

From the statistical point of view, neural network learning can be considered as parameter estimation of probability distribution model. In this work we use the following distribution model that is derived from the network defined by the eq.(1).

$$P(y|x,\theta) = \frac{1}{\sqrt{2\pi}\sigma} \exp(-\frac{(y-f(x;\theta))^2}{2\sigma^2}), \tag{2}$$

where σ^2 is a constant that depends on the generated examples.

Let θ^* be the true value of parameter that is unknown. $D = \{(x_i, y_i)|i = 1 \cdots n\}$ is a data set sampled from distribution $P(x, y|\theta^*) = P(y|x, \theta^*)P(x)$. Let $P(D|\theta)$ denote as:

$$P(D|\theta) = \prod_i^n P(x_i, y_i|\theta) = \prod_i^n P(y_i|x_i, \theta)P(x_i). \tag{3}$$

Under the definition above, learning the neural network can be considered as estimating most appropriate value of θ using D.

If θ^* is estimated by the eq.(3), we can use log likelihood as follows:

$$L(\theta) \quad = \quad \log P(D|\theta) = \sum_{i}^{n} \log P(y_i|x_i,\theta) + \sum_{i}^{n} \log P(x_i). \tag{4}$$

Combined with the eq.(2), $L(\theta)$ can be rewritten as:

$$L(\theta) \quad = \quad \sum_{i}^{n} -(y_i - f(x_i;\theta))^2 + C. \tag{5}$$

C is a constant independent of θ.

Therefore for the model of the eq.(3)it is clear that maximizing log likelihood $L(\theta)$ is equivalent to minimizing squared error.

On the other hand, the Bayesian approach,[2,3] can be considered in this context. In this approach, θ is treated as a random variable generated from probability distribution $P(\theta)$.

A posterior probability $P(\theta|D)$ represents a certainty of θ under D , and it is calculated with the Bayes theorem as:

$$P(\theta|D) \quad = \quad \frac{P(D|\theta)P(\theta)}{\int_\theta P(D|\theta)P(\theta)d\theta}. \tag{6}$$

In this approach, θ^* is estimated by maximizing the eq.(6) instead of $L(\theta)$. Then, we can use the following:

$$\begin{aligned} \log P(\theta|D) \quad &= \quad \log P(D|\theta) + \log P(\theta) - \log \int P(D|\theta)P(\theta)d\theta \\ &= \quad \sum_{i=1}^{n} -\{y_i - f(x_i;\theta)\}^2 + \log P(\theta) + C', \end{aligned} \tag{7}$$

where, C' is a constant independent of θ.

Maximizing the eq.(7) is equivalent to, what is called, the regularization method. In this context Regularization term corresponds to $\log P(\theta)$.

3 JEFFREYS' PRIOR

The Bayesian approach in the previous section requires selecting a prior $P(\theta)$.

If the prior $P(\theta)$ is selected correctly, an estimator that maximizes $P(\theta|D)$ is called "Bayes optimal". But it is not easy to give a correct prior before learning. Designing prior for neural networks is the key point of the Bayesian learning.[3] Whether a prior is appropriate or not depends on the target of learning. In this context various priors have been tested.[3] If we use "subjective prior", which depends on each particular problem, then

we should try various priors to find the best prior fit into the problem well.[2] However, here we do not discuss here about these approaches further.

On the other hand, "objective prior" is discussed to aim at getting a universal prior that could be applicable to any problem in this section. In the traditional Bayesian statistics, "non-informative prior" is proposed and studied for the case of only little prior knowledge being available. Jeffreys' prior[4] is a typical one. Jeffreys' prior is derived from the following the consideration.

Log likelihood curve is a function of θ. And it is determined by data D that depends on true parameter θ^*. If likelihood curve is translated only its location in the space of θ when θ^* is changed, this likelihood curve is called "data translated Likelihood". In this case, a locally uniform distribution satisfies the condition of the non-informative prior. On the other hand, when both the location and its shape of the likelihood curve are changed at every θ^*, the locally uniform distribution is not a non-informative prior. *

In the latter case, the metric ϕ is introduced to make the likelihood curve in the space of ϕ the data translated likelihood of ϕ. If proper ϕ is found out, we can construct a non-informative prior with the parameter transformation $\theta = g(\phi)$ and the random variable ϕ of uniform distribution.

Log likelihood about ϕ, $L(\phi)$ can be represented in the same manner as the eq.(4). Here, the Taylor expansion of $L(\phi)$ around $\phi^* = g^{-1}(\theta^*)$ by the second order, can be written as:

$$L(\phi) \approx L(\phi^*) + \frac{\partial L(\phi^*)}{\partial \phi}(\phi - \phi^*) + \frac{1}{2}\frac{\partial^2 L(\phi^*)}{\partial \phi^2}(\phi - \phi^*)^2. \tag{8}$$

The following constraints are required naturally:

$$L(\phi^*) = C,$$

$$\frac{\partial L(\phi^*)}{\partial \phi} = 0,$$

$$\frac{\partial^2 L(\phi^*)}{\partial \phi^2} = -nI(\phi^*).$$

$I(\theta)$ is the Fisher information which is written as:

$$I(\theta) \quad = \quad E\left[-\frac{\partial^2 \log P(D|\theta)}{\partial \theta^2}|\theta\right]. \tag{9}$$

Where, $E[\cdot]$ means taking expectation.

According to the foregoing, $L(\phi)$ can be written as:

$$L(\phi) \approx C - nI(\phi^*)(\phi^* - \phi)^2. \tag{10}$$

In order to make $L(\phi)$ the data translated likelihood, it is sufficient that $L(\phi)$ is the function of only $(\phi^* - \phi)$. Then the sufficient condition is that $I(\phi^*)$ in the eq.(10) is constant which is independent of ϕ^*.

This condition is clearly satisfied if we take:

*For further details of this discussion, see the reference.[4]

$$I(\phi) = I(\theta)\left(\frac{\partial\theta}{\partial\phi}\right)^2. \tag{11}$$

Therefore, the following is the sufficient condition for Jeffreys' prior:

$$\frac{\partial\theta}{\partial\phi} \propto [I(\theta)]^{-1/2}. \tag{12}$$

Because it is defined that $P(\phi)$ is a uniform distribution in the previous discussion, Jeffreys' prior which is denoted $P_J(\theta)$ is written as:

$$P_J(\theta) \propto [I(\theta)]^{1/2}. \tag{13}$$

Thus Jeffreys' prior can be written explicitly, but it can be calculated analytically only in the case of $P(D,\theta)$ belonging to a particular distribution family. In the case of the neural network model, it is not easy to give $I(\theta)$ as a simple function of θ.

4 NUMERICAL ANALYSIS AND APPROXIMATION

In the previous section, Jeffreys' prior could be given as a function of $I(\theta^*)$ making Jeffreys' prior more practically in the learning, we want to obtain simple relation between $P_J(\theta)$ and θ.

In this point of view we will show an approximation of Jeffreys' prior for neural networks. Here some assumptions are introduced to simplify the results.

If the true θ^* is given, Fisher information at θ^* is calculated numerically as follows:

$$I(\theta^*) = -\frac{1}{n}\frac{\partial^2 L(\theta)}{\partial\theta^2}|_{\theta^*}, \tag{14}$$

$$= \frac{1}{n}\sum_i^n \frac{\partial^2 (y_i - f(x_i|y,\theta))^2}{\partial\theta^2}|_{\theta^*}. \tag{15}$$

We suppose x is generated by locally uniform distribution $P(x)$ in the range of $[-10,10]$,y is generated by $P(y|x,\theta^*)$ and σ is small sufficiently. Because the analysis of the relation between Jeffreys' prior and each parameter of neural networks is important, the case of a single parameter is studied. We pick up only one parameter in neural networks as free variable θ, and fix other parameters. Under this condition Fisher information of each parameter can be calculated easily.

As a result, each Fisher information of the parameter a, b_i, v_i is constant in this framework, and Jeffreys' prior for them is uniform distribution. On the other hand Fisher information of w becomes non-linear function of w. Jeffreys' prior for w is not uniform distribution. The Jeffreys' prior can be calculated and shown in Figure1. This result shows that weight parameters of hidden units are more important than other parameters in the neural networks (1).This leads to the fact that the prior probability which expresses existence probability of w becomes smaller as $|w|$ becomes larger.

We can show the plot of $[I(w)]^{-1/2}$ by numerical calculation and it can be approximated as follows,

$$\frac{\partial w}{\partial \phi} = [I(w)]^{-1/2} \approx 10.0w - 1.0. \tag{16}$$

Combine with $w = g(\phi)$,the eq.(12) and the eq.(16),

$$\frac{\partial g(\phi)}{\partial \phi} \approx \alpha(g(\phi)), \tag{17}$$

where α is a constant.

To solve the above equation, we get the following transformation:

$$w = g(\phi) \approx \beta(e^{\alpha})^{\phi}, \tag{18}$$

where α, β are constants and w is considered as $w > 0$. We show the approximation of Jeffreys' prior as:

$$P_J(w) \propto [I(w)]^{1/2} \approx \frac{1}{\alpha|w|}. \tag{19}$$

We can show each error curve of the parameter w and ϕ (in Figure3 and Figure4). In the case of using the transformation the eq.(18), each error curve of ϕ keeps its shape independent of ϕ^* as changing of the true parameter. On the other hand, shapes of the error curves of w are different as each w^*.

5 APPLICATIONS

WEIGHT DECAY

Weight decay technique is known as a regularization method, and various decay terms have already been proposed. Many of these are, however given empirically. We try to derive the decay term from the approximated Jeffreys' prior(19).

To maximize $\log P(\theta|D)$ the following term:

$$-L(\theta) - \log P(\theta)$$

is minimizing by a learning algorithm.

Simple gradient decent algorithm updates θ as the following formula,

$$\delta\theta = \eta(\frac{\partial L(\theta)}{\partial \theta} + \frac{\partial \log P_J(\theta)}{\partial \theta}),$$

where η is the learning rate. Combined with the eq.(19), we can determine the weight decay term of w by:

$$\frac{\partial \log P_J(w_i)}{\partial w_i} \approx -\eta k/|w_i|,$$

where, k is determined by the normalization constant of $P_J(\theta)$. (Note:The range of θ must be given to make a prior properly, and $P(\theta)$ is normalized in this range.)

For other parameters, the weight decay term is set to 0 because $P_J(\theta)$ does not depends on θ as in the above discussion.

INITIAL CONDITION OF PARAMETERS

We can maximize the eq.(7) despite of the eq.(4) by the Bayesian approach. In this way, the initial condition of the parameters, which take large value about the eq.(7), is considered. For parameters of hidden units, w_i is given by the transform the eq.(18), where ϕ_i is sampled by uniform distribution. The initial condition of parameters except w are all determined by uniform distribution as ordinary back propagation, because $P_j(\theta)$ of these parameters are a uniform distribution. This setting method is used in the simple preliminary experiments,[5] and it has stable performance of learning for different targets.

6 CONCLUSION

In this paper, Jeffreys' prior has been discussed for neural networks. And it has been given by numerical calculation. It has been revealed that only in the case of the parameter of hidden units , Jeffreys' prior is different from uniform distribution. About the other parameters, Jeffreys' prior has given as a uniform distribution, therefore the Bayesian learning with this prior is equivalent to the ordinary cases.

We have obtained a simple transformation of the parameter that approximates Jeffreys' prior. Using this approximation, some learning techniques have been proposed, and the application of this transformation has worked effectively to make error curves corresponding to various examples stable. This feature is effective for the robustness of learning algorithms. In this study, we use rough approximation to solve the differential equation17 easily. More precise approximations are studied. Recently Clarke and Barron suggested that Jeffreys' prior is effective to obtain mini-max optimal.[6] Experimental studies for many applications are also important.

7 ACKNOWLEDGMENTS

Special thanks are due to Dr. Kawabata for discussions about Jeffreys' prior and all members of Mathematical Informatics section for their helpful communications. The author wishes to express my gratitude to M.Suwa, director of Information Science Division for affording an opportunity of this study. This work was supported in part by Real World Computing Program.

Figure 1: Plot of Jeffreys' prior $P_J(w)$ calculated numerically by eqs.(13) and (15).(Note: $w = 0$ is an irregular point, however it is plotted for convenience.)

Figure 2: $[I(\theta)]^{-1/2}$ and a approximation result($\theta = w$)

Figure 3: Plots of error curves of $\theta = w$, (X axis is θ, Y axis is $\sum_i^n \{f(x_i|\theta^*) - f(x_i|\theta)\}^2$)

Figure 4: Plots of error curves of ϕ, (X axis is ϕ , Y axis is $\sum_i^n \{f(x_i|\phi^*) - f(x_i|\phi)\}^2$)

8 REFERENCES

[1] Levin,E., Tishby,N. and Solla,S.:" A Statistical Approach to Learning and Generalization" , *Proc. IEEE vol.78 No.10*, pp.1568-1574(1990).

[2] Mackay,D.J.C.:"A Practical Bayesian Framework for Backpropagation Networks",*Neural Computation*,vol.4 No.3,pp.448-472,(1992).

[3] Wray,L.B., Andreas S.W.:" Bayesian Back-Propagation", *Complex Systems 5*, pp.603-643, (1991).

[4] Box,G. and Tiao,G.:"Bayesian Inference in Statistical Analysis", Addison Wesley,(1973).

[5] Motomura, Y.: "Improving Learning Performance of Layered Neural Networks with Multi Resolution Structure",Technical Report of IEICE,NC93-50 (in japanese)(1993).

[6] Clarke,B.S. and Barron,A.:"Jeffreys' Prior is asymptotically least favorable under entropy risk", *submitted to the JSPI*,(1992).