The Effects of Quantization on the Backpropagation Learning

Kazushi IKEDA Akihiro SUZUKI Kenji NAKAYAMA
Dept of Elec. Comp. Eng., Kanazawa Univ.
2-40-20 Kodatsuno, Kanazawa 920 Japan
kazushi@t.kanazawa-u.ac.jp

ABSTRACT

The effects of the quantization of the parameters of a learning machine are discussed. The learning coefficient should be as small as possible for a better estimate of parameters. On the other hand, when the parameters are quantized, it should be relatively larger in order to avoid the paralysis of learning originated from the quantization. How to choose the learning coefficient is given in this paper from the statistical point of view.

1. Introduction

In many of theoretical analyses of learning machines which realize an input and output relationship, *e.g.* multilayer perceptrons, their parameters are assumed to be analog and continuous, though they are digital and discrete in practical implementations because of the advantages that storage of parameters in digital memories can reduce the scale of circuits. Then, it is necessary to elucidate what happens when the parameters are quantized.

One important effect of quantization is that it makes the performance of the machine worse because the parameters change a little from the optimal [2,3,7,8]. And another important effect occurs in the learning process. In case that the parameters of the machine are modified by means of a stochastic gradient descent method, *e.g.* the backpropagation learning, the quantization causes paralysis of learning when the correction steps become smaller in absolute value than the resolution of the parameter. To avoid this, several algorithms based on parameter perturbation have been proposed [2,4,8] but they are rather *ad hoc* because they have no theoretical background and only move the parameters after the paralysis.

One method to overcome the paralysis is to make the learning coefficient large enough. But it simultaneously makes the training error increase [1,5,6]. In this work, the optimal learning coefficient and the variance of the trained parameters in that case are given from the statistical point of view, and it implies how much precision of the parameters is necessary in order that the machine has a desired precision.

2. Statistical Model of Quantization

Here, we consider the quantization error in two cases, one of which is the case of fixed-point representation, and the other is the case of floating-point representation.

In the case of the fixed-point representation, we assume that the quantization for b bits is done by the round-off of the $(b+1)$th bit including a bit for its sign. Then, the quantized parameter w' and the original parameter w have the relation

$$|w' - w| < w_{\max} \times 2^{-b} \qquad (1)$$

where b is the number of bits to represent the value including the bit for the sign, and $w_{\max}$ is the maximum value that can be represented in b bits.

In the case of the floating-point representation, we assume that the original parameter

0-7803-4122-8/97 $10.00©1997 IEEE

w is represented as the product of a mantissa between 1 and 10 in binary and an exponent 2^c, and that the quantization is done by the round-off of the $(b+1)$th bit of the mantissa. Then, w' and w have the relation $|w' - w| < 2^c \times 2^{-b}$ where b is the number of bits of the mantissa. Then, using

$$2^c < |w| < 2 \times 2^c,$$

we can derive the inequality

$$|w' - w| < |w| \times 2^{-b}. \qquad (2)$$

From the comparison of Equations (1) and (2), we can unify both representations by defining w_{M} as

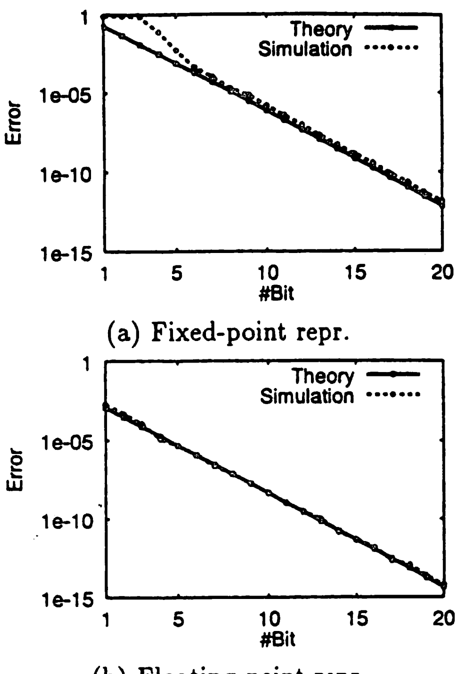
$$w_{\mathrm{M}} = \begin{cases} w_{\max} & \text{when fixed-point repr.,} \\ |w| & \text{when floating-point repr.,} \end{cases}$$

and then, the range of quantization error is written as $|w' - w| < w_{\mathrm{M}} \times 2^{-b}$.

We assume in the following that the quantization errors are independently uniformly distributed in the ranges. From this assumption, the mean and the variance of quantization error are derived as

$$\mathrm{E}\left[|w' - w|\right] = 0, \quad \mathrm{E}\left[|w' - w|^2\right] = \frac{1}{3} {w_{\mathrm{M}}}^2 2^{-2b},$$

respectively. Figure 1 shows the increase of the output error of the theoretical analysis under the assumption and that of the computer simulations in the case of a pattern classifier by a multilayer perceptron, which has 7 neurons in the input layer, 10 neurons in the hidden layer, and 1 neuron in the output layer, and which has learned given 3 input-output patterns by the backpropagation algorithm. They seem to agree well.



(a) Fixed-point repr.

(b) Floating-point repr.

Fig. 1: The increase of the output error of multilayer perceptrons.

3. Properties of Backpropagation Learning

We consider here the properties of the backpropagation learning, more general, the stochastic gradient descent method, that are elucidated in [1] and [6].

Assume that an input vector x given to a machine is independently chosen from a certain probability distribution having a density $p(x)$, and the machine outputs a vector y depending on both of the input x and the parameter vector w of the machine, that is, $y = F(x, w)$. Then, the ith given data is denoted by

$$(x_i, y_i) = (x_i, F(x_i, w_{\mathrm{true}}))$$

where w_{true} is the parameter the teacher machine has.

Let an error function which represents the distance between the output of the teacher machine and that of the student be denoted by

$$\begin{aligned} d(y', y; x) &= d(F(x, w_{\mathrm{true}}), F(x, w); x) \\ &= d(w; x), \end{aligned}$$

then learning is equivalent to minimizing the average $D(w; p(x))$ of the error function d on the input x, that is,

$$D(w; p(x)) = \int d(w; x) p(x) \, \mathrm{d}x.$$

In some cases, we do not know the distribution $p(x)$ and we have only a finite number of data. Then, we use the empirical distribution $p'(x) = \frac{1}{t} \sum_{j=1}^{t} \delta(x - x_j)$ constructed by

t given samples $(x_j, y_j), j = 1, \ldots, t$, instead of $p(x)$.

We define the optimal parameter $\boldsymbol{w}_{\text{opt}}$ and the stochastic gradient descent method as $\boldsymbol{w}_{\text{opt}} = \arg\min_{\boldsymbol{w}} D(\boldsymbol{w}; p(x))$ and

$$\boldsymbol{w}_{n+1} = \boldsymbol{w}_n + \Delta \boldsymbol{w}_n, \quad \Delta \boldsymbol{w}_n = -\eta \nabla d(\boldsymbol{w}_n; x),$$

respectively, where ∇ denotes a differential operator in respect of w, $\eta > 0$ is a learning coefficient, and n represents the number of iteration.

After learning enough, an estimated parameter $\boldsymbol{w}_{\text{est}}$ is obtained, that is, $\boldsymbol{w}_{\text{est}} = \lim_{n \to \infty} \boldsymbol{w}_n$. Since the estimated parameter $\boldsymbol{w}_{\text{est}}$ varies according to the samples given at the learning period, it is a stochastic variable which has a probability distribution density denoted by $p(\boldsymbol{w}_{\text{est}})$. Then, the next theorem holds:

Theorem 1 *If the initial value of the parameter w is appropriate, then*

$$\begin{aligned} \int (\boldsymbol{w}_{\text{est}} - \boldsymbol{w}_{\text{opt}}) p(\boldsymbol{w}_{\text{est}}) \, \mathrm{d}\boldsymbol{w}_{\text{est}} &= 0, \\ \int (\boldsymbol{w}_{\text{est}} - \boldsymbol{w}_{\text{opt}})^2 p(\boldsymbol{w}_{\text{est}}) \, \mathrm{d}\boldsymbol{w}_{\text{est}} &= \frac{\eta}{2} Q^{-1} G, \end{aligned}$$

where

$$\begin{aligned} G &= \int \nabla d(\boldsymbol{w}_{\text{opt}}; x) \nabla d(\boldsymbol{w}_{\text{opt}}; x)^T p(x) \, \mathrm{d}x, \\ Q &= \int \nabla\nabla d(\boldsymbol{w}_{\text{opt}}; x) p(x) \, \mathrm{d}x. \end{aligned}$$

The proof of this theorem is given in [1] and [6].

4. Backpropagation Learning with Quantization

As shown in two preceding sections, the variance of parameters originated in quantization and in stochasticity are $\frac{1}{3} w_{\text{M}}{}^2 2^{-2b} I$ and $\frac{\eta}{2} Q^{-1} G$, respectively. Then, if we assume that they are independent, the variance of parameters is

$$\frac{1}{3} w_{\text{M}}{}^2 2^{-2b} I + \frac{\eta}{2} Q^{-1} G.$$

It seems, therefore, the variance gets smaller up to $\mathrm{O}(2^{-2b})$ as the learning coefficient becomes smaller. In actual, however, learning paralyzes when the step $\Delta \boldsymbol{w}$ becomes smaller than $\frac{1}{2} w_{\text{M}} 2^{-b}$. To avoid the paralysis, the learning coefficient should be large enough, though that makes the variance of stochasticity large. Then, there exists an optimal learning coefficient to minimize the variance of parameters.

Let the learning coefficient η be $\mathrm{O}(2^{-k})$. Then, learning could achieve

$$\boldsymbol{w} - \boldsymbol{w}_{\text{opt}} = \begin{cases} \mathrm{O}(2^{-k/2}) & \text{if } k \leq 2b, \\ \mathrm{O}(2^{-b}) & \text{otherwise,} \end{cases}$$

unless the step $\Delta \boldsymbol{w}$ were quantized. Because the step $\Delta \boldsymbol{w}$ can be approximated as

$$\begin{aligned} \Delta \boldsymbol{w} &= -\eta \nabla d(\boldsymbol{w}; x) \\ &= -\eta \nabla^2 d(\boldsymbol{w}_{\text{opt}}; x) (\boldsymbol{w} - \boldsymbol{w}_{\text{opt}}) \end{aligned}$$

by Taylor expansion if $\boldsymbol{w} - \boldsymbol{w}_{\text{opt}}$ is small, $\Delta \boldsymbol{w} = \mathrm{O}(2^{-k}) (\boldsymbol{w} - \boldsymbol{w}_{\text{opt}})$ is satisfied and the learning stops when $\Delta \boldsymbol{w}$ is $\mathrm{O}(2^{-b})$, that is, when $\boldsymbol{w} - \boldsymbol{w}_{\text{opt}} = \mathrm{O}(2^{-b+k})$. In order to minimize the order of $\boldsymbol{w} - \boldsymbol{w}_{\text{opt}}$, k should be $2b/3$, Then, the following theorem is derived:

Theorem 2 *When the learning coefficient η is $\mathrm{O}(2^{-2b/3})$, the variance of parameters is minimized to $\mathrm{O}(2^{-2b/3})$.*

Reversely, we have to give the parameters a precision of $\mathrm{O}(2^{-b})$ if we need to estimate the parameters with a precision of $\mathrm{O}(2^{-b/3})$.

The result above holds even in the case that $\nabla d(\boldsymbol{w}_n; x)$ itself is quantized because

$$\begin{aligned} \Delta \boldsymbol{w} &= -\eta \nabla d(\boldsymbol{w}; x) + \varepsilon_{\text{Q}} \\ &= -\eta \nabla^2 d(\boldsymbol{w}_{\text{opt}}; x) (\boldsymbol{w} - \boldsymbol{w}_{\text{opt}}) + \varepsilon_{\text{Q}} \end{aligned}$$

where ε_{Q} is the quantization error which is at most 2^{-b}, and $\boldsymbol{w} - \boldsymbol{w}_{\text{opt}}$ has the precision of $2^{-b/3}$.

5. Computer Simulations

Computer simulations are done to confirm the theoretical result above. The student machine with 3 dimensional parameter w outputs

$$f(x; w) = \tanh \frac{x \cdot w}{2}$$

according to a three dimensional input x and is given the samples which consists of an input

x newly chosen subject to the uniform distribution in $[-1,1]^3$ and the according output $f(x_i; w_{\rm opt})+n$ of the teacher machine where n is a noise term chosen subject to $N(0, 0.0001)$. the samples are given until the student machine converges. Figure 2 shows the variance of the estimated parameter $w_{\rm est}$ when the student machine converges. the variance has the same slope as $-k$ and $k/2$ in the left and right hand, respectively. It agrees with the result of the theoretical analysis.

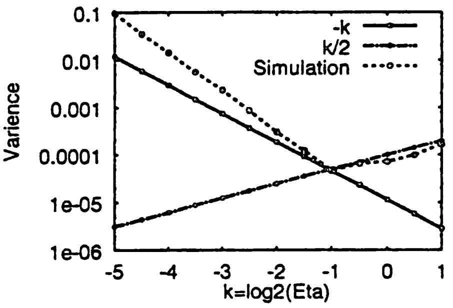


Fig. 2: Variance of the estimated parameter $w_{\rm est}$.

6. Conclusion

In this paper, the relation between the learning coefficient η and the precision of the estimated parameter $w_{\rm est}$ has been analyzed when the parameter is quantized with precision of 2^{-b}. Though the parameter has b-bit precision, the estimated parameter has only $b/3$-bit precision even in the optimal case.

References

[1] S. Amari, "Theory of adaptive pattern classifiers", *IEEE Trans. EC*, vol. 16, pp. 299–307, 1967.

[2] G. Duendar and K. Rose, "The effects of quantization on multilayer neural networks ", *IEEE Trans. NN*, vol. 6, pp. 1446–1451, 1995.

[3] K. Ikeda, A. Suzuki, and K. Nakayama, "The relation between the precision of the weights and the output error in multi-layer neural networks", NC 96-1, IEICE, 1996.

[4] A. J. Montalvo, P. W. Hollis, and J. J. Paolos, "On-chip learning with limited precision circuits", *Proc. Int'l Joint Conf. Neural Networks*, pp. 196–201, 1992.

[5] N. Murata, S. Yoshizawa, and S. Amari, "A criterion for determining the number of parameters in an artificial neural network model", *Artificial Neural Networks, Elsevier Science*, pp. 9–14, 1991.

[6] N. Murata, *A Statistical Asymptotic Theory of Learning*, Doctor's thesis, Univ. of Tokyo, 1992.

[7] Y. Xie and M. A. Jabri, "Analysis of the effects of quantization in multilayer neural networks using a statistical model", *IEEE Trans. NN*, vol. 3, pp. 334–338, 1992.

[8] Y. Xie and M. A. Jabri, "On the training of limited precision multilayer perceptrons ", *Proc. Int'l Joint Conf. Neural Networks*, pp. 942–947, 1992.