Analysis of the Effects of Quantization in Multi-Layer Neural Networks Using Statistical Model

Marwan Anwar Jabri

Neuromorphic LLC

Analysis of the Effects of Quantization in Multi-Layer Neural Networks Using Statistical Model

Yun Xie[*]
Department of Electronic Engineering
Tsinghua University
Beijing 100084, P.R.China

Marwan A. Jabri
School of Electrical Engineering
The University of Sydney
N.S.W. 2006, Australia

Abstract

A statistical quantization model is used to analyse the effects of quantization when digital technique is used to implement a real-valued feedforward multi-layer neural network. In this process, we introduce a parameter that we call "effective non-linearity coefficient" which is important in the study of the quantization effects. We develop, as function of the quantization parameters, general statistical formulations of the performance degradation of the neural network caused by quantization. Our formulation predicts (as intuitively one may think) that network's performance degradation gets worse when the number of bits is decreased; a change of the number of hidden units in a layer has no effect on the degradation; for a constant "effective non-linearity coefficient" and number of bits, an increase in the number of layers leads to worse performance degradation of the network; the number of bits in successive layers can be reduced if the neurons of the lower layer are non-linear.

I. Introduction

One of the first problems in the hardware implementation of artificial neural networks is to determine how many bits are necessary to represent physical states, parameters and variables, to ensure certain learning and generalization performance.

In the analysis of the effects of quantization on learning and generalisation, one of the complications is the non-linear transfer functions commonly used to represent neurons. Assumptions on the distributions of the neuron responses are necessary to relate quantization widths in different layers.

Not much work has been done in this area, although efforts that may be of interest, for example, are those that target the determination of a relationship between the number of hidden units, the number of hidden layers and the learning and generalisation capability of a network [1].

In this paper we use statistical models to represent the quantized values of weights and neurons a feedforward multi-layer neural network, and given some assumptions on their distributions, we investigate a relationship between bit-resolution, number of hidden units and layers, and the performance degradation of the network. The paper is structured as follows. In Section II. we define our statistical model and some terminology. Section III. develops for a three layer network the relationships between inputs and outputs, based on our statistical models. The assumptions on the neural network are stated. We also introduce in this section the "effective non-linearity coefficient" which plays an important role in the understanding of the role of non-linear units on bit-resolution. In Section IV. we reveal and analyse the relationship between bit-resolution, network architecture and the performance degradation of the network. Finally, in Section V. we present some conclusions.

II. Statistical Model of Quantization

The effects of quantization can be approached in many ways. For complex systems, the statistical model is a convenient method. The idea is that a quantized signal can be represented by the original signal plus a quantization error (noise), $e(n)$. We make the following assumptions:

- $e(n)$ is a stationary random process;
- $e(n)$ is independent of the signal;
- $e(n)$ is a white noise;
- Δ is the quantization width, $e(n)$ is uniformly distributed in $[-\Delta, \Delta]$, thus, the mean is zero and the variance is $\Delta^2/12$ denoted by σ_Δ^2.

This model does not give correct results in the case when the signal is highly self-correlated [2].

In the following $E\{z\}$ represents the expectation of z and σ_z^2 denotes the variance of z.

[*] Currently with School of Electrical Engineering, The University of Sydney.

III. Effects of Quantization

A. First Hidden Layer

In the following, x_k^0 is the input signal from input node k, w_{ik}^0 is the weight connecting node k in the input layer and node i in the first hidden layer, y_i^0 is the input of node i while x_i^1 is the output:

$$y_i^0 = \sum_{k=0}^{K_1-1} w_{ik}^0 x_k^0 \tag{1}$$

$$x_i^1 = f(y_i^0) \tag{2}$$

$f(.)$ is the non-linear transfer function of a node and the bias is treated as an input.

Assuming x_k^0 and w_{ik}^0 are quantized by N bits (one bit for sign), Δ_0 is the quantization width and the quantized value falls in $[-\Delta_0(2^{N-1}-1), \Delta_0(2^{N-1}-1)]$ which can be approximated to $[-\Delta_0 2^{N-1}, \Delta_0 2^{N-1}]$ when N is large enough, then:

$$\begin{aligned} y_i^0 &= \sum_{k=0}^{K_1-1} (w_{ik}^0 + \Delta w_{ik}^0)(x_k^0 + \Delta x_k^0) \\ &\simeq \sum_{k=0}^{K_1-1} w_{ik}^0 x_k^0 + \Delta y_i^0 \qquad (3) \end{aligned}$$

$$\Delta y_i^0 \overset{\text{def}}{=} \sum_{k=0}^{K_1-1} w_{ik}^0 \Delta x_k^0 + \sum_{k=0}^{K_1-1} x_k^0 \Delta w_{ik}^0 \tag{4}$$

Where the second order items are omitted. Δx_k^0 and Δw_{ik}^0 are quantization noises and are independent of each other, w_{ik}^0 and x_k^0. The corresponding variance is:

$$\sigma_{\Delta_0}^2 \overset{\text{def}}{=} \Delta_0^2/12 \tag{5}$$

Assuming x_k^0 and w_{ik}^0 are uniformly distributed in $[-\Delta_0 2^{N-1}, \Delta_0 2^{N-1}]$ and independent of each other we have:

$$E\{\Delta y_i^0\} = 0 \tag{6}$$

$$\begin{aligned} E\{(x_k^0)^2\} &= E\{(w_{ik}^0)^2\} \\ &= \int_{-\Delta_0 2^{N-1}}^{\Delta_0 2^{N-1}} \frac{x^2}{\Delta_0 2^N} dx \\ &= \frac{\Delta_0^2 2^{2N}}{12} \qquad (7) \end{aligned}$$

$$\begin{aligned} \sigma_{\Delta y_i^0}^2 &\overset{\text{def}}{=} E\{(\Delta y_i^0)^2\} \\ &= (\sum_{k=0}^{K_1-1} E\{(x_k^0)^2\} + \sum_{k=0}^{K_1-1} E\{(w_{ik}^0)^2\})\sigma_{\Delta_0}^2 \\ &= (2/3)K_1(\Delta_0 2^{N-1})^2 \sigma_{\Delta_0}^2 \\ &= (K_1 \Delta_0^4 2^{2N})/72 \\ &= \zeta_0 K_1 \Delta_0^4 2^{2N} \qquad (8) \end{aligned}$$

$$\zeta_0 = 1/72 \tag{9}$$

In order to relate quantization widths of different layers, we have to make some assumptions on the distribution of y_i^0.

$$\begin{aligned} E\{(x_k^0)^2(w_{ik}^0)^2\} &= E\{(x_k^0)^2\}E\{(w_{ik}^0)^2\} \\ &= (\Delta_0^4 2^{4N})/144 \end{aligned}$$

$$\begin{aligned} \sigma_{y_i^0}^2 &\overset{\text{def}}{=} E\{(y_i^0)^2\} \\ &= \sum_{k=0}^{K_1-1} E\{(x_k^0)^2(w_{ik}^0)^2\} \\ &= (K_1 \Delta_0^4 2^{4N})/144 \qquad (10) \end{aligned}$$

Where noise is omitted.

We define:

$$\begin{aligned} \max|y_i^0| &= \sqrt{3\sigma_{y_i^0}^2} \\ &= \sqrt{(3K_1\Delta_0^4 2^{4N})/144} \\ &= (\sqrt{K_1}\Delta_0^2 2^{2N})/(4\sqrt{3}) \\ &= \eta_0\sqrt{K_1}\Delta_0^2 2^{2N} \qquad (11) \end{aligned}$$

$$\eta_0 = 1/(4\sqrt{3}) \tag{12}$$

The distribution of y_i^0 is assumed to be uniform in $[-\max|y_i^0|, \max|y_i^0|]$. This distribution gives the same variance (power) of $\sigma_{y_i^0}^2$.

For simplicity, we use the non-linear transfer function as:

$$y = f(x) = \begin{cases} \Delta_1 2^{N-1} & x > \Delta_1 2^{N-1} \\ x & \text{In-between} \\ -\Delta_1 2^{N-1} & x < -\Delta_1 2^{N-1} \end{cases}$$

Δ_1 and N are quantization width and number of bits respectively in the quantization of the outputs of the first hidden layer and the weights between the first and the second hidden layers.

We define the "effective non-linearity coefficient" of the nodes in the first hidden layer as follows:

$$\begin{aligned} E_1 &\overset{\text{def}}{=} \frac{\max|y_i^0|}{\Delta_1 2^{N-1}} \\ &= \frac{\eta_0\sqrt{K_1}\Delta_0^2 2^{2N}}{\Delta_1 2^{N-1}} \\ &= 2\eta_0\sqrt{K_1}2^N\frac{\Delta_0^2}{\Delta_1} \qquad (13) \end{aligned}$$

$$\Delta_1 = \frac{2\eta_0\sqrt{K_1}\Delta_0^2 2^N}{E_1} \tag{14}$$

$$\begin{aligned} \sigma_{\Delta_1}^2 &\overset{\text{def}}{=} \frac{\Delta_1^2}{12} \\ &= \frac{4\eta_0^2 K_1 \Delta_0^2 2^{2N}}{E_1^2}\sigma_{\Delta_0}^2 \\ &= \frac{\eta_0^2}{3\zeta_0 E_1^2}\sigma_{\Delta y_i^0}^2 \qquad (15) \end{aligned}$$

Since neural networks are generally non-linear, E_1 is greater than unity.

Now we can derive the distribution of x_i^1. The probability density of x_i^1 in $(-\Delta_1 2^{N-1}, \Delta_1 2^{N-1})$ is:

$$p_{x_i^1} = \frac{1}{\Delta_1 2^N E_1} \tag{16}$$

and:

$$P(x_i^1 \in (-\Delta_1 2^{N-1}, \Delta_1 2^{N-1})) = \frac{1}{E_1} \tag{17}$$

$$\begin{aligned} P(x_i^1 = \Delta_1 2^{N-1}) &= P(x_i^1 = -\Delta_1 2^{N_1}) \\ &= \frac{E_1 - 1}{2E_1} \end{aligned} \tag{18}$$

The above two equations offer a method to measure E_1 in a neural network. We can also consider it as an other definition form of E_1.

Using this distribution, we get:

$$E\{(x_i^1)^2\} = \Delta_1^2 2^{2N-2} - \frac{\Delta_1^2}{3E_1} 2^{2N-1} \tag{19}$$

B. Second Hidden Layer

x_l^1 is the input signal from the first hidden layer, w_{il}^1 is the weight connecting node l in the first hidden layer and node i in the second hidden layer, y_i^1 is the input of node i in the second hidden layer and x_i^2 is its output:

$$y_i^1 = \sum_{l=0}^{K_2-1} w_{il}^1 x_l^1 \tag{20}$$

$$x_i^2 = f(y_i^1) \tag{21}$$

Just like in the first hidden layer, due to noise:

$$\begin{aligned} y_i^1 &= \sum_{l=0}^{K_2-1} (w_{il}^1 + \Delta w_{il}^1)(x_l^1 + \Delta y_l^1) \\ &\simeq \sum_{l=0}^{K_2-1} w_{il}^1 x_l^1 + \Delta y_i^1 \end{aligned} \tag{22}$$

$$\Delta y_i^1 \overset{\text{def}}{=} \sum_{l=0}^{K_2-1} w_{il}^1 \Delta x_l^1 + \sum_{l=0}^{K_2-1} x_l^1 \Delta w_{il}^1 \tag{23}$$

Δw_{il}^1 is the quantization noise with zero mean and variance of $\sigma_{\Delta_1}^2 = \Delta_1^2/12$. Because of non-linearity of $f(.)$:

$$\Delta x_l^1 = \begin{cases} \text{QN} + \Delta y_l^0 & x_l^1 \in (-\Delta_1 2^{N-1}, \Delta_1 2^{N-1}) \\ 0 & x_l^1 = -\Delta_1 2^{N-1}, or \Delta_1 2^{N-1} \end{cases}$$

Where QN is the quantization noise.

This means E_1^{-1} percent of x_l^1 has $\Delta x_l^1 \neq 0$.

Assuming the distribution of w_{il}^1 is uniform in $[-\Delta_1 2^{N-1}, \Delta_1 2^{N-1}]$:

$$E\{(w_{il}^1)^2\} = \frac{\Delta_1^2 2^{2N-2}}{3} \tag{24}$$

$$\begin{aligned} \sigma_{\Delta y_i^1}^2 &\overset{\text{def}}{=} E\{(\Delta y_i^1)^2\} \\ &= \frac{K_2}{E_1} E\{(w_{il}^1)^2\} \sigma_{\Delta x_l^1}^2 + K_2 E\{(x_l^1)^2\} \sigma_{\Delta_1}^2 \\ &= \frac{K_2 \Delta_1^2 2^{2N-2}}{3E_1} (\frac{3\zeta_0 E_1^2}{\eta_0^2} \sigma_{\Delta_1}^2 + \sigma_{\Delta_1}^2) \\ &\quad +(K_2 \Delta_1^2 2^{2N-2} - \frac{K_2 \Delta_1^2 2^{2N-1}}{E_1}) \sigma_{\Delta_1}^2 \\ &= \frac{1}{48} (\frac{\zeta_0 E_1}{\eta_0^2} - \frac{1}{3E_1} + 1) K_2 \Delta_1^4 2^{2N} \\ &= \zeta_1 K_2 \Delta_1^4 2^{2N} \end{aligned} \tag{25}$$

$$\zeta_1 \overset{\text{def}}{=} \frac{1}{48} (\frac{\zeta_0 E_1}{\eta_0^2} - \frac{1}{3E_1} + 1) \tag{26}$$

$$\begin{aligned} \sigma_{y_l^1}^2 &\overset{\text{def}}{=} E\{(y_l^1)^2\} \\ &= \sum_{l=0}^{K_2-1} E\{(w_{il}^1)^2\} E\{(x_l^1)^2\} \\ &= \frac{K_2 \Delta_1^4 2^{4N}}{48} - \frac{K_2 \Delta_1^4 2^{4N}}{72E1_1} \end{aligned} \tag{27}$$

Just like in the first hidden layer:

$$\begin{aligned} \max|y_l^1| &\overset{\text{def}}{=} \sqrt{3\sigma_{y_l^1}^2} \\ &= \sqrt{\frac{1}{16} - \frac{1}{24E_1}} \sqrt{K_2} \Delta_1^2 2^{2N} \\ &= \eta_1 \sqrt{K_2} \Delta_1^2 2^{2N} \end{aligned} \tag{28}$$

$$\eta_1 \overset{\text{def}}{=} \sqrt{\frac{1}{16} - \frac{1}{24E_1}} \tag{29}$$

The effective non-linear coefficient of the second hidden layer is defined as:

$$\begin{aligned} E_2 &\overset{\text{def}}{=} \frac{\max|y_l^1|}{\Delta_2 2^{N-1}} \\ &= \frac{2\eta_1 \sqrt{K_2} \Delta_1^2 2^N}{\Delta_2} \end{aligned} \tag{30}$$

Δ_2 and N are the quantization width and the number of bits in the quantization of the outputs of the second hidden layer and the weights between the second and the third hidden layers. From the above equation:

$$\Delta_2 = \frac{2\eta_1 \sqrt{K_2} \Delta_1^2 2^N}{E_2} \tag{31}$$

$$\begin{aligned} \sigma_{\Delta_2}^2 &\overset{\text{def}}{=} \frac{\Delta_2^2}{12} \\ &= \frac{4\eta_1^2 K_2 \Delta_1^2 2^{2N}}{E_2^2} \sigma_{\Delta_1}^2 \\ &= \frac{\eta_1^2}{3\zeta_1 E_2^2} \sigma_{\Delta y_l^1}^2 \end{aligned} \tag{32}$$

C. Third Hidden Layer

x_h^2 is the input signal from the second layer, w_{ih}^2 is the weight connecting node h in the second layer and node i in the third layer, y_i^2 is the input of node i in the third layer and x_i^3 is its output.

$$y_i^2 = \sum_{h=0}^{K_3-1} w_{ih}^2 x_h^2 \tag{33}$$

$$x_i^3 = f(y_i^2) \tag{34}$$

The same as in the second layer, we can derive:

$$y_i^2 \simeq \sum_{h=0}^{K_3-1} w_{ih}^2 x_h^2 + \Delta y_i^2 \tag{35}$$

$$\Delta y_i^2 = \sum_{h=0}^{K_3-1} w_{ih}^2 \Delta x_h^2 + \sum_{h=0}^{K_3-1} x_h^2 \Delta w_{ih}^2 \tag{36}$$

$$\begin{aligned}\sigma^2_{\Delta y_i^2} &\overset{\text{def}}{=} E\{(\Delta y_i^2)\} \\ &= \frac{1}{48}(\frac{\zeta_1 E_2}{\eta_1^2} - \frac{1}{3E_2} + 1)K_3\Delta_2^4 2^{2N} \\ &= \zeta_2 K_3 \Delta_2^4 2^{2N} \qquad (37)\end{aligned}$$

$$\sigma^2_{y_i^2} = \frac{K_3\Delta_2^4 2^{4N}}{48} - \frac{K_3\Delta_2^4 2^{4N}}{72E_2} \tag{38}$$

$$\begin{aligned}\max|y_i^2| &\overset{\text{def}}{=} \sqrt{3\sigma^2_{y_i^2}} \\ &= \sqrt{\frac{1}{16} - \frac{1}{24E_2}}(\sqrt{K_3}\Delta_2^2 2^{2N}) \\ &= \eta_2\sqrt{K_3}\Delta_2^2 2^{2N} \qquad (39)\end{aligned}$$

$$\zeta_2 \overset{\text{def}}{=} \frac{1}{48}(\frac{\zeta_1 E_2}{\eta_1^2} - \frac{1}{3E_2} + 1) \tag{40}$$

$$\eta_2 \overset{\text{def}}{=} \sqrt{\frac{1}{16} - \frac{1}{24E_2}} \tag{41}$$

D. Generalisation to n layers

Following the similar assumptions made above we can generalize the above analysis to higher hidden layers. For the nth hidden layer, y_i^{n-1} is the input of node i, K_n is the number of nodes, E_n is the "effective non-linearity coefficient", and Δ_{n-1} and N are the quantization width and the number of bits in the quantization of the outputs of the $(n-1)$th hidden layer and the weights between the $(n-1)$th and the nth hidden layers, we have:

$$E_n = \frac{2\eta_{n-1}\sqrt{K_n}\Delta_{n-1}^2 2^N}{\Delta_n} \tag{42}$$

$$\eta_n = \sqrt{1/16 - 1/(24E_n)} \tag{43}$$

$$\zeta_n = \frac{1}{48}(\frac{\zeta_{n-1}E_n}{\eta_{n-1}^2} - \frac{1}{3E_n} + 1) \tag{44}$$

$$\sigma^2_{y_i^{n-1}} = K_n\Delta_{n-1}^4 2^{2N}(\frac{1}{48} - \frac{1}{72E_{n-1}}) \tag{45}$$

$$\sigma^2_{\Delta y_i^{n-1}} = \zeta_{n-1}K_n\Delta_{n-1}^4 2^{2N} \tag{46}$$

IV. Results Analysis

A. Signal to Noise Ratio

We can define the signal to noise ratios in the second and the third hidden layers.

In the second hidden layer, the signal to noise ratio R_2^p is:

$$\begin{aligned}R_2^p &\overset{\text{def}}{=} \frac{\sigma^2_{y_i^1}}{\sigma^2_{\Delta y_i^1}} \\ &= \frac{\frac{1}{3}\eta_1^2 K_3\Delta_1^4 2^{4N}}{\zeta_1 K_3 \Delta_1^4 2^{4N}} \\ &= \frac{\eta_1^2 2^{2N}}{3\zeta_1} \qquad (47)\end{aligned}$$

Similarly in the third hidden layer:

$$R_3^p = \frac{\eta_2^2 2^{2N}}{3\zeta_2} \tag{48}$$

We define:

$$\max|\Delta y_i^1| = \sqrt{3\sigma^2_{\Delta y_i^1}} \tag{49}$$

We can define the signal to noise ratio in another way:

$$\begin{aligned}R_2^m &\overset{\text{def}}{=} \frac{\max|y_i^1|}{\max|\Delta y_i^1|} \\ &= \sqrt{R_2^p} \\ &= \frac{\eta_1 2^N}{\sqrt{3\zeta_1}} \qquad (50)\end{aligned}$$

Similarly we have:

$$R_3^m = \frac{\eta_2 2^N}{\sqrt{3\zeta_2}} \tag{51}$$

First we consider the case where the second layer is the output layer. we assume Δy_i^1 has an uniform distribution in $[-\max|\Delta y_i^1|, \max|\Delta y_i^1|]$, the output nodes of the network are assumed to have the function of:

$$y = f(x) = \begin{cases} 1 & x > 0 \\ -1 & x < 0 \end{cases}$$

Now we can calculate the probability P_2^f that an output node gives a wrong output because of the quantization noise:

$$\begin{aligned}P_2^f &= 2\int_0^{\max|\Delta y_i^1|} \frac{1}{2\max|y_i^1|} \\ &\quad \int_{-\max|\Delta y_i^1|}^{-x} \frac{1}{2\max|\Delta y_i^1|}dydx \\ &= \frac{\max|\Delta y_i^0|}{4\max|y_i^1|} \\ &= \frac{1}{4R_2^m} \qquad (52)\end{aligned}$$

If we replace the assumption of uniform distribution of Δy_i^1 by such an approximation that if y_i^1 falls in $[-\max|\Delta y_i^1|, \max|\Delta y_i^1|]$ an error occurs in the output, then

$$P_2^f = \frac{\max|\Delta y_i^1|}{\max|y_i^1|} = \frac{1}{R_2^m} \qquad (53)$$

If the third layer is the output layer the results are similar:

$$P_3^f = \frac{1}{4R_3^m} \qquad (54)$$

Or:

$$P_3^f = \frac{1}{R_3^m} \qquad (55)$$

It is clear that the signal to noise ratio is directly associated with the performance degradation of the network.

B. *The Effects of Number of Bits*

From equations (47) and (48) it is clear that increasing the number of bits results in the improvement of signal to noise ratios and the ratio is independent of the number of nodes in each layer. This means that if we change the number of nodes in a layer the signal to quantization noise ratio in the output of the layer remains the same, although the network's performance may or may not degrade.

Then, is it necessary for all the layers to have the same number of bits? Let us consider the second layer. Suppose we use M bits to represent the output of the first hidden layer, then equation (13) will become:

$$E_1 = \frac{\max|y_i^0|}{\Delta_1 2^{M-1}} = \frac{2\eta_0\sqrt{K_1}\Delta_0^2 2^{2N-M}}{\Delta_1} \qquad (56)$$

$$\Delta_1 = \frac{2\eta_0\sqrt{K_1}\Delta_0^2 2^{2N-M}}{E_1} \qquad (57)$$

From equations (8) and (57) we have:

$$\sigma_{\Delta_1}^2 = \frac{\Delta_1^2}{12} = \frac{\eta_0^2 2^{2N-2M}}{3\zeta_0 E_1^2}\sigma_{\Delta y_i^0}^2 \qquad (58)$$

$\sigma_{\Delta y_i^0}^2$ is the variance of the input noise to the second hidden layer and $\sigma_{\Delta_1}^2$ is the variance of quantization noise produced by the second hidden layer.

If we want $\sigma_{\Delta_1}^2 = \sigma_{\Delta y_i^0}^2$, which means that the power of newly introduced noise in the output of the first hidden layer is the same as the power of the original noise in the input (this is requirement is commonly used in signal processing systems), then

$$\frac{\eta_0^2 2^{2N-2M}}{3\zeta_0 E_1^2} = 1 \qquad (59)$$

We get:

$$M = N - \frac{1}{2} - \log_2 E_1 \qquad (60)$$

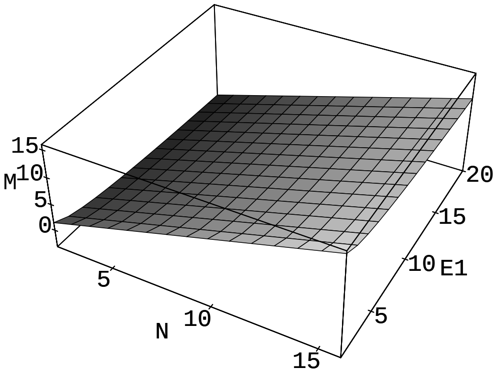

Figure 1: The relationship of M,N and E_1

Since $E_1 > 1$, $N < M$. The greater E_1 is, the more M is less than N. The reason is: as E_1 increases, less unprocessed information is passed to the second layer, so less bits are needed to represent the information. So from layer to layer the necessary number of bits decreases.

C. *The Effects of the Number of Layers*

This is not so explicit. Combining the definition of ζ and η, the signal to noise ratios becomes:

$$\begin{aligned} R_2^p &= \frac{(3E_1-2)2^{2N}}{2E_1^2+3E_1-1} \\ R_3^p &= \end{aligned}$$

$$\frac{(9E_1E_2-6E_1-6E_2+4)2^{2N}}{2E_1^2E_2^2+3E_1E_2^2-E_2^2+9E_1E_2-6E_2-3E_1+2}$$

Table 1 shows that as the number of layers and the non-linearity of the nodes increase, which indicates that the network needs to extract the small changes in the input signal, the signal to noise ratio decreases with a fixed number of bits. In this case more bits are needed to keep the signal to noise ratio constant.

Table 1: Effects of the number of layers on signal-to-noise ratio.

	2 layer	3 layer
$E_1 = E_2$	$R_2^p/2^{2N}$	$R_3^p/2^{2N}$
2	0.30	0.22
3	0.25	0.17
4	0.23	0.13
6	0.18	0.07

D. *About the Non-Linearity Coefficient*

E_i is the only parameter which makes our neural network different from linear ones. It is a measure of a neural network's non-linearity with respect to its inputs. According to the above analysis the performance of a network is closely associated with this parameter.

Actually we can interpret E_i as a measure of the permeability of a node. It controls the amount of information that is allowed to "pass through" and the amount that is compressed (abstracted) into one bit by the node after it produces its input weighted sum. Therefore the "effective non-linearity coefficient" controls the non-linear amount of information flowing in the network from its input to its output.

V. Conclusions

In this paper for feedforward multi-layer neural networks, we have developed relationships using statistical models between the quantization width of weights and neuron input/output states, and network architecture (number of hidden units and layers) and the performance of the network. To develop the relationships, assumptions on the distribution of the response of hidden units are necessary in order to relate distributions between successive layers. Following the assumptions, the "ve developed the general formulation of the probability (P^f) that an output node of the network gives a result different from the output when no quantization is involved. Our formulation predicts (as intuitively one may think) that P^f increases when the number of bits is decreased; a change of the number of hidden units in a layer has no effect on P^f; for a constant "Effective Non-linearity Coefficient" and number of bits, an increase in the number of layers leads to an increase in P^f; the number of bits in successive layers can be reduced if the neurons of the higher layer have an "effective non-linearity coefficient" that is greater or equal to one (non-linear neurons).

VI. Acknowledgment

This work was supported in part by the Australian Research Council and the Australian Telecommunication and Electronic Research Board. One of us (YX) acknowledges the support of National Education Committee of P.R.China.

References

[1] A. Krogh J. Hertz and R. G. Palmer. *Introduction to the Theory of Neural Computation.* Addison Wesley, 1990.

[2] A. V. Oppenheim and R. W. Schafer. *'Digital Signal Processing'*, chapter 9. Prentice-Hall, INC., Englewood Cliffs, New Jersey, 1975.