

Cramer Rao Maximum A-Posteriori Bounds on Neural Network Training Error for Non-Gaussian Signals and Parameters

Michael T. Manry, Cheng-Hsiung Hsieh, Michael S. Dawson, and Adrian K. Fung
Department of Electrical Engineering
University of Texas at Arlington
Arlington, Texas 76019

Steven J. Apollo
Lockheed Fort Worth Company, Mail Zone 2615
P.O. Box 748
Fort Worth, Texas 76101

Abstract

Previously, it has been shown that neural networks approximate minimum mean square estimators. In minimum mean square estimation, an estimate $\hat{\theta}'$ of the M -dimensional random parameter vector θ is obtained from a noisy N -dimensional input vector \mathbf{y} where \mathbf{y} has an additive noise component \mathbf{e} . For the Cramer-Rao maximum a-posteriori bounds on the variance of elements of $\hat{\theta}' - \theta$ to be tight, two necessary conditions are that the mapping from \mathbf{y} to θ is bijective and that \mathbf{e} and θ are both Gaussian. In this paper, we relax the second condition and develop bounds on the variances of elements of $\hat{\theta}' - \theta$ for the case where the input noise vector \mathbf{e} and the random parameter vector θ are non-Gaussian. First, we use linear transformations to obtain a new parameter vector ϕ from θ and a new input vector \mathbf{x} from \mathbf{y} . The parameter vector ϕ and \mathbf{n} , the noise component of the vector \mathbf{x} , are approximately Gaussian because of the central limit theorem, so Cramer-Rao maximum a-posteriori bounds on the variance of elements of $\hat{\phi}' - \phi$ can be tight. Lastly, an inverse transformation produces bounds on variances of elements of $\hat{\theta}' - \theta$. It is shown in examples that the bounds are tight when a neural network is used to estimate parameters of noisy exponential waveforms.

Published in: *Internat. Journ. of Intelligent Control and Syst.*, vol. 1, no. 3, 1996, pp. 381-391.

I. Introduction

Neural networks have been used to estimate a parameter vector $\boldsymbol{\theta}$ from a noisy input vector \mathbf{y} in many applications including remote sensing [1-3], forecasting of power loads for electric utilities [4-6], and in the design of nonlinear feedback controllers in robotics [7-10]. Recently, it has been shown that the training error for such neural network estimators is minimized when the neural network approximates the minimum mean square estimator $E[\boldsymbol{\theta}|\mathbf{y}]$ [11,12]. The performance of the estimator is characterized by the Cramer-Rao maximum a-posteriori (CRM) bound, which is obtained from the maximum a-posteriori (MAP) Fisher information matrix (FIM), $\mathbf{J}_\theta^{\text{MAP}}$ [11-13].

In neural network applications, the bounds represent target values for the network training error (mean-squared error). When this target is reached, training can be stopped. Failure of the training error to reach the bounds alerts the user to the fact that further or better training is necessary, that a larger network is required, or that the mapping from \mathbf{y} to $\boldsymbol{\theta}$ is not bijective.

When the elements of $\boldsymbol{\theta}$ are not jointly Gaussian, the Fisher information matrix may be impossible to calculate, or the bounds may be too small. Let $\boldsymbol{\theta}'$ denote an estimate of $\boldsymbol{\theta}$. In this paper, we develop bounds on the variances of elements of $\boldsymbol{\theta}' - \boldsymbol{\theta}$ for the case where the input signal \mathbf{y} and the parameter vector $\boldsymbol{\theta}$ are not limited to having Gaussian probability density functions (pdfs). This will facilitate the extension of the CRM bounds to applications such as power load forecasting, which is part of the closed loop control of utility power systems. In section II, we review some relevant details of the CRM bounds, and motivate our work. In section III, we describe a transformation approach for approximately converting non-Gaussian vectors \mathbf{y} and $\boldsymbol{\theta}$ into Gaussian vectors. In section IV, bounds are first found for the transformed parameters and then the bounds are transformed to the non-Gaussian case. Simulations that demonstrate our results are presented in section V.

II. Theory

In this section we review the CRM bounds for the general case, the additive Gaussian noise and Gaussian parameter case. We then look at the case of where θ is jointly uniform.

A. General Case

In minimum mean square estimation [13], an M-dimensional random parameter vector θ is to be estimated from an N-dimensional noisy input vector \mathbf{y} . Before CRM bounds can be calculated, we need to obtain the MAP Fisher information matrix (FIM), $\mathbf{J}_\theta^{\text{MAP}}$ [11-13], whose elements are defined as

$$\begin{aligned}
 J_\theta^{\text{MAP}}(i,j) &= E_\theta[J_\theta^{\text{MLE}}(i,j)] + E_\theta\left[\frac{\partial \Lambda_\theta^{\text{AP}}}{\partial \theta_i} \frac{\partial \Lambda_\theta^{\text{AP}}}{\partial \theta_j}\right], \\
 J_\theta^{\text{MLE}}(i,j) &\equiv E_c\left[\frac{\partial \Lambda_\theta^{\text{MLE}}}{\partial \theta_i} \frac{\partial \Lambda_\theta^{\text{MLE}}}{\partial \theta_j}\right], \\
 \Lambda_\theta^{\text{MLE}} &\equiv \ln(p_{\mathbf{y}|\mathbf{s}'}(\mathbf{y}|\mathbf{s}')), \\
 \Lambda_\theta^{\text{AP}} &\equiv \ln(p_\theta(\theta))
 \end{aligned} \tag{1}$$

where $E_\theta[\cdot]$ denotes expected value over the parameter vector θ , $E_c[\cdot]$ denotes expected value over the noise, and MLE denotes maximum likelihood estimation. Let $(J_\theta^{\text{MAP}})^{ij}$ denote an element of $(\mathbf{J}_\theta^{\text{MAP}})^{-1}$. Then [13],

$$\text{var}(\theta_i' - \theta_i) \geq (J_\theta^{\text{MAP}})^{ii} \tag{2}$$

where θ_i' can be any estimate of θ_i .

B. Gaussian Noise and Parameters

For the additive noise case, elements of \mathbf{y} are modelled as

$$y(n) = s'(\boldsymbol{\theta}, n) + e(n) \quad (3)$$

where $s'(\boldsymbol{\theta}, n)$ is the n th element of the deterministic signal vector \mathbf{s} and $e(n)$ is the n th element of the additive noise vector \mathbf{e} , which has the covariance matrix \mathbf{C}_e . The independent variable n may or may not represent discrete time. If $\boldsymbol{\theta}$ is Gaussian and if \mathbf{e} is additive and Gaussian, the elements of $\mathbf{J}_0^{\text{MAP}}$ in (1) can now be evaluated as

$$E_0[J_0^{\text{MLE}}(i,j)] = E_0\left[\left(\frac{\partial \mathbf{s}'}{\partial \theta_i}\right)^T \mathbf{C}_e^{-1} \left(\frac{\partial \mathbf{s}'}{\partial \theta_j}\right)\right], \quad (4)$$

$$E_0\left[\frac{\partial \Lambda_0^{\text{AP}}}{\partial \theta_i} \frac{\partial \Lambda_0^{\text{AP}}}{\partial \theta_j}\right] = d_0(i,j)$$

where $d_0(i,j)$ denotes an element of \mathbf{C}_0^{-1} , where \mathbf{C}_0 is the M by M covariance matrix of the parameter vector $\boldsymbol{\theta}$. Note that for the Gaussian case, the limit of the bounds in (2), as the noise variance increases, yields

$$\text{var}(\theta_i' - \theta_i) \geq \text{var}(\theta_i) \quad (5)$$

which is satisfied with equality when $\theta_i' = E[\theta_i]$. This property of the CRM bounds, which holds for the Gaussian parameter case, is called the *limit property*. CRM bounds must have the limit property to be tight. This property then, is a necessary condition for tightness.

C. Uniform Parameter Case

In many applications parameters are often bounded or nonnegative, and therefore non-

Gaussian. Consider the case where elements of $\boldsymbol{\theta}$ are statistically independent and uniform. In order to make the pdf $p_{\theta}(\boldsymbol{\theta})$ differentiable, we give Gaussian tails to each of its elements as

$$p_{\theta}(\boldsymbol{\theta}) = \frac{1}{1+\sqrt{2\pi}\sigma} \left[u\left(\frac{1}{2}-|\boldsymbol{\theta}|\right) + e^{-\left(\boldsymbol{\theta}-\frac{1}{2}\text{sgn}(\boldsymbol{\theta})\right)^2/2\sigma^2} u\left(|\boldsymbol{\theta}|-\frac{1}{2}\right) \right]$$

Evaluating the a-priori part of (1), we get

$$E_{\theta} \left[\frac{\partial \Lambda_0^{AP}}{\partial \theta_i} \frac{\partial \Lambda_0^{AP}}{\partial \theta_j} \right] = \frac{\sqrt{2\pi}}{\sigma} \cdot \delta(i-j) \quad (7)$$

In the limit as we let σ^2 go to zero, the term in (7) goes to infinity and the CRM bounds go to zero. For the uniform parameter case and many other cases where $p_{\theta}(\boldsymbol{\theta})$ is zero for finite θ_i , the CRM bounds lack the limit property, are equal to zero, and are clearly useless.

In the remainder of this paper, our goal is to develop new bounds on $\text{var}(\hat{\theta}_i' - \theta_i)$ which have the limit property and are potentially tight, even when \mathbf{y} and $\boldsymbol{\theta}$ are non-Gaussian.

III. Transformations of Input and Parameter Vectors

Assume that the input vector \mathbf{y} and parameter vector $\boldsymbol{\theta}$ are non-Gaussian, but that \mathbf{y} has additive noise. It is well-known that the stochastic Cramer-Rao bounds are usually not tight for this case. Our goals here are to convert the input vector \mathbf{y} and parameter vector $\boldsymbol{\theta}$ to a new input vector \mathbf{x} and parameter vector $\boldsymbol{\phi}$ with approximately Gaussian probability density functions (pdfs).

A. Input Signal Vector

Assume that the input vector \mathbf{y} is put through a linear transformation, as $\mathbf{x} = \mathbf{B} \cdot \mathbf{y}$, before it

is fed into the estimator. The signal and noise components of \mathbf{x} are $\mathbf{s}(\boldsymbol{\theta})$ and \mathbf{n} respectively. The noise covariance matrix \mathbf{C}_n is

$$\mathbf{C}_n = \mathbf{B} \cdot \mathbf{C}_e \cdot \mathbf{B}^T \quad (8)$$

The vector \mathbf{x} is approximately Gaussian because of the central limit theorem [14]. The matrix \mathbf{B} can be chosen in at least two ways. First, it can be a rectangular transformation matrix of size N by N' , used for compressing the inputs down to a manageable number, while minimizing the degradation of the estimates [15]. In this case $N' \ll N$. Second, \mathbf{B} may represent the weights which feed the input vector \mathbf{y} into net functions [16] of a multilayer perceptron (MLP). \mathbf{C}_n is used in the MLE log likelihood function for θ , which is

$$\Lambda_{\theta}^{MLE} = \ln(p_{\mathbf{x}|\theta}(\mathbf{x}|\theta)) = C - \frac{1}{2}(\mathbf{x} - \mathbf{s}(\theta))^T \mathbf{C}_n^{-1} (\mathbf{x} - \mathbf{s}(\theta)) \quad (9)$$

where C denotes a constant.

B. Parameter Vector

Following the same procedure used for the input vectors, we want to transform the parameter vectors as $\phi = \mathbf{A} \cdot \theta$. The covariance matrix for ϕ is easily shown to be

$$\mathbf{C}_{\phi} = \mathbf{A} \cdot \mathbf{C}_{\theta} \cdot \mathbf{A}^T \quad (10)$$

Here, we want to constrain \mathbf{A} such that the elements of ϕ are approximately statistically independent, and \mathbf{C}_{θ} is diagonal. We choose the matrix \mathbf{A} as

$$\mathbf{A} = \mathbf{P} \cdot \mathbf{S} \quad (11)$$

where \mathbf{S} denotes a diagonal matrix which normalizes the elements of θ to unit variance and where the matrix \mathbf{P} denotes an orthogonal matrix such as the DCT (discrete cosine transform [17])

transformation matrix. Clearly then, many matrices \mathbf{A} exist. As the dimension M of $\boldsymbol{\theta}$ increases, $\boldsymbol{\phi}$ becomes Gaussian via the central limit theorem.

IV. Steps in the Derivation

In this section, our goal is to find the MAP FIM for $\boldsymbol{\phi}$ and use it to find an approximate FIM for $\boldsymbol{\theta}$.

A. Fisher Information Matrix for Transformed Parameters

In order to find the MAP FIM for $\boldsymbol{\phi}$, we rewrite the FIM expressions in (1) and (4), using \mathbf{x} and $\boldsymbol{\phi}$ instead of \mathbf{y} and $\boldsymbol{\theta}$. Relevant conditional pdfs are found as

$$\begin{aligned} p_{x|\boldsymbol{\theta}}(x|\boldsymbol{\theta}) &= p_n(x-s(\boldsymbol{\theta})) \\ p_{x|\boldsymbol{\phi}}(x|\boldsymbol{\phi}) &= p_{x|\boldsymbol{\theta}}(x|\mathbf{A}^{-1}\boldsymbol{\phi}) \end{aligned} \quad (12)$$

The MLE log likelihood function for $\boldsymbol{\theta}$ in (9) is now rewritten for $\boldsymbol{\phi}$ as

$$\begin{aligned} \Lambda_{\boldsymbol{\phi}}^{MLE} &= \ln(p_{x|\boldsymbol{\phi}}(\mathbf{x}|\boldsymbol{\phi})) \\ &= C - \frac{1}{2}(\mathbf{x}-s(\mathbf{A}^{-1}\boldsymbol{\phi}))^T \mathbf{C}_n^{-1}(\mathbf{x}-s(\mathbf{A}^{-1}\boldsymbol{\phi})) \end{aligned} \quad (13)$$

Similarly, the a priori log likelihood function for $\boldsymbol{\phi}$, which is

$$\Lambda_{\boldsymbol{\phi}}^{AP} = \ln(p_{\boldsymbol{\phi}}) \quad (14)$$

can be approximated via the central limit theorem as

$$\begin{aligned}
\Lambda_{\phi}^{AP} &\approx -\frac{1}{2}(\phi - m_{\phi})^T C_{\phi}^{-1} (\phi - m_{\phi}) \\
&= -\frac{1}{2} \sum_{m=1}^M \sum_{n=1}^M (\phi_m - m_{\phi}(m)) b_{\phi}(m,n) (\phi_n - m_{\phi}(n))
\end{aligned} \tag{15}$$

where $b_{\phi}(m,n)$ denotes an element of the matrix C_{ϕ}^{-1} and $m_{\phi}(m)$ denotes an element of the mean vector $\mathbf{m}_{\phi} = E[\phi]$.

The next step in the derivation is to find the MAP FIM $\mathbf{J}_{\phi}^{\text{MAP}}$ for parameter vector ϕ .

Elements of $\mathbf{J}_{\phi}^{\text{MAP}}$ are found as

$$J_{\phi}(i,j) = E_{\phi} [E_n [\frac{\partial \Lambda_{\phi}^{\text{MLE}}}{\partial \phi_i} \frac{\partial \Lambda_{\phi}^{\text{MLE}}}{\partial \phi_j}]] + E_{\phi} [\frac{\partial \Lambda_{\phi}^{\text{AP}}}{\partial \phi_i} \frac{\partial \Lambda_{\phi}^{\text{AP}}}{\partial \phi_j}] \tag{16}$$

where $E_n[\cdot]$ denotes expected value over the noise in \mathbf{x} . Using,

$$\frac{\partial \Lambda_{\phi}^{\text{MLE}}}{\partial \phi_i} = \left(\frac{\partial s(A^{-1}\phi)}{\partial \phi_i} \right)^T C_n^{-1} (x - s(A^{-1}\phi)) , \tag{17}$$

$$\frac{\partial \Lambda_{\phi}^{\text{MLE}}}{\partial \phi_j} = (x - s(A^{-1}\phi))^T C_n^{-1} \left(\frac{\partial s(A^{-1}\phi)}{\partial \phi_j} \right) , \tag{18}$$

the MLE portion of the MAP FIM is found as

$$\begin{aligned}
& E_{\phi} [E_n (\frac{\partial \Lambda_{\phi}^{MLE}}{\partial \phi_i}) (\frac{\partial \Lambda_{\phi}^{MLE}}{\partial \phi_j})] \\
& E_{\phi} [(\frac{\partial s(A^{-1}\phi)}{\partial \phi_i})^T C_n^{-1} E_n [(x-s(A^{-1}\phi))(x-s(A^{-1}\phi))^T] C_n^{-1} (\frac{\partial s(A^{-1}\phi)}{\partial \phi_j})] \\
& = E_{\phi} [(\frac{\partial s(A^{-1}\phi)}{\partial \phi_i})^T C_n^{-1} (\frac{\partial s(A^{-1}\phi)}{\partial \phi_j})]
\end{aligned} \tag{19}$$

Then,

$$\begin{aligned}
\frac{\partial s(\mathbf{A}^{-1}\boldsymbol{\phi})}{\partial \phi_i} \mathbf{C}_n^{-1} \left(\frac{\partial s(\mathbf{A}^{-1}\boldsymbol{\phi})}{\partial \phi_j} \right) &= E_{\boldsymbol{\phi}} \left[\left(\sum_{k=1}^n \left(\frac{\partial s}{\partial \theta_k} \right)^T d_{ki} \right) \mathbf{C}_n^{-1} \left(\sum_{m=1}^n \left(\frac{\partial s}{\partial \theta_m} \right) \right) \right. \\
&= \sum_{k=1}^n \sum_{m=1}^n d_{ki} d_{mj} E_0 \left[\left(\frac{\partial s(\boldsymbol{\theta})}{\partial \theta_k} \right)^T \mathbf{C}_n^{-1} \left(\frac{\partial s(\boldsymbol{\theta})}{\partial \theta_m} \right) \right] \\
&= \sum_{k=1}^N \sum_{m=1}^N d_{ki} d_{mj} E_0 [J^{\text{MLE}}(k,m)]
\end{aligned} \tag{20}$$

where d_{mj} denotes an element of \mathbf{A}^{-1} and where

$$J_0^{\text{MLE}}(k,m) \equiv \left(\frac{\partial s(\boldsymbol{\theta})}{\partial \theta_k} \right)^T \mathbf{C}_n^{-1} \left(\frac{\partial s(\boldsymbol{\theta})}{\partial \theta_m} \right) \tag{21}$$

Note that $\mathbf{J}_0^{\text{MLE}}$ is a *pseudo-MLE FIM* that would be the correct MLE FIM if \mathbf{n} were Gaussian.

Next, we need to find the a-priori part of $\mathbf{J}_\phi^{\text{MAP}}$. Using the Gaussian approximation of the pdf of $\boldsymbol{\phi}$,

$$\begin{aligned}
\frac{\partial \Lambda_\phi^{\text{AP}}}{\partial \phi_i} &\approx -\frac{1}{2} \sum_{n=1}^M b_\phi(i,n) (\phi_n - m_\phi(n)) - \frac{1}{2} \sum_{n=1}^M (\phi_n - m_\phi(n)) b_\phi(i,n) \\
&= -\sum_{n=1}^M b_\phi(i,n) (\phi_n - m_\phi(n))
\end{aligned} \tag{22}$$

$$\frac{\partial \Lambda_\phi^{\text{AP}}}{\partial \phi_j} \approx -\sum_{m=1}^M b_\phi(j,m) (\phi_m - m_\phi(m)) \tag{23}$$

Elements of the a-priori part of the MAP FIM are found as

$$E_{\Phi} \left[\frac{\partial \Lambda_{\Phi}^{AP}}{\partial \phi_i} \frac{\partial \Lambda_{\Phi}^{AP}}{\partial \phi_j} \right] \approx \sum_{m=1}^M \sum_{n=1}^M b_{\Phi}(i,m) b_{\Phi}(j,n) c_{\Phi}(n,m) = b_{\Phi}(i,j) \quad (24)$$

where $c_{\Phi}(m,n)$ denotes an element of matrix \mathbf{C}_{Φ} . Using (20), (21) and (24), $\mathbf{J}_{\Phi}^{\text{MAP}}$ is written as

$$\mathbf{J}_{\Phi}^{\text{MAP}} = (\mathbf{A}^T)^{-1} E_{\theta} [\hat{\mathbf{J}}_{\theta}^{\text{MLE}}] \mathbf{A}^{-1} + (\mathbf{A}^T)^{-1} \mathbf{C}_{\theta}^{-1} \mathbf{A}^{-1} \quad (25)$$

B. Pseudo Fisher Information Matrix

The MAP FIM $\mathbf{J}_{\Phi}^{\text{MAP}}$ in (25), which is used to find bounds on variances of elements of the vector $\Phi' - \Phi$, can be rewritten as

$$\mathbf{J}_{\Phi}^{\text{MAP}} = (\mathbf{A}^T)^{-1} \hat{\mathbf{J}}_{\theta}^{\text{MAP}} \mathbf{A}^{-1} \quad (26)$$

where $\hat{\mathbf{J}}_{\theta}^{\text{MAP}}$ is defined using (25) and (26) as

$$\begin{aligned} \hat{\mathbf{J}}_{\theta}^{\text{MAP}} &= \mathbf{A}^T \mathbf{J}_{\Phi}^{\text{MAP}} \mathbf{A} \\ &= E_{\theta} [\hat{\mathbf{J}}_{\theta}^{\text{MLE}}] + \mathbf{C}_{\theta}^{-1} \end{aligned} \quad (27)$$

Note that the matrix $\hat{\mathbf{J}}_{\theta}^{\text{MAP}}$ is not an approximation of $\mathbf{J}_{\theta}^{\text{MAP}}$ and is not used to find the true CRM bounds, which are likely to be very small, if not zero, and therefore useless. For lack of a better name then, $\hat{\mathbf{J}}_{\theta}^{\text{MAP}}$ is termed the *pseudo-MAP Fisher information matrix* for θ .

Lastly, we need to establish that $\hat{\mathbf{J}}_{\theta}^{\text{MAP}}$ has a useful relationship to our estimation error variances. Let $\mathbf{C}_{\theta' - \theta}$ and $\mathbf{C}_{\Phi' - \Phi}$ denote covariance matrices for $\theta' - \theta$ and $\Phi' - \Phi$ respectively. We can get

$$\mathbf{C}_{\phi'-\phi} = \mathbf{A} \cdot \mathbf{C}_{\theta'-\theta} \cdot \mathbf{A}^T \quad (28)$$

It is obvious that off-diagonal elements of $\mathbf{C}_{\theta'-\theta}$ and $\mathbf{C}_{\phi'-\phi}$ are zero. The estimation error variances can be written as

$$\begin{aligned} \text{var}(\phi'_i - \phi_i) &= [\mathbf{C}_{\phi' - \phi}]_{ii} \\ \text{var}(\theta'_i - \theta_i) &= [\mathbf{C}_{\theta' - \theta}]_{ii} \end{aligned} \quad (29)$$

where $[\mathbf{W}]_{ii}$ denotes the i th diagonal element of the matrix \mathbf{W} .

Cramer-Rao bounds for parameter vector $\boldsymbol{\phi}$ are diagonal elements of $(\mathbf{J}_{\boldsymbol{\phi}}^{\text{MAP}})^{-1}$ so

$$\text{var}(\phi'_i - \phi_i) \geq [(\mathbf{J}_{\boldsymbol{\phi}}^{\text{MAP}})^{-1}]_{ii} \quad (30)$$

which can be rewritten as

$$[\mathbf{A} \mathbf{C}_{\theta' - \theta} \mathbf{A}^T]_{ii} \geq [\mathbf{A} (\mathbf{J}_{\boldsymbol{\theta}}^{\text{MAP}})^{-1} \mathbf{A}^T]_{ii} \quad (31)$$

When the bounds in (30) are tight, we can use the fact that $(\mathbf{J}_{\boldsymbol{\phi}}^{\text{MAP}})^{-1} = \mathbf{C}_{\phi' - \phi}$ to write

$$\begin{aligned} \mathbf{A} \mathbf{C}_{\theta' - \theta} \mathbf{A}^T &= \mathbf{A} (\mathbf{J}_{\boldsymbol{\theta}}^{\text{MAP}})^{-1} \mathbf{A}^T \\ \mathbf{C}_{\theta' - \theta} &= (\mathbf{J}_{\boldsymbol{\theta}}^{\text{MAP}})^{-1} \end{aligned} \quad (32)$$

so

$$\text{var}(\theta'_i - \theta_i) \geq [(\mathbf{J}_{\boldsymbol{\theta}}^{\text{MAP}})^{-1}]_{ii} \quad (33)$$

We have now shown that $\mathbf{J}_{\boldsymbol{\theta}}^{\text{MAP}}$ can be used to obtain bounds for non-Gaussian parameter vectors when M is sufficiently large.

In summary, the bounding process has two steps as follows.

(1) A transform matrix \mathbf{B} is found which converts to observation vector \mathbf{y} into an observation vector \mathbf{x} which is approximately Gaussian.

(2) The pseudo-MAP FIM $\mathbf{J}_{\boldsymbol{\theta}}^{\text{MAP}}$ is calculated using $\mathbf{C}_{\boldsymbol{\theta}}$, as if $\boldsymbol{\theta}$ were Gaussian, and the bounds

are found.

Interestingly enough, the transform matrix \mathbf{A} drops out of the derivation and never has to be found. Also, the pseudo-MAP FIM $\mathbf{J}_\theta^{\text{MAP}}$ is the same as the MAP FIM used when the parameter vector θ is Gaussian, so the bounds given in (32) have the limit property. Unlike \mathbf{A} , the matrix \mathbf{B} is more than just a theoretical tool in our derivations.

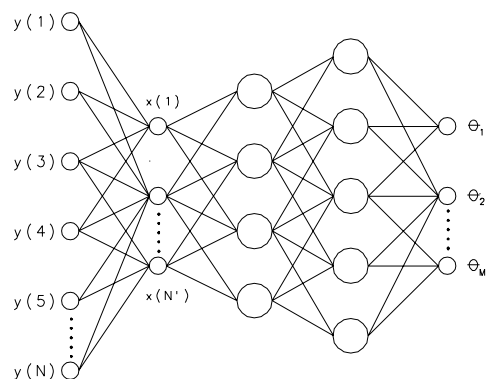


Figure 1. MLP Parameter Estimation Network

Compression of neural network inputs is critical if one wants to speed up training and eliminate useless information [12,15]. A system which uses this compression concept is shown in Fig. 1. In the figure, the input layer weights transform the N -dimensional vector \mathbf{y} into the N' -dimensional vector \mathbf{x} via the matrix \mathbf{B} . The remaining layers process \mathbf{x} into θ' .

V. Example

As an example we generated 5,000 training patterns, each of which had 128 noisy input samples ($N=128$) of the form

$$y(n) = A \cdot e^{-n/\tau} + n(n)$$

where A and τ had uniform pdfs which varied respectively from 1 to 2 and 10 to 20 for τ . The noise $n(n)$ was also uniform with zero-mean and a standard deviation of .1. The 128-sample waveforms, $y(n)$, were first compressed down to N' -sample waveforms $x(n)$ (as in Fig. 1.), using the discrete Karhunen-Loeve transform (KLT) [17], where N' varied from 1 to 8. Assuming that \mathbf{x} and θ were

Gaussian, we calculated the CRM bounds as a function of the number of transform features, N' . MLP neural networks of size N' -20-10-2 were trained via output weight optimization (OWO) [1,3]. In figures 2 and 3, we plot the CRM bounds and neural network training errors versus N' . Also plotted is the final CRM bound for the case where all 128 features are used. Clearly, the bounds provide useful limits on MLP training error, even though the data was not Gaussian. We repeated this example for the discrete Fourier transform (DFT) [17]. Results for A and τ are shown respectively in figures 4 and 5. As with the KLT, the bounds are tight, in spite of the fact that the data is not Gaussian. Figures 2 through 5 also reveal that the KLT is far better for compressing waveforms than the DFT, if the goal is to estimation of exponential parameters.

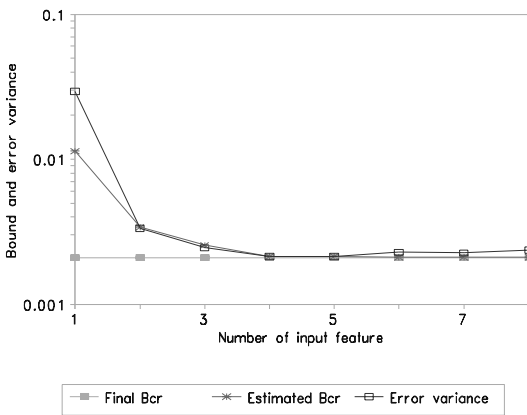


Figure 2. Bounds on $\text{var}(A-A')$ and MLP Training Error for the KLT

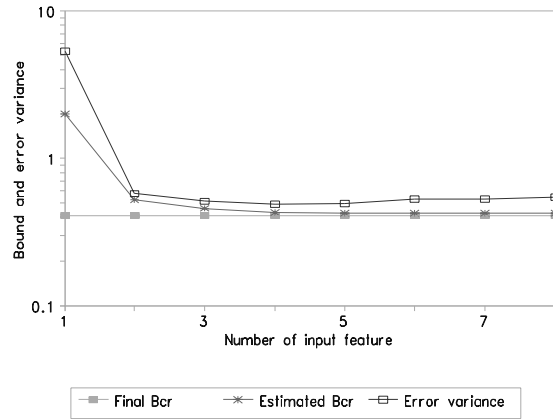


Figure 3. Bounds on $\text{var}(\tau-\tau')$ and MLP Training Error for the KLT

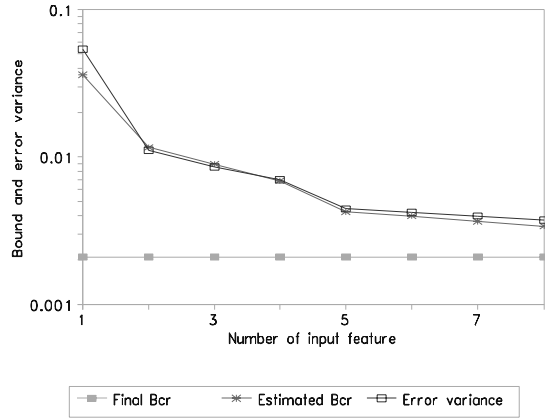


Figure 4. Bounds on $\text{var}(A-A')$ and MLP Training Error for the DFT

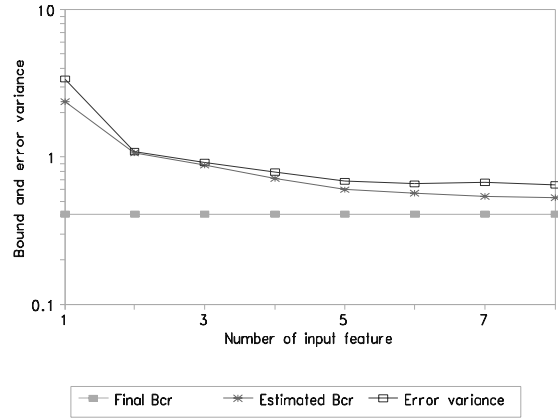


Figure 5. Bounds on $\text{var}(\tau-\tau')$ and MLP Training Error for the DFT

VI. Conclusions

In this paper we have developed CRM bounds on error variance of parameter estimates for the case of non-Gaussian additive noise and non-Gaussian parameters. These bounds have the limit property, which is a necessary condition for tightness. As seen in an example, the bounds can be tight, and can indicate limits on neural network training errors. Much work remains before the CRM bounds can be applied to applications in control and prediction. Most importantly, we must find ways to calculate valid bounds when the deterministic signal model is not bijective.

Acknowledgement

This work was funded by NASA under Grant NAGW-3091, by the NSF under grant IRI-9216545, by EPRI under grant RP 8030-09, and by a grant from the state of Texas.

VII. References

- [1] M.S. Dawson, A.K. Fung, and M.T. Manry, "Surface Parameter Retrieval Using Fast Learning Neural Networks," *Remote Sensing Reviews*, Vol. 7, pp. 1-18, 1993.
- [2] X. Jiang, Mu-Song Chen, M.T. Manry, M.S. Dawson, A.K. Fung, "Analysis and Optimization of Neural Networks for Remote Sensing," *Remote Sensing Reviews*, vol. 9, pp. 97-114, 1994.
- [3] M.T. Manry, S.J. Apollo, L.S. Allen, W.D. Lyle, W. Gong, M.S. Dawson, and A.K. Fung, "Fast Training of Neural Networks for Remote Sensing," *Remote Sensing Reviews*, vol. 9, pp. 77-96, 1994.
- [4] A. Khotanzad, R-C Hwang, and D. Maratukulam, "Hourly Load Forecasting by Neural Networks," *IEEE PES Winter Meeting*, Columbus Ohio, February 1993.
- [5] K. Liu, S. Subbarayan, R.R. Shoults, M.T. Manry C. Kwan, F.L. Lewis, and J. Naccarino, "Comparison of Very Short-Term Load Forecasting Techniques," *IEEE Power Engineering Society Summer Meeting*, Portland, OR, 1995.
- [6] K. Liu, S. Subbarayan, R.R.Shoults, M.T.Manry C.Kwan, F.L.Lewis, and J.Naccarino, "Comparison of Very Short-Term Load Forecasting Techniques," *IEEE Transactions on Power Systems*, to appear.

- [7] K.S. Narendra, "Adaptive Control Using Neural Networks," *Neural Networks for Control*, pp. 115-142, MIT Press, 1991.
- [8] L.G. Kraft and D.P. Campagna, "A Summary Comparison of CMAC Neural Networks and Traditional Adaptive Control Systems," *Neural Networks and Control*, pp. 143-169, MIT Press, 1991.
- [9] F.L. Lewis, K. Liu, and A. Yesildirek, "Neural net robot controller with guaranteed tracking performance," *IEEE Trans. Neural Networks*, pp. 703-710, May 1995.
- [10] S. Jagannathan and F.L. Lewis, "Multilayer discrete-time neural net controller with guaranteed performance," *IEEE Trans. Neural Networks*, to appear, 1995.
- [11] Q. Yu, S.J. Apollo, and M.T. Manry, "MAP Estimation and the Multilayer Perceptron," *Proceedings of the 1993 IEEE Workshop on Neural Networks for Signal Processing*, Linthicum Heights, Maryland, Sept. 6-9, 1993, pp. 30-39.
- [12] M.T. Manry, S.J. Apollo, and Q. Yu, "Minimum Mean Square Estimation and the Multilayer Perceptron," submitted to *Neurocomputing*.
- [13] H. L. Van Trees, *Detection, Estimation, and Modulation Theory - Part I*, New York, NY: John Wiley and Sons, 1968.
- [14] Berry, A.C. "The accuracy of the Gaussian Approximation to the sums of Independent Variables", *Trans. Amer. Math. Soc.* vol. 49, 1941, pp.122-136.
- [15] S.J. Apollo, M.T. Manry, L.S. Allen, and W.D. Lyle, "Optimality of transforms for parameter estimation," *Conference Record of the Twenty-Sixth Annual Asilomar Conference on Signals, Systems, and Computers*, Oct. 1992, vol. 1, pp. 294-298.
- [16] D.E. Rumelhart, G.E. Hinton, and R.J. Williams, "Learning internal representations by error propagation," in D.E. Rumelhart and J.L. McClelland (Eds.), *Parallel Distributed Processing*, Vol. I, Cambridge, Massachusetts: The MIT Press, 1986.
- [17] D.F. Elliot and K.R. Rao, *Fast Transforms: Algorithms, Analyses, Applications*, Academic Press, 1982.