

## ACKNOWLEDGEMENTS

The author thanks Dr. Sumio Watanabe for many valuable discussions, and the anonymous reviewers for their helpful comments.

## REFERENCES

- Ahlfors, L. V. (1966). *Complex Analysis* (pp.101-172). New York, NY: McGraw-Hill.
- Akaike, H. (1974). A new Look at the statistical model identification. *IEEE Transactions on Automatic Control*, **19**(6), 716-723.
- Amari, S. (1985). *Differential-geometrical methods in statistics*. Lecture Notes in Statistics **28**. Springer-Verlag.
- Barron, A. R. (1993). Universal approximation bounds for superpositions of a sigmoidal function. *IEEE Transactions on Information Theory*, **39**, 930-945.
- Chen, A. M., Liu, H., & Hecht-Nielsen, R. (1993). On the geometry of feedforward neural network error surfaces. *Neural Computation*, **5**, 910-927.
- Cramér, H. (1946). *Mathematical method of statistics* (pp.497-506). Princeton, NJ: Princeton University Press.
- Cybenko, G. (1989). Approximation by superpositions of a sigmoidal function. *Mathematics of Control, Signals and Systems*, **2**(4), 303-314.
- Funahashi, K. (1989). On the approximate realization of continuous mapping by neural networks. *Neural Networks*, **2**, 183-192.
- Hagiwara, K., Toda, N., & Usui, S. (1993). On the problem of applying AIC to determine the structure of a layered feed-forward neural network. *Proceedings of 1993 International Joint Conference on Neural Networks*, 2263-2266.
- Hornik, K., Stinchcombe, M., & White, H. (1989). Multi-layer feed-forward networks are universal approximators. *Neural Networks*, **2**, 359-366.
- Kůrková, V., & Kainen, P. C. (1994). Function equivalent feedforward neural networks. *Neural Computation*, **6**, 543-558.
- Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning internal representations by error propagation. In Rumelhart, D.E., McClelland, J.L., & the PDP Research Group (Eds.), *Parallel distributed processing*, Vol.1 (pp.318-362). Cambridge, MA: MIT Press.
- Sussmann, H. J. (1992). Uniqueness of the weights for minimal feedforward nets with a given input-output map. *Neural Networks*, **5**, 589-593.
- Watanabe, S. & Fukumizu, K. (1995). Probabilistic design of layered neural networks based on their unified framework. *IEEE Transactions on Neural Networks*, **6**(3), 691-702.
- White, H. (1989). Learning in artificial neural networks: A statistical perspective. *Neural Computation*, **1**, 425-464.

- (1) there exists  $j$  such that  $\mathbf{u}_j = \mathbf{o}$ .
- (2) there exists  $j$  such that  $w_{1j} = \dots = w_{Mj} = 0$ .
- (3) there exist different  $j_1$  and  $j_2$  such that  $(\mathbf{u}_{j_1}^T, \zeta_{j_1}) = (\mathbf{u}_{j_2}^T, \zeta_{j_2})$  or  $(\mathbf{u}_{j_1}^T, \zeta_{j_1}) = -(\mathbf{u}_{j_2}^T, \zeta_{j_2})$ .

#### Case (1)

We assume  $j = 1$  without loss of generality. In this case,  $\frac{\partial \mathbf{f}}{\partial w_{11}} = (\sigma(\zeta_1), 0, \dots, 0)^T$  and  $\frac{\partial \mathbf{f}}{\partial \eta_1} = (1, 0, \dots, 0)^T$ . Hence

$$-\frac{1}{\sigma(\zeta_1)} \frac{\partial \mathbf{f}}{\partial w_{11}} + \frac{\partial \mathbf{f}}{\partial \eta_1} = \mathbf{o}. \quad (35)$$

#### Case (2)

In this case,  $\frac{\partial \mathbf{f}}{\partial \zeta_j} = \mathbf{o}$ . Evidently, the functions  $\frac{\partial \mathbf{f}}{\partial \theta_a}$  are linearly dependent.

#### Case (3)

We will prove only the case of  $(\mathbf{u}_{j_1}, \zeta_{j_1}) = (\mathbf{u}_{j_2}, \zeta_{j_2})$ . The other case can be proved in the same way. We assume  $j_1 = 1$  and  $j_2 = 2$  without loss of generality. In this case  $\frac{\partial \mathbf{f}}{\partial w_{11}} = \frac{\partial \mathbf{f}}{\partial w_{12}} = (\sigma(\mathbf{u}_1 \cdot \mathbf{x} + \zeta_1), 0, \dots, 0)^T$ . Clearly the functions  $\frac{\partial \mathbf{f}}{\partial \theta_a}$  are linearly dependent. This completes the proof of the main theorem. **Q.E.D.**

The main technique of our proof is to analyze the singularities of the complex function. This method is also applicable to the kinds of networks that have an activation function whose analytic continuation has isolated singularities in  $\mathbf{C}$ . If irreducibility is properly defined for these kinds of networks, an analogy of our main theorem can be obtained in a same fashion.

This method can also be applied to show the fact that the hidden unit functions of an irreducible perceptron network are linearly independent. This fact plays an essential part in the proof of Sussmann's theorem. He showed the independence by expanding the sigmoidal function as an infinite series and considering its coefficients carefully. Our method can replace the discussion on the coefficients with the analysis of the singularities.

## 5 CONCLUSION

We elucidate a useful condition ensuring that the Fisher information matrix of a three-layer perceptron network is positive definite: we show that the irreducibility defined by Sussmann (1992) is equivalent to regularity of the matrix. This implies that if the Fisher information matrix of a three-layer perceptron network is singular, we should search for redundant hidden units and remove them until the network becomes irreducible. This reduction procedure helps us apply many statistical techniques that require regularity of the information matrix.

It is extremely important to elucidate the asymptotic behavior of the estimator when a network in use has redundant hidden units. Our theorem will be helpful in doing this.

Our method essentially depends on complex analytic properties of the sigmoidal function, and it is important to clarify the regularity conditions for other kinds of feed-forward networks. As the example in Section 3 shows, minimality is not always a sufficient condition. Determining the conditions of activation functions in which minimality means regularity of the matrix will be an interesting problem.

From eq.(26) we have  $\Psi_i^{(l)}(t) = 0$  for all  $t \in \mathbf{R}$ . Hence, by Proposition 1,

$$\Psi_i^{(l)}(z) = 0, \quad \text{for } \forall z \in D^{(l)}. \quad (29)$$

Consequently, by Proposition 2 all the points in  $S_j^{(l)}$  are removable singularities.

We assume  $|m_1^{(l)}| \geq |m_2^{(l)}| \geq \dots \geq |m_H^{(l)}| > 0$  without loss of generality. We set

$$p_j^{(l)} := \frac{\pi\sqrt{-1} - \zeta_j}{m_j^{(l)}} \in S_j^{(l)}. \quad (30)$$

First we discuss the singularity at  $p_1^{(l)}$ . In the case that  $|m_1^{(l)}| > |m_2^{(l)}|$ , clearly  $p_1^{(l)}$  is not contained in  $S_j^{(l)}$  for  $j \geq 2$ . Assume  $|m_j^{(l)}| = |m_1^{(l)}|$  for some  $j \geq 2$ . If  $m_j^{(l)} = m_1^{(l)}$ , recalling the fact  $m_j^{(l)} + \zeta_j \neq m_1^{(l)} + \zeta_1$  for  $j \geq 2$ , we have  $\zeta_j \neq \zeta_1$  and therefore,  $p_1^{(l)} \notin S_j^{(l)}$ . Likewise, if  $m_j^{(l)} = -m_1^{(l)}$ , we obtain  $\zeta_j \neq -\zeta_1$  and  $p_1^{(l)} \notin S_j^{(l)}$ . Thus we conclude  $p_1^{(l)} \notin S_j^{(l)}$  for  $j \geq 2$ .

We rewrite  $\Psi_i^{(l)}(z)$  as

$$\Psi_i^{(l)}(z) = w_{i1} \left( \sum_{k=1}^L \beta_{1k} x_k^{(l)} z + \beta_{10} \right) \sigma'(m_1^{(l)} z + \zeta_1) + \alpha_{i1} \sigma(m_1^{(l)} z + \zeta_1) + \phi_{i,2}^{(l)}(z), \quad (31)$$

where

$$\phi_{i,2}^{(l)}(z) = \alpha_{i0} + \sum_{j \geq 2} \alpha_{ij} \sigma(m_j^{(l)} z + \zeta_j) + \sum_{j \geq 2} w_{ij} \left( \sum_{k=1}^L \beta_{jk} x_k^{(l)} z + \beta_{j0} \right) \sigma'(m_j^{(l)} z + \zeta_j). \quad (32)$$

The point  $p_1^{(l)}$  is a regular point of  $\phi_{i,2}^{(l)}(z)$ , while  $\sigma(m_1^{(l)} z + \zeta_1)$  has a pole of order 1 at  $p_1^{(l)}$  and  $\sigma'(m_1^{(l)} z + \zeta_1)$  has a pole of order 2 at  $p_1^{(l)}$ . Since  $p_1^{(l)}$  is a removable singularity of  $\Psi_i^{(l)}(z)$  and  $w_{i1} \left( \sum_k \beta_{1k} x_k^{(l)} z + \beta_{10} \right)$  does not have an isolated zero point at  $p_1^{(l)}$ , we have

$$\begin{aligned} w_{i1} \sum_{k=1}^L \beta_{1k} x_k^{(l)} &= w_{i1} \beta_{10} = 0, \\ \alpha_{i1} &= 0. \end{aligned} \quad (33)$$

As a result, we have  $\Psi_i^{(l)}(z) = \phi_{i,2}^{(l)}(z)$ . Applying the same argument successively to  $p_2^{(l)}, p_3^{(l)}, \dots$ , we finally obtain

$$\begin{aligned} w_{ij} \sum_{k=1}^L \beta_{jk} x_k^{(l)} &= 0, & \forall i, \forall j, \forall l, \\ w_{ij} \beta_{j0} &= 0, & \forall i, \forall j, \\ \alpha_{ij} &= 0, & \forall i, \forall j, \\ \alpha_{i0} &= 0, & \forall i. \end{aligned} \quad (34)$$

By the assumption of irreducibility, for any  $j$  there exists  $i$  such that  $w_{ij} \neq 0$ . Picking such  $i$ , we obtain  $\sum_k \beta_{jk} x_k^{(l)} = 0$  and  $\beta_{j0} = 0$ . Since  $\langle \mathbf{x}^{(1)}, \dots, \mathbf{x}^{(L)} \rangle$  form a basis of  $\mathbf{R}^L$ , we have  $\beta_{jk} = 0$ . Now we conclude that all the linear coefficients of eq.(23) are zero.

[  $\Rightarrow$  ] We will prove that if  $\mathbf{f}(\cdot; \boldsymbol{\theta})$  is reducible, then the functions  $\frac{\partial \mathbf{f}}{\partial \theta_i}$  are linearly dependent. By the definition of reducibility, one of the following conditions is satisfied:

## Proof of the main theorem

[  $\Leftarrow$  ] Let  $\mathbf{f}(\cdot; \boldsymbol{\theta})$  be an irreducible three-layer perceptron network. From Lemma 3, there exists a basis of  $\mathbf{R}^L$ ,  $\langle \mathbf{x}^{(1)}, \dots, \mathbf{x}^{(L)} \rangle$ , that satisfies the conditions stated in the lemma. By lemma 2, we have to prove if

$$\sum_{i=1}^M \sum_{j=1}^H \alpha_{ij} \frac{\partial \mathbf{f}}{\partial w_{ij}} + \sum_{i=1}^M \alpha_{i0} \frac{\partial \mathbf{f}}{\partial \eta_i} + \sum_{j=1}^H \sum_{k=1}^L \beta_{jk} \frac{\partial \mathbf{f}}{\partial u_{jk}} + \sum_{j=1}^H \beta_{j0} \frac{\partial \mathbf{f}}{\partial \zeta_j} = \mathbf{o}, \quad (23)$$

then all the coefficients,  $\alpha_{ij}$ ,  $\alpha_{i0}$ ,  $\beta_{jk}$ , and  $\beta_{j0}$ , are zero.

First we calculate the derivatives of  $\mathbf{f}$ :

$$\begin{aligned} \frac{\partial f^{i_1}}{\partial w_{i_2 j}} &= \delta_{i_1 i_2} \sigma(\mathbf{u}_j \cdot \mathbf{x} + \zeta_j), \\ \frac{\partial f^{i_1}}{\partial \eta_{i_2}} &= \delta_{i_1 i_2}, \\ \frac{\partial f^i}{\partial u_{jk}} &= w_{ij} \sigma'(\mathbf{u}_j \cdot \mathbf{x} + \zeta_j) x_k, \\ \frac{\partial f^i}{\partial \zeta_j} &= w_{ij} \sigma'(\mathbf{u}_j \cdot \mathbf{x} + \zeta_j), \end{aligned} \quad (24)$$

where  $\delta_{ab}$  denotes Kronecker's delta. Hence, for  $1 \leq i \leq M$ ,

$$\begin{aligned} \sum_{j=1}^H \alpha_{ij} \sigma(\mathbf{u}_j \cdot \mathbf{x} + \zeta_j) + \alpha_{i0} + \sum_{j=1}^H \sum_{k=1}^L \beta_{jk} w_{ij} \sigma'(\mathbf{u}_j \cdot \mathbf{x} + \zeta_j) x_k \\ + \sum_{j=1}^H \beta_{j0} w_{ij} \sigma'(\mathbf{u}_j \cdot \mathbf{x} + \zeta_j) = 0. \end{aligned} \quad (25)$$

We substitute  $\mathbf{x}^{(l)} t$  ( $t \in \mathbf{R}$ ) for  $\mathbf{x}$  in eq.(25), and use the notation  $m_j^{(l)} := \mathbf{u}_j \cdot \mathbf{x}^{(l)}$ . We have  $m_j^{(l)} \neq 0$  and  $m_{j_1}^{(l)} + \zeta_{j_1} \neq \pm(m_{j_2}^{(l)} + \zeta_{j_2})$  for  $j_1 \neq j_2$ . Then, for all  $t \in \mathbf{R}$

$$\begin{aligned} \sum_{j=1}^H \alpha_{ij} \sigma(m_j^{(l)} t + \zeta_j) + \alpha_{i0} + \sum_{j=1}^H \sum_{k=1}^L \beta_{jk} w_{ij} \sigma'(m_j^{(l)} t + \zeta_j) x_k^{(l)} t \\ + \sum_{j=1}^H \beta_{j0} w_{ij} \sigma'(m_j^{(l)} t + \zeta_j) = 0, \quad (1 \leq i \leq M, 1 \leq l \leq L). \end{aligned} \quad (26)$$

We fix  $l$  for a while. We set

$$S_j^{(l)} := \{z \in \mathbf{C} \mid z = \frac{(2n+1)\pi\sqrt{-1} - \zeta_j}{m_j^{(l)}}, n \in \mathbf{Z}\}. \quad (27)$$

Clearly the points in  $S_j^{(l)}$  are the singularities of  $\sigma(m_j^{(l)} z + \zeta_j)$ . Let  $D^{(l)} := \mathbf{C} - \cup_j S_j^{(l)}$ . Note that  $\mathbf{R} \subset D^{(l)}$ . We define holomorphic functions on  $D^{(l)}$  as follows:

$$\begin{aligned} \Psi_i^{(l)}(z) &:= \sum_{j=1}^H \alpha_{ij} \sigma(m_j^{(l)} z + \zeta_j) + \alpha_{i0} + \sum_{j=1}^H \sum_{k=1}^L \beta_{jk} w_{ij} \sigma'(m_j^{(l)} z + \zeta_j) x_k^{(l)} z \\ &+ \sum_{j=1}^H \beta_{j0} w_{ij} \sigma'(m_j^{(l)} z + \zeta_j), \quad (1 \leq i \leq M). \end{aligned} \quad (28)$$

We define the complex sigmoidal function on  $\mathbf{C}$  by

$$\sigma(z) = \frac{1}{1 + e^{-z}}, \quad (z \in \mathbf{C}). \quad (22)$$

The singularities of  $\sigma$  are  $\{z \in \mathbf{C} \mid z = (2n + 1)\pi\sqrt{-1}, n \in \mathbf{Z}\}$ , all of which are poles of order 1. This is clear from the fact that the set of all the zeros of  $1 + e^{-z}$  is  $\{(2n + 1)\pi\sqrt{-1} \mid n \in \mathbf{Z}\}$ , and each of them is of order 1.

Next we review fundamental propositions in complex analysis.

**Proposition 1** *Let  $f$  be a holomorphic function on a connected open set  $D$  in  $\mathbf{C}$ , and  $p$  be a point in  $D$ . If there exists a sequence  $\{p_n\}_{n=1}^{\infty}$  in  $D$  such that  $p_n \neq p$ ,  $\lim_{n \rightarrow \infty} p_n = p$ , and  $f(p_n) = 0$  for all  $n \in \mathbf{N}$ , then  $f(z) = 0$  for all  $z \in D$ .*

(For the proof, see Ahlfors 1966.)

**Proposition 2** *Let  $f$  be a holomorphic function which has an isolated singularity at  $p$ . Then the following equivalence relations hold:*

- (1)  $p$  is a removable singularity  $\Leftrightarrow \lim_{\substack{z \rightarrow p \\ z \neq p}} f(z) \in \mathbf{C}$ .
- (2)  $p$  is a pole  $\Leftrightarrow \lim_{\substack{z \rightarrow p \\ z \neq p}} |f(z)| = \infty$ .
- (3)  $p$  is an essential singularity  $\Leftrightarrow \lim_{\substack{z \rightarrow p \\ z \neq p}} f(z)$  does not exist and  $\lim_{\substack{z \rightarrow p \\ z \neq p}} |f(z)| \neq \infty$ .

(See Ahlfors 1966.)

We prepare a lemma on three-layer perceptron networks for the proof of the main theorem.

**Lemma 3** *Let  $\mathbf{f}(\cdot; \boldsymbol{\theta})$  be an irreducible three-layer perceptron network. Then there exists a basis of  $\mathbf{R}^L$ ,  $\langle \mathbf{x}^{(1)}, \dots, \mathbf{x}^{(L)} \rangle$ , that satisfies the following conditions:*

- (1)  $\mathbf{u}_j \cdot \mathbf{x}^{(l)} \neq 0$  for all  $j$  and  $l$ .
- (2)  $\mathbf{u}_{j_1} \cdot \mathbf{x}^{(l)} + \zeta_{j_1} \neq \pm(\mathbf{u}_{j_2} \cdot \mathbf{x}^{(l)} + \zeta_{j_2})$  for  $j_1 \neq j_2$  and for all  $l$ .

**[Proof]** For  $1 \leq j \leq H$ , we set

$$C_j := \{(\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(L)}) \in \mathbf{R}^{L \times L} \mid 1 \leq l \leq L, \mathbf{u}_j \cdot \mathbf{x}^{(l)} = 0\},$$

and for  $j_1 \neq j_2$ ,

$$B_{j_1 j_2} := \{(\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(L)}) \in \mathbf{R}^{L \times L} \mid 1 \leq l \leq L, (\mathbf{u}_{j_1} - \mathbf{u}_{j_2}) \cdot \mathbf{x}^{(l)} + (\zeta_{j_1} - \zeta_{j_2}) = 0, \text{ or } (\mathbf{u}_{j_1} + \mathbf{u}_{j_2}) \cdot \mathbf{x}^{(l)} + (\zeta_{j_1} + \zeta_{j_2}) = 0\}.$$

By the assumption of irreducibility,  $C_j$  and  $B_{j_1 j_2}$  is a union of a finite number of hyperplanes in  $\mathbf{R}^{L \times L}$ . Let  $U := \{(\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(L)}) \in \mathbf{R}^{L \times L} \mid \mathbf{x}^{(1)}, \dots, \mathbf{x}^{(L)} \text{ are linearly independent}\}$ .  $U$  is an open dense set in  $\mathbf{R}^{L \times L}$ . Thus  $U - ((\cup_j C_j) \cup (\cup_{j_1 \neq j_2} B_{j_1 j_2}))$  is an open dense set in  $\mathbf{R}^{L \times L}$ . Choose  $(\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(L)}) \in U - ((\cup_j C_j) \cup (\cup_{j_1 \neq j_2} B_{j_1 j_2}))$ , then they form a basis of  $\mathbf{R}^L$  that satisfy the stated conditions. **Q.E.D.**

the minimality of a network does not always imply the regularity of the matrix. We do not yet know the characterization of activation functions that makes our main theorem hold.

### Example

Let  $X$  and  $Y$  be one-dimensional. For an activation function we set

$$\rho(t) = t + \exp(-t^2). \quad (16)$$

We investigate feed-forward networks with 2 hidden units and the activation function  $\rho$ . We write

$$f(x; \boldsymbol{\theta}) = \sum_{j=1}^2 w_j \rho(u_j x + \zeta_j) + \eta, \quad (17)$$

where  $\boldsymbol{\theta} = (w_1, w_2, \eta, u_1, u_2, \zeta_1, \zeta_2)$ . We define the parameter  $\boldsymbol{\theta}_0$  as follows:

$$\begin{aligned} w_1 &= w_2 = 1, \\ u_1 &= 1, \\ u_2 &= -1, \\ \zeta_1 &= \zeta_2 = \eta = 0. \end{aligned} \quad (18)$$

Then we obtain

$$\begin{aligned} f(x; \boldsymbol{\theta}_0) &= \rho(x) + \rho(-x) \\ &= 2 \exp(-x^2). \end{aligned} \quad (19)$$

In this case,  $f(\cdot; \boldsymbol{\theta}_0)$  is minimal, yet the Fisher information matrix is singular at  $\boldsymbol{\theta}_0$ . To prove that the network is minimal, suppose  $f(\cdot; \boldsymbol{\theta}_0)$  is written by a network with 1 hidden unit, that is,  $f(x; \boldsymbol{\theta}_0) = \alpha \rho(ax + b) + \beta$ . Then we have  $\alpha(ax + b + \exp\{-(ax + b)^2\}) + \beta = 2 \exp(-x^2)$ . Considering the limit for  $x \rightarrow \infty$ , we know  $\alpha a = 0$ . If  $\alpha = 0$ , then  $\beta = 2 \exp(-x^2)$ , and if  $a = 0$ , then  $\alpha(b + \exp(-b^2)) = \exp(-x^2)$ . Since both cases never happen, we conclude  $f(\cdot; \boldsymbol{\theta}_0)$  is minimal. Next, we have

$$\begin{aligned} \frac{\partial f(x; \boldsymbol{\theta}_0)}{\partial \zeta_1} &= 1 - 2x \exp(-x^2), \\ \frac{\partial f(x; \boldsymbol{\theta}_0)}{\partial \zeta_2} &= 1 + 2x \exp(-x^2), \\ \frac{\partial f(x; \boldsymbol{\theta}_0)}{\partial \eta} &= 1. \end{aligned} \quad (20)$$

Hence

$$\frac{\partial f(x; \boldsymbol{\theta}_0)}{\partial \zeta_1} + \frac{\partial f(x; \boldsymbol{\theta}_0)}{\partial \zeta_2} - 2 \frac{\partial f(x; \boldsymbol{\theta}_0)}{\partial \eta} = 0. \quad (21)$$

This implies that the Fisher information matrix is singular.

## 4 MATHEMATICAL PROOFS

In this section we will give a complete proof of the main theorem. The main idea of our proof is to consider the analytic continuation of the linear combination of hidden unit functions. It has much analytic information in its singularities.

Since  $\mathbf{f}$  and  $\mathbf{g}$  are both irreducible, the mapping  $c$  is necessarily injective. This contradicts the assumption  $H' < H$  and completes the proof. **Q.E.D.**

Irreducibility is defined by excepting three simple situations under which a network can be made smaller. The above theorem asserts that no other situations can reduce the number of hidden units without changing the input-output map. The concept of irreducibility thus gives us an important viewpoint for analyzing redundancy of network parameters.

### 3 MAIN THEOREM

The main result of this paper is as follows:

**Theorem 1** *Assume that the input density  $q$  is positive and continuous. Then the Fisher information matrix of a three-layer perceptron network is positive definite if and only if the network is irreducible.*

Combining this theorem with Sussmann's result, we have

**Corollary 1** *Assume that the input density  $q$  is positive and continuous. Then the Fisher information of a three-layer perceptron network is positive definite if and only if the network is minimal.*

Statistical theories assuming the regularity of an information matrix are sometimes applied improperly to multilayer perceptron networks, which can have a singular information matrix. The above theorem suggests a practical method of avoiding this: if one finds the Fisher information matrix singular, one should search for redundant hidden units and remove them until the network becomes irreducible.

In this paper, we discuss networks with the identity output-unit activation function. However, the result for these networks can be easily extended to networks with the sigmoidal output-unit activation function. To see this, let  $\mathbf{f}(\cdot; \boldsymbol{\theta})$  be a multi-layer perceptron network, and define  $\mathbf{g}(\cdot; \boldsymbol{\theta})$  by

$$g^i(\mathbf{x}; \boldsymbol{\theta}) := \sigma(f^i(\mathbf{x}; \boldsymbol{\theta})), \quad (1 \leq i \leq M). \quad (14)$$

Then we have

$$\frac{\partial g^i(\mathbf{x}; \boldsymbol{\theta})}{\partial \theta_a} = \sigma'(f^i(\mathbf{x}; \boldsymbol{\theta})) \frac{\partial f^i(\mathbf{x}; \boldsymbol{\theta})}{\partial \theta_a}. \quad (15)$$

Because  $\sigma'(t)$  is positive, it is easy to obtain that  $\frac{\partial g^i}{\partial \boldsymbol{\theta}_a}$  are linearly independent if and only if  $\frac{\partial f^i}{\partial \boldsymbol{\theta}_a}$  are linearly independent. Since lemma 1 and 2 hold also for  $\mathbf{g}(\cdot; \boldsymbol{\theta})$ , the Fisher information matrix of  $\mathbf{g}(\cdot; \boldsymbol{\theta})$  is regular if and only if that of  $\mathbf{f}(\cdot; \boldsymbol{\theta})$  is regular. Thus we can concentrate our discussion on networks with the identity output-unit activation function.

It is a very interesting problem to elucidate the regularity conditions for general feed-forward networks. Intuitively, it seems natural that if a network is minimal, the information matrix of the network is regular. As the following example shows, however, for some activation functions

### 2.3 Functionally Equivalent Networks

Sussmann (1992) shows a condition that ensures two three-layer perceptron networks have the same input-output map, using the concept of minimality and irreducibility. This idea is relevant to our main theorem. We state a slight extension of his theorem here.

**Theorem [Sussmann]** *A three-layer perceptron network is irreducible if and only if it is minimal.*

In the case of a network with one-dimensional output and tanh as the activation function, the theorem was proved by Sussmann (1992, Corollary). Using Sussmann's result, we can easily prove the above theorem in the general case.

**[Proof]** First, note that  $\sigma(t) = (1 + \tanh(t/2))/2$ . We define a map from the parameter space of three-layer perceptron networks to that of tanh networks by

$$\begin{aligned} w_{ij} &\mapsto \frac{1}{2}w_{ij}, & \eta_i &\mapsto \frac{1}{2}(\eta_i + w_{i1} + w_{i2} + \dots + w_{iH}), \\ u_{jk} &\mapsto \frac{1}{2}u_{jk}, & \zeta_j &\mapsto \frac{1}{2}\zeta_j. \end{aligned} \quad (10)$$

This gives one-to-one correspondence of two kinds of networks that preserves irreducibility and minimality. It is clear that Sussmann's result holds also for three-layer perceptron networks with a one-dimensional output unit and the sigmoidal hidden-unit activation function.

The proof that a minimal network is irreducible goes in exactly the same way as the one-dimensional case in Sussmann (1992), so we omit it. Let  $F_{[i]} := \{j \mid 1 \leq j \leq H, w_{ij} \neq 0\}$ . For each  $i$  we define a three-layer perceptron network  $f_{[i]}(\cdot; \boldsymbol{\theta})$  by

$$f_{[i]}(\mathbf{x}; \boldsymbol{\theta}_{[i]}) = \sum_{j \in F_{[i]}} w_{ij} \sigma(\mathbf{u}_j \cdot \mathbf{x} + \zeta_j) + \eta_i, \quad (11)$$

where  $\boldsymbol{\theta}_{[i]} = \{(w_{ij}, \eta_i, u_{jk}, \zeta_j) \mid j \in F_{[i]}\}$ . This is the  $i$ th underlying subnetwork with a single output unit and  $\#F_{[i]}$  hidden units.

Suppose  $\mathbf{f}(\cdot; \boldsymbol{\theta})$  is not minimal, then there exists a network  $\mathbf{g}(\cdot; \boldsymbol{\omega})$  with  $H'$  hidden units ( $H' < H$ ) such that  $\mathbf{f}(\mathbf{x}; \boldsymbol{\theta}) = \mathbf{g}(\mathbf{x}; \boldsymbol{\omega})$  for all  $\mathbf{x} \in X$ . We can assume  $\mathbf{g}$  is irreducible by eliminating redundant hidden units if necessary. We set

$$g_{[i]}(\mathbf{x}; \boldsymbol{\omega}) = \sum_{h=1}^{H'} a_{ih} \sigma(\mathbf{b}_h \cdot \mathbf{x} + \beta_h) + \gamma_i. \quad (12)$$

We define  $G_{[i]} := \{h \mid 1 \leq h \leq H', a_{ih} \neq 0\}$  and the underlying subnetworks  $g_{[i]}(\mathbf{x}; \boldsymbol{\omega}_{[i]})$  in the same way as  $f_{[i]}$ . Note that all the underlying subnetworks are irreducible if the original network is irreducible.

We have  $f_{[i]} = g_{[i]}$ , then by the sigmoidal version of Sussmann's result on single output nets (Sussmann, 1992, Theorem 1), there exists a one-to-one correspondence,  $c_i : F_{[i]} \rightarrow G_{[i]}$ , such that  $(\mathbf{u}_j, \zeta_j) = \pm(\mathbf{b}_{c_i(j)}, \beta_{c_i(j)})$ . If  $j$  is contained in different  $F_{[i]}$  and  $F_{[i']}$ , the corresponding hidden units  $c_i(j)$  and  $c_{i'}(j)$  must be identical since the original network  $\mathbf{g}$  is irreducible. Therefore, the following function is well defined.

$$\begin{aligned} c : \{1, \dots, H\} &\longrightarrow \{1, \dots, H'\}, \\ c(j) &= c_i(j) \quad \text{if } j \in F_{[i]}. \end{aligned} \quad (13)$$

We assume that Gaussian noise is added to the output. It is well known (Watanabe & Fukumizu, 1995) that the maximum likelihood estimator under the Gaussian noise assumption is equal to the least-squares estimator. Therefore, the above definition of the probability density is natural when we consider feed-forward neural networks from the statistical viewpoint.

We show simple lemmas to characterize the regularity of a Fisher information matrix.

**Lemma 1** *Let  $q$  be a probability density function on  $X$ , and  $V$  be a  $M \times M$  positive definite symmetric matrix. Let  $\mathbf{f}(\cdot; \boldsymbol{\theta})$  be a feed-forward network with an activation function of class  $C^1$ . Then  $I(\boldsymbol{\theta})$ , the Fisher information matrix of the network  $\mathbf{f}(\cdot; \boldsymbol{\theta})$  with respect to  $q$  and  $V$ , is given by*

$$I_{ab}(\boldsymbol{\theta}) = \int_X I_{ab}(\mathbf{x}; \boldsymbol{\theta}) q(\mathbf{x}) d\mathbf{x}, \quad (6)$$

where

$$I_{ab}(\mathbf{x}; \boldsymbol{\theta}) = \left( \frac{\partial \mathbf{f}(\mathbf{x}; \boldsymbol{\theta})}{\partial \theta_a} \right)^T V^{-1} \left( \frac{\partial \mathbf{f}(\mathbf{x}; \boldsymbol{\theta})}{\partial \theta_b} \right). \quad (7)$$

[Proof] Integrating on  $\mathbf{y}$ , we easily obtain the above formula.

**Q.E.D.**

**Lemma 2** *Let  $\mathbf{f}(\cdot; \boldsymbol{\theta})$  be a feed-forward network with an activation function of class  $C^1$ . Let  $q$  be a probability density function on  $X$ , and  $V$  be a  $M \times M$  positive definite symmetric matrix. If  $q$  is continuous and  $q(\mathbf{x}) > 0$  for all  $\mathbf{x} \in X$ , then  $I(\boldsymbol{\theta})$ , the Fisher information matrix of the network  $\mathbf{f}(\cdot; \boldsymbol{\theta})$  w.r.t.  $q$  and  $V$ , is strictly positive definite if and only if the vector valued functions  $\frac{\partial \mathbf{f}(\mathbf{x}; \boldsymbol{\theta})}{\partial \theta_a}$  ( $1 \leq a \leq H(M+1) + L(H+1)$ ) are linearly independent over  $\mathbf{R}$ .*

[Proof] It suffices to show that  $I(\boldsymbol{\theta})$  is not positive definite if and only if the derivatives are linearly dependent. The matrix  $I(\boldsymbol{\theta})$  is not positive definite if and only if there exists a non-zero vector  $(v^a)$  that satisfies

$$\sum_a \sum_b v^a I_{ab}(\boldsymbol{\theta}) v^b = 0. \quad (8)$$

Using lemma 1, the left hand side of eq.(8) is calculated as

$$\sum_a \sum_b v^a I_{ab}(\boldsymbol{\theta}) v^b = \int_X \left( \sum_a v^a \frac{\partial \mathbf{f}(\mathbf{x}; \boldsymbol{\theta})}{\partial \theta_a} \right)^T V^{-1} \left( \sum_b v^b \frac{\partial \mathbf{f}(\mathbf{x}; \boldsymbol{\theta})}{\partial \theta_b} \right) q(\mathbf{x}) d\mathbf{x}. \quad (9)$$

By the assumption that  $q$  is continuous and positive and that  $V$  is positive definite, we easily conclude that  $I(\boldsymbol{\theta})$  is singular if and only if the derivatives are linearly dependent. **Q.E.D.**

**Remark:** From this lemma, we know that regularity of the information matrix does not depend on a choice of  $V$  and  $q$  if  $q$  is continuous and positive. Hereafter, we assume  $q$  is always positive and continuous, and we omit  $V$  and  $q$  when we discuss the regularity.

Although multilayer perceptron networks have been analyzed from the statistical viewpoint (White, 1989; Watanabe & Fukumizu 1994), the regularity conditions of the Fisher information matrix have not been clarified yet. We will present a useful condition in the next section.

**Definition 2** Let  $\mathbf{f}(\cdot; \boldsymbol{\theta})$  be a three-layer feed-forward network with  $H$  hidden units. We call  $\mathbf{f}(\cdot; \boldsymbol{\theta})$  minimal if no networks with fewer hidden units have the same input-output map as  $\mathbf{f}(\cdot; \boldsymbol{\theta})$ ; that is, if for any network  $\mathbf{g}(\cdot; \boldsymbol{\omega})$  with  $H'$  hidden units ( $H' < H$ ) and the same activation function there exists  $\mathbf{x} \in X$  such that  $\mathbf{f}(\mathbf{x}; \boldsymbol{\theta}) \neq \mathbf{g}(\mathbf{x}; \boldsymbol{\omega})$ .

In this paper,  $\sigma(t)$  denotes the sigmoidal function defined by

$$\sigma(t) = \frac{1}{1 + e^{-t}}. \quad (3)$$

We define a *three-layer perceptron network* as a three-layer feed-forward network with the activation function  $\sigma$ . We consider only networks with the identity output-unit activation function. However, our main theorem is applicable to networks with the sigmoidal output-unit activation function, as we explain later.

We slightly extend Sussmann's definition of irreducibility. Sussmann (1992) discussed networks with a single output unit. We define irreducibility for a multilayer perceptron network with multiple output units.

**Definition 3** A three-layer perceptron network with  $H$  hidden units is said to be irreducible if it satisfies the following three conditions:

- (1)  $\mathbf{u}_j \neq \mathbf{0}$ , for  $1 \leq j \leq H$ .
- (2)  $(w_{1j}, \dots, w_{Mj})^T \neq \mathbf{0}^T$ , for  $1 \leq j \leq H$ .
- (3) For any two different indices  $j_1$  and  $j_2$ ,  $(\mathbf{u}_{j_1}^T, \zeta_{j_1}) \neq \pm(\mathbf{u}_{j_2}^T, \zeta_{j_2})$ .

## 2.2 Fisher Information Matrix

Next we discuss the Fisher information matrix of a neural network.

**Definition 4** Let  $\Theta$  be a domain in  $\mathbf{R}^S$ . Let  $\{p(\mathbf{z}; \boldsymbol{\theta}) \mid \mathbf{z} \in \mathbf{R}^n, \boldsymbol{\theta} \in \Theta\}$  be a family of probability density functions of class  $C^1$  on  $\boldsymbol{\theta}$ . Then, the matrix

$$I_{ab}(\boldsymbol{\theta}) = \int_{\mathbf{R}^n} \frac{\partial \log p(\mathbf{z}; \boldsymbol{\theta})}{\partial \theta_a} \frac{\partial \log p(\mathbf{z}; \boldsymbol{\theta})}{\partial \theta_b} p(\mathbf{z}; \boldsymbol{\theta}) d\mathbf{x}, \quad (1 \leq a, b \leq S) \quad (4)$$

is called the Fisher information matrix at  $\boldsymbol{\theta} \in \Theta$ .

A symmetric  $n \times n$  matrix  $A = (a_{ij})$  is called *semi-positive definite* if  $\mathbf{x}^T A \mathbf{x} \geq 0$  for all  $\mathbf{x} \in \mathbf{R}^n$ , and called (*strictly*) *positive definite* if  $\mathbf{x}^T A \mathbf{x} > 0$  for any non-zero vector  $\mathbf{x}$ . A Fisher information matrix is necessarily semi-positive definite but is not always positive definite. If the Fisher information matrix  $I(\boldsymbol{\theta}_0)$  at the true parameter  $\boldsymbol{\theta}_0$  is positive definite, it essentially determines the asymptotic behavior of the maximum likelihood estimator  $\hat{\boldsymbol{\theta}}_N$ , where  $N$  is the number of data. Asymptotic theory shows that the limiting distribution of  $\sqrt{N}(\hat{\boldsymbol{\theta}}_N - \boldsymbol{\theta}_0)$  is the multivariate normal distribution with a covariance matrix  $I(\boldsymbol{\theta}_0)^{-1}$ . Therefore, it is very important to elucidate the regularity conditions of Fisher information matrixes.

**Definition 5** Let  $\mathbf{f}(\cdot; \boldsymbol{\theta})$  be a feed-forward network. Let  $q$  be a probability density function on  $X$ , and  $V$  be a  $M \times M$  positive definite symmetric matrix. The Fisher information matrix of the network  $\mathbf{f}(\cdot; \boldsymbol{\theta})$  with respect to  $q$  and  $V$  is defined by the Fisher information matrix of the family of densities  $\{p(\mathbf{x}, \mathbf{y}; \boldsymbol{\theta})\}$  at  $\boldsymbol{\theta}$ , where

$$p(\mathbf{x}, \mathbf{y}; \boldsymbol{\theta}) = \frac{1}{(2\pi)^{M/2} |V|^{1/2}} \exp\left\{-\frac{1}{2}(\mathbf{y} - \mathbf{f}(\mathbf{x}; \boldsymbol{\theta}))^T V^{-1}(\mathbf{y} - \mathbf{f}(\mathbf{x}; \boldsymbol{\theta}))\right\} q(\mathbf{x}). \quad (5)$$

matrix of the parametrized distribution essentially determines the asymptotic behavior of the estimator (Cramér, 1946). This matrix works as a Riemannian metric in information geometry (Amari, 1985) and plays an essential part in the derivation of Akaike’s Information Criterion (AIC, Akaike, 1974). A Fisher information matrix is necessarily semi-positive definite by definition, but is not always regular or positive definite. Many of the statistical techniques based on the asymptotic theory, however, require the matrix to be positive definite. In many linear problems, such as polynomial approximation, the Fisher information matrixes are always positive definite under general conditions. In multilayer perceptron networks, however, the Fisher information matrix can be singular, and therefore, AIC cannot always be applied properly (Hagiwara et al. 1993). Thus it is essential to elucidate the conditions in which a multilayer perceptron network has a positive definite Fisher information matrix.

Several investigators have recently studied equi-output conditions for multilayer perceptron networks (Sussmann, 1992; Chen et al., 1993; Kůrková, 1994) and their results give hints about how to analyze the regularity conditions of a Fisher information matrix. Sussmann (1992) introduces the concept of minimality and irreducibility, and clarifies the condition in which two three-layer perceptron networks have the same input-output map. He calls a network “minimal” if its input-output map cannot be obtained from another network with fewer hidden units. On the other hand, “reducibility” is defined by listing three very simple situations under which a network could be made smaller, and calling a network “irreducible” if none of these situations occurs. Sussmann proves that irreducibility is equivalent to minimality, giving us an insight into the effectiveness of network parameters.

Our main result is a theorem asserting that the Fisher information matrix of a three-layer perceptron network is positive definite if and only if the network is irreducible; that is, if it is minimal. From this theorem we know that a network with a singular Fisher information matrix can be reduced to one with a positive definite Fisher Information matrix by removing redundant hidden units. We prove the theorem in a rigorously mathematical way with the help of some fundamental propositions in complex analysis.

This paper is organized as follows. Section 2 introduces the required definitions and simple lemmas, Section 3 presents our main theorem and explains its meaning, and Section 4 mathematically proves the main theorem. Section 5 concludes with a brief summary, a statement of the significance of this work, and an indication of related problems that remain to be studied.

## 2 DEFINITIONS AND PRELIMINARIES

### 2.1 Definition of Neural Networks

Let  $X = \mathbf{R}^L$  and  $Y = \mathbf{R}^M$ . Let  $\Theta = \{\boldsymbol{\theta} = (w_{11}, \dots, w_{MH}, \eta_1, \dots, \eta_M, u_{11}, \dots, u_{HL}, \zeta_1, \dots, \zeta_H) \in \mathbf{R}^{M(H+1)+H(L+1)}\}$ . In the following discussion, we write

$$\mathbf{u}_j := (u_{j1}, u_{j2}, \dots, u_{jL})^T. \quad (1)$$

**Definition 1** Let  $X$ ,  $Y$ , and  $\Theta$  be as above. Let  $\rho$  be a function on  $\mathbf{R}$ . A three-layer feed-forward network with  $H$  hidden units and the activation function  $\rho$  is a parametrized function  $\mathbf{f}(\cdot; \boldsymbol{\theta}) : X \rightarrow Y$ , ( $\boldsymbol{\theta} \in \Theta$ ) defined by

$$f^i(\mathbf{x}; \boldsymbol{\theta}) = \sum_{j=1}^H w_{ij} \rho \left( \sum_{k=1}^L u_{jk} x_k + \zeta_j \right) + \eta_i, \quad (1 \leq i \leq M). \quad (2)$$

We define a minimal network, following Sussmann (1992), as:

To appear in *Neural Networks*

## A Regularity Condition of the Information Matrix of a Multilayer Perceptron Network

Kenji Fukumizu

Information and Communication R&D Center, Ricoh Co., Ltd.  
3-2-3 Shin-yokohama, Kohoku-ku, Yokohama 222 Japan  
E-mail: fuku@ic.rdc.ricoh.co.jp

**Abstract** – The Fisher information matrix of a multi-layer perceptron network can be singular at certain parameters, and in such cases many statistical techniques based on asymptotic theory cannot be applied properly. In this paper, we prove rigorously that the Fisher information matrix of a three-layer perceptron network is positive definite if and only if the network is irreducible; that is, if there is no hidden unit that makes no contribution to the output and there is no pair of hidden units that could be collapsed to a single unit without altering the input-output map. This implies that a network that has a singular Fisher information matrix can be reduced to a network with a positive definite Fisher information matrix by eliminating redundant hidden units.

**Keywords** – Multilayer perceptron, Parametric estimation, Information matrix, Irreducibility, Minimality, Sigmoidal function.

### 1 INTRODUCTION

Feed-forward neural networks, described as a family of parametrized functions  $\{\mathbf{f}(\mathbf{x}; \boldsymbol{\theta})\}$ , have been applied in a variety of areas and have recently been extensively analyzed theoretically. As Watanabe and Fukumizu (1995) show, it is useful to describe neural networks as a parametric family of probability density functions. From the statistical viewpoint, the least squares learning in a network can be considered as nonlinear regression which approximates  $E[\mathbf{y}|\mathbf{x}]$ , the conditional expectation of the output  $\mathbf{y}$  given an input  $\mathbf{x}$  (White, 1989; Watanabe & Fukumizu, 1995). The least squares estimator of  $\boldsymbol{\theta}$  is equal to the maximum likelihood estimator, which has been extensively analyzed in statistics. White (1989) reviews learning in neural networks in detail from the statistical viewpoint and provides useful insights into network learning methods. The statistical approach to analyzing neural networks is thus clearly an effective one.

Of the many feed-forward models, the multilayer perceptron model (Rumelhart et al., 1986) is one that has shown considerable utility in many applications. From the theoretical viewpoint, it has been proved (Cybenko, 1989; Funahashi, 1989; Hornik et al., 1989) that a three-layer perceptron network can approximate an arbitrary continuous function on a compact set with any desired accuracy if an infinite number of hidden units are available. It has also been proved that three-layer perceptron networks are better approximators than many other conventional models, such as polynomials, in that we need fewer parameters to approximate an arbitrary function in a certain class (Barron, 1993). These theoretical results show the feasibility of multilayer perceptron networks.

In this paper, we discuss the Fisher information matrix of a multilayer perceptron network. In parametric estimation like the learning in feed-forward networks, the Fisher information