行政院國家科學委員會專題研究計畫 成果報告

相關性隱藏節點與學習演算法
研究成果報告(精簡版)

計 畫 類 別 ：個別型
計 畫 編 號 ：NSC 95-2416-H-004-049-
執 行 期 間 ：95 年 08 月 01 日至 96 年 07 月 31 日
執 行 單 位 ：國立政治大學資訊管理學系

計 畫 主 持 人：蔡瑞煌

計畫參與人員：學士級-專任助理：楊佳鳳

報 告 附 件 ：出席國際會議研究心得報告及發表論文

處 理 方 式 ：本計畫可公開查詢

中 華 民 國 96 年 07 月 30 日

1. The Construction Procedure

Huang and Babri (1998) propose an elegant construction method to set up a real-valued single-hidden layer feed-forward neural network (SLFN) with N hidden nodes that successfully learns N distinct samples with zero error. In a *correlated* real-valued single-hidden layer feed-forward neural network (SLFN), the weight vectors in the input layer of all its hidden nodes are linearly dependent. Tsaih and Wan (2007) realize that the SLFN constructed by Huang and Babri (1998) is correlated. They further show that the correlated SLFN has the property of hyperplane preimages. The correlated SLFN provides a hyperplane-preimage approach for the nonlinear regression problem with the assumption of linear preimage. Such usages motivate a study of the construction procedure for creating a correlated SLFN with less than N hidden nodes that perfectly fits N distinct samples.

The proposed construction procedure will initially set up one hidden node and then recruit (add) more (linearly dependent) hidden nodes during the learning process. In the literature, there are some similar procedures; for instance, the tiling algorithm for binary-valued layered feed-forward neural networks (cf. Me'zard and Nadal, 1989), the cascade-correlation algorithm (cf. Fahlman and C. Lebiere, 1990), and the upstart algorithm for binary-valued layered feed-forward neural networks (cf. Frean, 1990). In contrast to these researches, this study copes with the correlated real-valued SLFN.

In the context of *estimation*, the response y equates $f(\mathbf{x}, \mathbf{w}) + \delta$ where $\mathbf{w}$ is the parameter vector and δ is the error term. Usually, the function form of f is predetermined and fixed during the process of deriving its associated $\mathbf{w}$ from a given data set of observations $\{({}_1\mathbf{x}, {}_1y), \ldots, ({}_N\mathbf{x}, {}_Ny)\}$, with ${}_cy$ being the observed response corresponding to the c^{th} observation ${}_c\mathbf{x}$.

The least squares estimator (LSE) is one of the most popular methods for estimating. If $\hat{\mathbf{w}}$ denotes any estimate of $\mathbf{w}$, then LSE is defined to minimize $\sum_{c=1}^{N} {}_ce^2$, where

$$ {}_ce = {}_cy - f({}_c\mathbf{x}, \hat{\mathbf{w}}). \qquad (1)$$

The generalized delta rule proposed in (Rumelhart, Hilton, and Williams, 1986) for the learning process of SLFN is a kind of (nonlinear) LSE. The LSE, however, is known to be very sensitive to outliers.

In the literature of linear regression analysis, there are two approaches of dealing with outlier problems: deletion diagnostics and robust estimators (cf. Rousseeuw and Leroy (1987, page 8)). The diagnostic approach assesses the influence of an individual observation or a subset of observations to the LSE. The diagnostic approach is useful to assess the adequacy of the underlying assumption and to identify unexpected characteristics of the data. One way for the diagnostics is to identify the observations that leave the largest change in the diagnostic quantity (cf. (Cook and Weisberg, 1982)(Atkinson, 1985)) when they are excluded from the fitted data set.

As for the robust statistics approach, the robustness analysis (cf. (Hampel, 1986)) limits the attention to a "trimmed" sum of squared residuals instead of adding all the squared residuals as in the LSE. If only the first q of those ordered squared residuals are included in the summation, then the least trimmed squares (LTS) estimator is defined as

$$ \textit{Minimize} \ \sum_{c=1}^{q} {}_{[c]}e^2, \qquad (2)$$

where $_{[c]}e^2$ denotes the ordered squared residuals; that is, $_{[1]}e^2 \le {}_{[2]}e^2 \le \ldots \le {}_{[N]}e^2$. Zaman, Rousseeuw, and Orthan (2001) suggest that $\lfloor 0.75N \rfloor$[1] is a reasonable value for q in most empirical studies.

Atkinson and Cheng (1999) adapt the forward search algorithm proposed in (Atkinson, 1994) to develop the LTS estimates. The forward search algorithm consists of randomly adopting an (initial) subset of m+1 observations to fit the linear regression model, ordering the residuals of all N observations, and then augmenting the subset gradually by including extra observations based upon the smallest squared residuals principle.

The C-step[2] of Rousseeuw and Van Driessen (2002) can release quite fast a series of subsets of observations whose corresponding total squared residuals are refined gradually. The last subset results in a good linear fitting function which is an approximation of the LTS estimator.

2. The Mapping Requirement and Notations

An SLFN provides a nonlinear mapping between $\mathbf{x}$ and y, whose form is $y = f(\mathbf{x})$. f is a nonlinear function whose parameters (i.e., weights and biases) are derived from a given data set of mapping samples $\{(\mathbf{x}^1, t^1), \ldots, (\mathbf{x}^N, t^N)\}$ with $\mathbf{x}^{c_1} \ne \mathbf{x}^{c_2}$, $c_1 \ne c_2$, and with t^c the target value of y corresponding to $\mathbf{x}^c$. $\mathbf{x} \equiv (x_1, x_2, \ldots, x_m)^{\mathrm{T}} \in R^m$ where x_j is the j^{th} input component, with j from 1 to m.

Hereafter, m and p denote the numbers of adopted input and hidden nodes, respectively; w_{j0}^2 stands for the bias of the j^{th} hidden node; $\mathbf{w}_j^2 \equiv (w_{j1}^2, w_{j2}^2, \ldots, w_{jm}^2)^{\mathrm{T}}$ for the weights between the j^{th} hidden node and input layer; w_0^3 for the bias of the output node; and $\mathbf{w}^3 \equiv (w_1^3, w_2^3, \ldots, w_p^3)^{\mathrm{T}}$ for the weights between the output node and all hidden nodes. Characters in bold represent column vectors; the superscript T indicates transposition.

Let the $tanh(t)$ activation function be used by all hidden nodes and a linear activation function be used by the output node. Thus, given the c^{th} sample $\mathbf{x}^c$, the activation value of the j^{th} hidden node $a^c(w_{j0}^2, \mathbf{w}_j^2)$ and the output value y^c are as follows:

$$a^c(w_{j0}^2, \mathbf{w}_j^2) \equiv tanh\left(w_{j0}^2 + \sum_{i=1}^{m} w_{ji}^2 x_i^c\right); \qquad (3)$$

$$y^c \equiv w_0^3 + \sum_{j=1}^{p} w_j^3 a^c(w_{j0}^2, \mathbf{w}_j^2). \qquad (4)$$

Given the N mapping samples, let $\mathbf{v}(w_{j0}^2, \mathbf{w}_j^2) \equiv (a^1(w_{j0}^2, \mathbf{w}_j^2), a^2(w_{j0}^2, \mathbf{w}_j^2), \ldots, a^N(w_{j0}^2, \mathbf{w}_j^2))^{\mathrm{T}} \in (-1,1)^N$ be the responding vector of the j^{th} hidden node with the c^{th} component being $\mathbf{a}^c(w_{j0}^2, \mathbf{w}_j^2)$. Furthermore, let $\mathbf{1}$ be a $N \times 1$ vector with all components 1 and $\mathbf{T} \equiv (t^1, t^2, \ldots, t^N)^{\mathrm{T}} \in R^{\mathrm{N}}$. Thus, the set of simultaneous equations $w_0^3 + \sum_{j=1}^{p} w_j^3 tanh(w_{j0}^2 + \sum_{i=1}^{m} w_{ji}^2 x_i^c) = t^c \ \forall\ c = 1, \ldots, N$ is equivalent to system (5), which states that $\mathbf{T}$ is in the space spanned by $\mathbf{1}$ and p responding vectors, $\{\mathbf{v}(w_{j0}^2, \mathbf{w}_j^2), j = 1, \ldots, p\}$.

$$w_0^3 \mathbf{1} + \sum_{j=1}^{p} w_j^3 \mathbf{v}(w_{j0}^2, \mathbf{w}_j^2) = \mathbf{T}. \qquad (5)$$

---

[1] Hereafter, $\lfloor x \rfloor$ is the largest integer not larger than x.

[2] C stands for "concentration". The idea of C-step has been implemented in the built-in function *lts.reg* of Splus which is a commercial statistical computing package published by MathSoft Co..

Hereafter, let ${}_{\mathbf{a}}\mathbf{R}_{\mathbf{b}} \equiv \mathbf{a} - \frac{\mathbf{b}^{\mathrm{T}}\mathbf{a}}{\|\mathbf{b}\|}\frac{\mathbf{b}}{\|\mathbf{b}\|}$ denote the residual of vector **a** regarding **b** after the part parallel with the vector **b** has been taken away. $\frac{\mathbf{b}^{\mathrm{T}}\mathbf{a}}{\|\mathbf{b}\|}\frac{\mathbf{b}}{\|\mathbf{b}\|}$ is the projection of **a** in the direction of **b**; and ${}_{\mathbf{a}}\mathbf{R}_{\mathbf{b}}$ is orthogonal to **b** since $\mathbf{b}^{\mathrm{T}}\,{}_{\mathbf{a}}\mathbf{R}_{\mathbf{b}} = \mathbf{b}^{\mathrm{T}}(\mathbf{a} - \frac{\mathbf{b}^{\mathrm{T}}\mathbf{a}}{\|\mathbf{b}\|}\frac{\mathbf{b}}{\|\mathbf{b}\|}) = 0$.

Similarly, let ${}_{\mathbf{a}}\mathbf{R}_{\{\mathbf{b}^1,\ldots,\mathbf{b}^k\}} \equiv {}_{\mathbf{a}}\mathbf{R}_{\{\mathbf{b}^1,\ldots,\mathbf{b}^{k-1}\}} - \frac{({}_{\mathbf{a}}\mathbf{R}_{\{\mathbf{b}^1,\ldots,\mathbf{b}^{k-1}\}})^{\mathrm{T}}\,{}_{\mathbf{b}^k}\mathbf{R}_{\{\mathbf{b}^1,\ldots,\mathbf{b}^{k-1}\}}}{\|{}_{\mathbf{b}^k}\mathbf{R}_{\{\mathbf{b}^1,\ldots,\mathbf{b}^{k-1}\}}\|}\frac{{}_{\mathbf{b}^k}\mathbf{R}_{\{\mathbf{b}^1,\ldots,\mathbf{b}^{k-1}\}}}{\|{}_{\mathbf{b}^k}\mathbf{R}_{\{\mathbf{b}^1,\ldots,\mathbf{b}^{k-1}\}}\|}$ denote the residual of vector **a** regarding the ordered sequence of linearly independent vectors $\{\mathbf{b}^1,\ldots,\mathbf{b}^k\}$ after the part in the (sub-)space spanned by $\{\mathbf{b}^1,\ldots,\mathbf{b}^k\}$ has been taken away. $\frac{({}_{\mathbf{a}}\mathbf{R}_{\{\mathbf{b}^1,\ldots,\mathbf{b}^{k-1}\}})^{\mathrm{T}}\,{}_{\mathbf{b}^k}\mathbf{R}_{\{\mathbf{b}^1,\ldots,\mathbf{b}^{k-1}\}}}{\|{}_{\mathbf{b}^k}\mathbf{R}_{\{\mathbf{b}^1,\ldots,\mathbf{b}^{k-1}\}}\|}\frac{{}_{\mathbf{b}^k}\mathbf{R}_{\{\mathbf{b}^1,\ldots,\mathbf{b}^{k-1}\}}}{\|{}_{\mathbf{b}^k}\mathbf{R}_{\{\mathbf{b}^1,\ldots,\mathbf{b}^{k-1}\}}\|}$ is the projection of ${}_{\mathbf{a}}\mathbf{R}_{\{\mathbf{b}^1,\ldots,\mathbf{b}^{k-1}\}}$ in the direction of ${}_{\mathbf{b}^k}\mathbf{R}_{\{\mathbf{b}^1,\ldots,\mathbf{b}^{k-1}\}}$; and ${}_{\mathbf{a}}\mathbf{R}_{\{\mathbf{b}^1,\ldots,\mathbf{b}^k\}}$ is orthogonal to ${}_{\mathbf{b}^k}\mathbf{R}_{\{\mathbf{b}^1,\ldots,\mathbf{b}^{k-1}\}}$ since $({}_{\mathbf{a}}\mathbf{R}_{\{\mathbf{b}^1,\ldots,\mathbf{b}^k\}})^{\mathrm{T}}\,{}_{\mathbf{b}^k}\mathbf{R}_{\{\mathbf{b}^1,\ldots,\mathbf{b}^{k-1}\}} = ({}_{\mathbf{a}}\mathbf{R}_{\{\mathbf{b}^1,\ldots,\mathbf{b}^{k-1}\}} - \frac{({}_{\mathbf{a}}\mathbf{R}_{\{\mathbf{b}^1,\ldots,\mathbf{b}^{k-1}\}})^{\mathrm{T}}\,{}_{\mathbf{b}^k}\mathbf{R}_{\{\mathbf{b}^1,\ldots,\mathbf{b}^{k-1}\}}}{\|{}_{\mathbf{b}^k}\mathbf{R}_{\{\mathbf{b}^1,\ldots,\mathbf{b}^{k-1}\}}\|}\frac{{}_{\mathbf{b}^k}\mathbf{R}_{\{\mathbf{b}^1,\ldots,\mathbf{b}^{k-1}\}}}{\|{}_{\mathbf{b}^k}\mathbf{R}_{\{\mathbf{b}^1,\ldots,\mathbf{b}^{k-1}\}}\|})^{\mathrm{T}}\,{}_{\mathbf{b}^k}\mathbf{R}_{\{\mathbf{b}^1,\ldots,\mathbf{b}^{k-1}\}} = 0$. By definition we have

Lemma 1. Furthermore, $\mathbf{a} = \frac{\mathbf{b}^{1\mathrm{T}}\mathbf{a}}{\|\mathbf{b}^1\|}\frac{\mathbf{b}^1}{\|\mathbf{b}^1\|} + \sum_{j=2}^{k}\frac{({}_{\mathbf{a}}\mathbf{R}_{\{\mathbf{b}^1,\ldots,\mathbf{b}^{j-1}\}})^{\mathrm{T}}\,{}_{\mathbf{b}^j}\mathbf{R}_{\{\mathbf{b}^1,\ldots,\mathbf{b}^{j-1}\}}}{\|{}_{\mathbf{b}^j}\mathbf{R}_{\{\mathbf{b}^1,\ldots,\mathbf{b}^{j-1}\}}\|}\frac{{}_{\mathbf{b}^j}\mathbf{R}_{\{\mathbf{b}^1,\ldots,\mathbf{b}^{j-1}\}}}{\|{}_{\mathbf{b}^j}\mathbf{R}_{\{\mathbf{b}^1,\ldots,\mathbf{b}^{j-1}\}}\|} + {}_{\mathbf{a}}\mathbf{R}_{\{\mathbf{b}^1,\ldots,\mathbf{b}^k\}}$ and **Lemma 2** lists some properties associated with the above proposed residual vector.

Lemma 1: If **a** is linearly dependent with **b**, then ${}_{\mathbf{a}}\mathbf{R}_{\mathbf{b}} = \mathbf{0}$. Similarly, if **a** can be linearly represented by the set of vectors $\{\mathbf{b}^1,\ldots,\mathbf{b}^k\}$,[3] then ${}_{\mathbf{a}}\mathbf{R}_{\{\mathbf{b}^1,\ldots,\mathbf{b}^k\}} = \mathbf{0}$.

Lemma 2: (i) ${}_{\mathbf{a}}\mathbf{R}_{\mathbf{b}}$ is orthogonal to **b**. (ii) ${}_{\mathbf{a}}\mathbf{R}_{\{\mathbf{b}^1,\ldots,\mathbf{b}^k\}}$ is orthogonal to the subspace spanned by the set of linearly independent vectors $\{\mathbf{b}^1,\ldots,\mathbf{b}^k\}$. (iii) If ${}_{\mathbf{a}}\mathbf{R}_{\{\mathbf{b}^1,\ldots,\mathbf{b}^k\}} = \mathbf{0}$, then ${}_{\mathbf{a}}\mathbf{R}_{\{\mathbf{b}^1,\ldots,\mathbf{b}^{k+1}\}} = \mathbf{0}$.

3. The Proposed Construction Procedure

Table 1 presents the proposed procedure for constructing a correlated SLFN appropriate for fitting the mapping embedded in $\{(\mathbf{x}^1, t^1), \ldots, (\mathbf{x}^N, t^N)\}$.

Table 1. The proposed deterministic procedure for constructing an appropriate correlated SLFN for the mapping requirement of $\{(\mathbf{x}^1, t^1), \ldots, (\mathbf{x}^N, t^N)\}$. $\mathbf{T} \equiv (t^1, t^2, \ldots, t^N)^{\mathrm{T}}$ and $\mathbf{v}(w_0,\mathbf{w}) \equiv (a^1(w_0,\mathbf{w}), a^2(w_0,\mathbf{w}), \ldots, a^N(w_0,\mathbf{w}))^{\mathrm{T}}$ is a $N \times 1$ vector with the c^{th} component being $a^c(w_0,\mathbf{w}) \equiv tanh(\mathbf{w}^{\mathrm{T}}\mathbf{x}^c + w_0)$, in which $\mathbf{w} \equiv (w_1, w_2, \ldots, w_m)^{\mathrm{T}}$.

Step 1: Calculate ${}_{\mathbf{T}}\mathbf{R}_{\mathbf{1}}$. If ${}_{\mathbf{T}}\mathbf{R}_{\mathbf{1}} = \mathbf{0}$, then (i) claim that the fitting job requests no hidden node; (ii) set the bias of output node as $\frac{\mathbf{T}^{\mathrm{T}}\mathbf{1}}{\mathbf{1}^{\mathrm{T}}\mathbf{1}}$ and the weight vector between the output

[3] Namely, **a** is in the (sub-)space spanned by $\{\mathbf{b}^1,\ldots,\mathbf{b}^k\}$.

node and the input layer as **0**; and (iii) stop.

Step 2: Apply the C-step to all N observations to obtain the m+1 input samples that are linearly independent. Let $\mathbf{I}(m+1)$ be the set of indices of these samples and $\mathbf{I}(N)$ be the set of indices of all samples.

Step 3: Calculate $\tilde{t}^{\,c}$ which equates $tanh^{-1}(\frac{t^c - \min_{c\in\mathbf{I}(N)} t^c + 1}{\max_{c\in\mathbf{I}(N)} t^c - \min_{c\in\mathbf{I}(N)} t^c + 2})$ from $\{t^c: \forall\, c \in \mathbf{I}(m+1)\}$.

Next, apply the linear regression method to the data set $\{(\mathbf{x}^c, \tilde{t}^{\,c}): \forall\, c \in \mathbf{I}(m+1)\}$ to get a set of m+1 weights.

Step 4: Set one hidden node in the network whose values of w_{10}^2 and $\mathbf{w}_1^2$ are assigned as the values of the weights obtained in Step 3, and initial values of w_0^3 and w_1^3 are assigned as $\min_{c\in\mathbf{I}(N)} t^c - 1$ and $\max_{c\in\mathbf{I}(N)} t^c - \min_{c\in\mathbf{I}(N)} t^c + 2$, respectively. Then set $\gamma_1 = 1$ and $p = 1$.

Step 5: If $w_0^3\mathbf{1} + \sum_{j=1}^{p} w_j^3 \mathbf{v}(w_{j0}^2, \mathbf{w}_j^2) = \mathbf{T}$, then (i) claim that the fitting job requests p hidden nodes with the bias and weights being the above w_0^3, w_{j0}^2, $\mathbf{w}_j^2$, and w_j^3 for all $1 \le j \le p$; and (ii) stop.

Step 6: If $w_0^3\mathbf{1} + \sum_{j=1}^{p} w_j^3 \mathbf{v}(w_{j0}^2, \mathbf{w}_j^2) \neq \mathbf{T}$, then solve $\min_{w_0,\gamma} \| {}_{\mathrm{T}}\mathrm{R}_{\{\mathbf{1},\mathbf{v}(w_{10}^2,\mathbf{w}_1^2),\ldots,\mathbf{v}(w_{p0}^2,\mathbf{w}_p^2),\mathbf{v}(w_0,\gamma\mathbf{w}_1^2)\}} \|^2$ and let $(w_0^*, \gamma^*) \equiv \arg(\min_{w_0,\gamma} \| {}_{\mathrm{T}}\mathrm{R}_{\{\mathbf{1},\mathbf{v}(w_{10}^2,\mathbf{w}_1^2),\ldots,\mathbf{v}(w_{p0}^2,\mathbf{w}_p^2),\mathbf{v}(w_0,\gamma\mathbf{w}_1^2)\}} \|^2)$.

Step 7: set $\gamma_{p+1} = \gamma^*$, $w_{p+1,0}^2 = w_0^*$, $\mathbf{w}_{p+1}^2 = \gamma_{p+1}\mathbf{w}_1^2$, and $p+1 \rightarrow p$.

Step 8: Apply the linear regression method to the data set $\{((a^c(w_{10}^2, \mathbf{w}_1^2), a^c(w_{20}^2, \mathbf{w}_2^2), \ldots, a^c(w_{p0}^2, \mathbf{w}_p^2))^{\mathrm{T}}, t^c): \forall\, c \in \mathbf{I}(N)\}$ to get values of w_0^3 and w_j^3, $\forall\, j = 1, \ldots, p$. Then go to Step 5.

Step 2 releases m+1 linearly independent input samples through applying the C-step to all N input samples.[4] Let $\mathbf{I}(m+1)$ be the set of indices of these (linearly independent) input samples. Step 3 calculates $\tilde{t}^{\,c}$ via $tanh^{-1}(\frac{t^c - \min_{c\in\mathbf{I}(N)} t^c + 1}{\max_{c\in\mathbf{I}(N)} t^c - \min_{c\in\mathbf{I}(N)} t^c + 2})$ $\forall\, c \in \mathbf{I}(m+1)$. Then Step 3 applies the linear regression method to the data set $\{(\mathbf{x}^c, \tilde{t}^{\,c}): \forall\, c \in \mathbf{I}(m+1)\}$ to get the unique solution of $(w_{10}^2, \mathbf{w}_1^2)$ of system (6), which is a system of m+1 linear equations in m+1 unknowns.

[4] The choices of the subset at Step 2 can be adapted by other considerations (cf. Stromberg (1993)).

$$w_{10}^2 + \sum_{j=1}^{m} w_{1j}^2 \, x_j^c = \tilde{t}^{\,c} \quad \forall\, c \in \mathbf{I}(m+1). \qquad (6)$$

Step 4 sets up the network with one hidden node whose values of $(w_{10}^2, \mathbf{w}_1^2)$ are assigned as obtained in Step 3. The initial values of w_0^3 and w_1^3 are assigned as $\min_{c\in\mathbf{I}(N)} t^c - 1$ and $\max_{c\in\mathbf{I}(N)} t^c - \min_{c\in\mathbf{I}(N)} t^c + 2$, respectively. According to Rousseeuw and Van Driessen (2002), this setup network renders ${}_ce^2 = 0 \;\forall\; c \in \mathbf{I}(m+1)$ and is a good approximation of the LTS estimator.

Step 5 denotes the stopping criterion of the proposed procedure.

The minimization in Step 6 and the assignment in Step 7 determine the bias and weights for the connections of the input nodes to the most newly recruited hidden node. All biases and weights for the connections of the input nodes to the previously recruited hidden nodes are unchanged. Furthermore, the assignment of Step 7 renders the constructed SLFN correlated since $\mathbf{w}_j^2 = \gamma_j \mathbf{w}_1^2$, $j = 1, \ldots, p$.

4. The Correctness of the Proposed Procedure

We now prove that the correlated SLFN constructed by the procedure stated in Table 1 meets the mapping requirement of $\{(\mathbf{x}^1, t^1), \ldots, (\mathbf{x}^N, t^N)\}$ without error.

Tsaih and Wan (2007) state that, for any given $\{(\mathbf{x}^c, t^c): \forall\, c = 1, \ldots, N\}$, there exists a set of $\{(w_{j0}^2, \gamma_j \mathbf{w}_1^2),\ j = 1, \ldots, N\text{-}1\}$ such that the associated square matrix $(\mathbf{1},\ \mathbf{v}(w_{10}^2, \mathbf{w}_1^2),\ \ldots,\ \mathbf{v}(w_{N-1,0}^2, \gamma_{N-1}\mathbf{w}_1^2))$ is invertible and the mapping requirement is achieved perfectly. Therefore, we have **Lemma 3** and the proposed procedure will stop at any p with $0 \le p \le N\text{-}1$.

Lemma 3: If $p < N\text{-}1$, then there always exist w_0 and γ such that ${}_{\mathbf{v}(w_0,\gamma\mathbf{w}_1^2)}\mathrm{R}_{\{\mathbf{1},\mathbf{v}(w_{10}^2,\mathbf{w}_1^2),\ldots,\mathbf{v}(w_{p0}^2,\mathbf{w}_p^2)\}} \neq \mathbf{0}$.

Proof of Lemma 3: Suppose there is a $p < N\text{-}1$ such that there are no w_0 and γ to render ${}_{\mathbf{v}(w_0,\gamma\mathbf{w}_1^2)}\mathrm{R}_{\{\mathbf{1},\mathbf{v}(w_{10}^2,\mathbf{w}_1^2),\ldots,\mathbf{v}(w_{p0}^2,\mathbf{w}_p^2)\}} \neq \mathbf{0}$. In other words, $p < N\text{-}1$ and there are no w_0 and γ such that $\mathbf{v}(w_0, \gamma\mathbf{w}_1^2)$ is linearly independent with $\{\mathbf{1},\ \mathbf{v}(w_{10}^2, \mathbf{w}_1^2),\ \mathbf{v}(w_{20}^2, \mathbf{w}_2^2),\ \ldots,\ \mathbf{v}(w_{p0}^2, \mathbf{w}_p^2)\}$. This contradicts with the statement of Tsaih and Wan (2007). **Q.E.D.**

When the procedure stops at Step 1, the SLFN constructed at Step 1 meets the mapping requirement of $\{(\mathbf{x}^1, t^1), \ldots, (\mathbf{x}^N, t^N)\}$ without error since ${}_{\mathbf{T}}\mathrm{R}_{\mathbf{1}} = \mathbf{0}$. On the other hand, it is obvious to have **Lemma 4**, which states the necessary condition of a SLFN with p hidden nodes appropriate for the mapping requirement of $\{(\mathbf{x}^1, t^1), \ldots, (\mathbf{x}^N, t^N)\}$. Thus the stopping criterion stated in Step 5 is suitable.

Lemma 4: Regarding the mapping samples of $\{(\mathbf{x}^1, t^1), \ldots, (\mathbf{x}^N, t^N)\}$, the SLFN with p hidden nodes is appropriate for the mapping requirement if $w_0^3 \mathbf{1} + \sum_{j=1}^{p} w_j^3 \mathbf{v}(w_{j0}^2, \mathbf{w}_j^2) = \mathbf{T}$.

Consider the case of $w_0^3 \mathbf{1} + \sum_{j=1}^{p-1} w_j^3 \mathbf{v}(w_{j0}^2, \mathbf{w}_j^2) \neq \mathbf{T}$ and $w_0^3 \mathbf{1} + \sum_{j=1}^{p} w_j^3 \mathbf{v}(w_{j0}^2, \mathbf{w}_j^2) = \mathbf{T}$. Namely, ${}_{\mathbf{T}}\mathrm{R}_{\{\mathbf{1},\mathbf{v}(w_{10}^2,\mathbf{w}_1^2),\ldots,\mathbf{v}(w_{j0}^2,\mathbf{w}_j^2)\}} \neq \mathbf{0}$ for each $1 \le j \le p\text{-}1$ and ${}_{\mathbf{T}}\mathrm{R}_{\{\mathbf{1},\mathbf{v}(w_{10}^2,\mathbf{w}_1^2),\ldots,\mathbf{v}(w_{p0}^2,\mathbf{w}_p^2)\}} = \mathbf{0}$. For each $1 \le j \le p\text{-}1$, from calculation, we have ${}_{\mathbf{T}}\mathrm{R}_{\{\mathbf{1},\mathbf{v}(w_{10}^2,\mathbf{w}_1^2),\ldots,\mathbf{v}(w_{j0}^2,\mathbf{w}_j^2),\mathbf{v}(w_0,\gamma\mathbf{w}_1^2)\}} = {}_{\mathbf{T}}\mathrm{R}_{\{\mathbf{1},\mathbf{v}(w_{10}^2,\mathbf{w}_1^2),\ldots,\mathbf{v}(w_{j0}^2,\mathbf{w}_j^2)\}} - \frac{({}_{\mathbf{T}}\mathrm{R}_{\{\mathbf{1},\mathbf{v}(w_{10}^2,\mathbf{w}_1^2),\ldots,\mathbf{v}(w_{j0}^2,\mathbf{w}_j^2)\}})^{\mathrm{T}}\ {}_{\mathbf{v}(w_0,\gamma\mathbf{w}_1^2)}\mathrm{R}_{\{\mathbf{1},\mathbf{v}(w_{10}^2,\mathbf{w}_1^2),\ldots,\mathbf{v}(w_{j0}^2,\mathbf{w}_j^2)\}}}{\| {}_{\mathbf{v}(w_0,\gamma\mathbf{w}_1^2)}\mathrm{R}_{\{\mathbf{1},\mathbf{v}(w_{10}^2,\mathbf{w}_1^2),\ldots,\mathbf{v}(w_{j0}^2,\mathbf{w}_j^2)\}} \|}$ and $\frac{{}_{\mathbf{v}(w_0,\gamma\mathbf{w}_1^2)}\mathrm{R}_{\{\mathbf{1},\mathbf{v}(w_{10}^2,\mathbf{w}_1^2),\ldots,\mathbf{v}(w_{j0}^2,\mathbf{w}_j^2)\}}}{\| {}_{\mathbf{v}(w_0,\gamma\mathbf{w}_1^2)}\mathrm{R}_{\{\mathbf{1},\mathbf{v}(w_{10}^2,\mathbf{w}_1^2),\ldots,\mathbf{v}(w_{j0}^2,\mathbf{w}_j^2)\}} \|}$

$\| \ {}_{\mathbf{T}}R_{\{\mathbf{1},\mathbf{v}(w_{10}^2,\mathbf{w}_1^2),\ldots,\mathbf{v}(w_{j0}^2,\mathbf{w}_j^2),\mathbf{v}(w_0,\gamma\mathbf{w}_1^2)\}} \ \|^2 = \| \ {}_{\mathbf{T}}R_{\{\mathbf{1},\mathbf{v}(w_{10}^2,\mathbf{w}_1^2),\ldots,\mathbf{v}(w_{j0}^2,\mathbf{w}_j^2)\}} \ \|^2 \ (1-(\frac{({}_{\mathbf{T}}R_{\{\mathbf{1},\mathbf{v}(w_{10}^2,\mathbf{w}_1^2),\ldots,\mathbf{v}(w_{j0}^2,\mathbf{w}_j^2)\}})^{\mathrm{T}}}{\|{}_{\mathbf{T}}R_{\{\mathbf{1},\mathbf{v}(w_{10}^2,\mathbf{w}_1^2),\ldots,\mathbf{v}(w_{j0}^2,\mathbf{w}_j^2)\}}\|}$

$\frac{\mathbf{v}(w_0,\gamma\mathbf{w}_1^2)R_{\{\mathbf{1},\mathbf{v}(w_{10}^2,\mathbf{w}_1^2),\ldots,\mathbf{v}(w_{j0}^2,\mathbf{w}_j^2)\}}}{\|_{\mathbf{v}(w_0,\gamma\mathbf{w}_1^2)}R_{\{\mathbf{1},\mathbf{v}(w_{10}^2,\mathbf{w}_1^2),\ldots,\mathbf{v}(w_{j0}^2,\mathbf{w}_j^2)\}}\|})^2)$. Hence we have Lemma 5. Furthermore, suppose ${}_{\mathbf{T}}R_{\{\mathbf{1},\mathbf{v}(w_{10}^2,\mathbf{w}_1^2),\ldots,\mathbf{v}(w_{p0}^2,\mathbf{w}_p^2)\}} \neq \mathbf{0}$, $\| \ {}_{\mathbf{T}}R_{\{\mathbf{1},\mathbf{v}(w_{10}^2,\mathbf{w}_1^2),\ldots,\mathbf{v}(w_{p0}^2,\mathbf{w}_p^2),\mathbf{v}(w_0^*,\gamma^*\mathbf{w}_1^2)\}} \|^2 = 0$ if and only if $(\frac{({}_{\mathbf{T}}R_{\{\mathbf{1},\mathbf{v}(w_{10}^2,\mathbf{w}_1^2),\ldots,\mathbf{v}(w_{p0}^2,\mathbf{w}_p^2)\}})^{\mathrm{T}}}{\|{}_{\mathbf{T}}R_{\{\mathbf{1},\mathbf{v}(w_{10}^2,\mathbf{w}_1^2),\ldots,\mathbf{v}(w_{p0}^2,\mathbf{w}_p^2)\}}\|}$

$\frac{\mathbf{v}(w_0^*,\gamma^*\mathbf{w}_1^2)R_{\{\mathbf{1},\mathbf{v}(w_{10}^2,\mathbf{w}_1^2),\ldots,\mathbf{v}(w_{p0}^2,\mathbf{w}_p^2)\}}}{\|_{\mathbf{v}(w_0^*,\gamma^*\mathbf{w}_1^2)}R_{\{\mathbf{1},\mathbf{v}(w_{10}^2,\mathbf{w}_1^2),\ldots,\mathbf{v}(w_{p0}^2,\mathbf{w}_p^2)\}}\|})^2 = 1$, which implies the following **Lemma 6** is true.

Lemma 5: If ${}_{\mathbf{T}}R_{\{\mathbf{1},\mathbf{v}(w_{10}^2,\mathbf{w}_1^2),\ldots,\mathbf{v}(w_{j0}^2,\mathbf{w}_j^2)\}} \neq \mathbf{0}$, then $\min_{w_0,\gamma} \| \ {}_{\mathbf{T}}R_{\{\mathbf{1},\mathbf{v}(w_{10}^2,\mathbf{w}_1^2),\ldots,\mathbf{v}(w_{j0}^2,\mathbf{w}_j^2),\mathbf{v}(w_0,\gamma\mathbf{w}_1^2)\}} \|^2 > \| \ {}_{\mathbf{T}}R_{\{\mathbf{1},\mathbf{v}(w_{10}^2,\mathbf{w}_1^2),\ldots,\mathbf{v}(w_{j0}^2,\mathbf{w}_j^2)\}} \|^2$.

Proof of Lemma 5: From **Lemma 3**, there always exists w_0 and γ such that ${}_{\mathbf{v}(w_0,\gamma\mathbf{w}_1^2)}R_{\{\mathbf{1},\mathbf{v}(w_{10}^2,\mathbf{w}_1^2),\ldots,\mathbf{v}(w_{j0}^2,\mathbf{w}_j^2)\}}$ is a non-zero vector. Now ${}_{\mathbf{T}}R_{\{\mathbf{1},\mathbf{v}(w_{10}^2,\mathbf{w}_1^2),\ldots,\mathbf{v}(w_{j0}^2,\mathbf{w}_j^2),\mathbf{v}(w_0,\gamma\mathbf{w}_1^2)\}} = {}_{\mathbf{T}}R_{\{\mathbf{1},\mathbf{v}(w_{10}^2,\mathbf{w}_1^2),\ldots,\mathbf{v}(w_{j0}^2,\mathbf{w}_j^2)\}} - \frac{({}_{\mathbf{T}}R_{\{\mathbf{1},\mathbf{v}(w_{10}^2,\mathbf{w}_1^2),\ldots,\mathbf{v}(w_{j0}^2,\mathbf{w}_j^2)\}})^{\mathrm{T}}{}_{\mathbf{v}(w_0,\gamma\mathbf{w}_1^2)}R_{\{\mathbf{1},\mathbf{v}(w_{10}^2,\mathbf{w}_1^2),\ldots,\mathbf{v}(w_{j0}^2,\mathbf{w}_j^2)\}}}{\|_{\mathbf{v}(w_0,\gamma\mathbf{w}_1^2)}R_{\{\mathbf{1},\mathbf{v}(w_{10}^2,\mathbf{w}_1^2),\ldots,\mathbf{v}(w_{j0}^2,\mathbf{w}_j^2)\}}\|} \frac{\mathbf{v}(w_0,\gamma\mathbf{w}_1^2)R_{\{\mathbf{1},\mathbf{v}(w_{10}^2,\mathbf{w}_1^2),\ldots,\mathbf{v}(w_{j0}^2,\mathbf{w}_j^2)\}}}{\|_{\mathbf{v}(w_0,\gamma\mathbf{w}_1^2)}R_{\{\mathbf{1},\mathbf{v}(w_{10}^2,\mathbf{w}_1^2),\ldots,\mathbf{v}(w_{j0}^2,\mathbf{w}_j^2)\}}\|}$ and ${}_{\mathbf{v}(w_0,\gamma\mathbf{w}_1^2)}R_{\{\mathbf{1},\mathbf{v}(w_{10}^2,\mathbf{w}_1^2),\ldots,\mathbf{v}(w_{j0}^2,\mathbf{w}_j^2)\}}$ and ${}_{\mathbf{T}}R_{\{\mathbf{1},\mathbf{v}(w_{10}^2,\mathbf{w}_1^2),\ldots,\mathbf{v}(w_{j0}^2,\mathbf{w}_j^2),\mathbf{v}(w_0,\gamma\mathbf{w}_1^2)\}}$ are orthogonal to each other. Thus, $\| \ {}_{\mathbf{T}}R_{\{\mathbf{1},\mathbf{v}(w_{10}^2,\mathbf{w}_1^2),\ldots,\mathbf{v}(w_{j0}^2,\mathbf{w}_j^2)\}} \|^2 > \| \ {}_{\mathbf{T}}R_{\{\mathbf{1},\mathbf{v}(w_{10}^2,\mathbf{w}_1^2),\ldots,\mathbf{v}(w_{j0}^2,\mathbf{w}_j^2),\mathbf{v}(w_0,\gamma\mathbf{w}_1^2)\}} \|^2$. In other words, the optimization of $\min_{w_0,\gamma} \| \ {}_{\mathbf{T}}R_{\{\mathbf{1},\mathbf{v}(w_{10}^2,\mathbf{w}_1^2),\ldots,\mathbf{v}(w_{j0}^2,\mathbf{w}_j^2),\mathbf{v}(w_0,\gamma\mathbf{w}_1^2)\}} \|^2$ leads to a non-zero ${}_{\mathbf{v}(w_0^*,\gamma^*\mathbf{w}_1^2)}R_{\{\mathbf{1},\mathbf{v}(w_{10}^2,\mathbf{w}_1^2),\ldots,\mathbf{v}(w_{j0}^2,\mathbf{w}_j^2)\}}$, in which $(w_0^*,\gamma^*) \equiv \arg(\min_{w_0,\gamma} \| \ {}_{\mathbf{T}}R_{\{\mathbf{1},\mathbf{v}(w_{10}^2,\mathbf{w}_1^2),\ldots,\mathbf{v}(w_{j0}^2,\mathbf{w}_j^2),\mathbf{v}(w_0,\gamma\mathbf{w}_1^2)\}} \|^2)$ and $\| \ {}_{\mathbf{T}}R_{\{\mathbf{1},\mathbf{v}(w_{10}^2,\mathbf{w}_1^2),\ldots,\mathbf{v}(w_{j0}^2,\mathbf{w}_j^2),\mathbf{v}(w_0^*,\gamma^*\mathbf{w}_1^2)\}} \|^2 < \| \ {}_{\mathbf{T}}R_{\{\mathbf{1},\mathbf{v}(w_{10}^2,\mathbf{w}_1^2),\ldots,\mathbf{v}(w_{j0}^2,\mathbf{w}_j^2)\}} \|^2$.

Q.E.D.

Lemma 6: Suppose $\| \ {}_{\mathbf{T}}R_{\{\mathbf{1},\mathbf{v}(w_{10}^2,\mathbf{w}_1^2),\ldots,\mathbf{v}(w_{p0}^2,\mathbf{w}_p^2)\}} \|^2 \neq 0$. $\| \ {}_{\mathbf{T}}R_{\{\mathbf{1},\mathbf{v}(w_{10}^2,\mathbf{w}_1^2),\ldots,\mathbf{v}(w_{p0}^2,\mathbf{w}_p^2),\mathbf{v}(w_0^*,\gamma^*\mathbf{w}_1^2)\}} \|^2 = 0$ if and only if ${}_{\mathbf{T}}R_{\{\mathbf{1},\mathbf{v}(w_{10}^2,\mathbf{w}_1^2),\ldots,\mathbf{v}(w_{p0}^2,\mathbf{w}_p^2)\}}$ and ${}_{\mathbf{v}(w_0^*,\gamma^*\mathbf{w}_1^2)}R_{\{\mathbf{1},\mathbf{v}(w_{10}^2,\mathbf{w}_1^2),\ldots,\mathbf{v}(w_{p0}^2,\mathbf{w}_p^2)\}}$ are parallel.

Lemma 7: The (sub-)space spanned by the set of linearly independent vectors $\{\mathbf{b}^1,\ldots,\mathbf{b}^k\}$ is equivalent with the one spanned by the set of orthonormal vectors $\{\frac{\mathbf{b}^1}{\|\mathbf{b}^1\|}, \frac{\mathbf{b}^2 R_{\mathbf{b}^1}}{\|_{\mathbf{b}^2}R_{\mathbf{b}^1}\|}, \ldots, \frac{\mathbf{b}^k R_{\{\mathbf{b}^1,\ldots,\mathbf{b}^{k-1}\}}}{\|_{\mathbf{b}^k}R_{\{\mathbf{b}^1,\ldots,\mathbf{b}^{k-1}\}}\|}\}$.

Proof of Lemma 7: Let us prove by induction. It is trivial for the case of any set of two linearly independent vectors $\{\mathbf{b}^1, \mathbf{b}^2\}$ since $\mathbf{b}^2R_{\mathbf{b}^1}$ is orthogonal to $\mathbf{b}^1$ and $\mathbf{b}^1R_{\mathbf{b}^2}$ is orthogonal to $\mathbf{b}^2$.

Now consider the case of any set of k+1 linearly independent vectors $\{\mathbf{b}^1,\ldots,\mathbf{b}^{k+1}\}$ with $k \geq 2$. Assume the subspace spanned by the subset of vectors $\{\mathbf{b}^1,\ldots,\mathbf{b}^k\}$ is equivalent with the one spanned by the set of orthonormal vectors $\{\frac{\mathbf{b}^1}{\|\mathbf{b}^1\|}, \frac{\mathbf{b}^2 R_{\mathbf{b}^1}}{\|_{\mathbf{b}^2}R_{\mathbf{b}^1}\|}, \ldots, \frac{\mathbf{b}^k R_{\{\mathbf{b}^1,\ldots,\mathbf{b}^{k-1}\}}}{\|_{\mathbf{b}^k}R_{\{\mathbf{b}^1,\ldots,\mathbf{b}^{k-1}\}}\|}\}$. The

vector $\mathbf{b}^{k+1}$ can be represented as $\frac{(\mathbf{b}^{k+1})^{\mathrm{T}}\mathbf{b}^1}{\|\mathbf{b}^1\|}\frac{\mathbf{b}^1}{\|\mathbf{b}^1\|} + \sum_{j=2}^{k}\frac{(\mathbf{b}^{k+1}\mathrm{R}_{\{\mathbf{b}^1,\ldots,\mathbf{b}^{j-1}\}})^{\mathrm{T}}\mathbf{b}^j\mathrm{R}_{\{\mathbf{b}^1,\ldots,\mathbf{b}^{j-1}\}}}{\|\mathbf{b}^j\mathrm{R}_{\{\mathbf{b}^1,\ldots,\mathbf{b}^{j-1}\}}\|}\frac{\mathbf{b}^j\mathrm{R}_{\{\mathbf{b}^1,\ldots,\mathbf{b}^{j-1}\}}}{\|\mathbf{b}^j\mathrm{R}_{\{\mathbf{b}^1,\ldots,\mathbf{b}^{j-1}\}}\|} + \mathbf{b}^{k+1}\mathrm{R}_{\{\mathbf{b}^1,\ldots,\mathbf{b}^k\}}$, in which $\mathbf{b}^{k+1}\mathrm{R}_{\{\mathbf{b}^1,\ldots,\mathbf{b}^k\}}$ is a non-zero vector since $\mathbf{b}^{k+1}$ is linearly independent with the set of vectors $\{\mathbf{b}^1,\ldots,\mathbf{b}^k\}$. Thus the (sub-)space spanned by the set of vectors $\{\mathbf{b}^1,\ldots,\mathbf{b}^{k+1}\}$ is equivalent with the one spanned by the set of vectors $\{\frac{\mathbf{b}^1}{\|\mathbf{b}^1\|}, \frac{\mathbf{b}^2\mathrm{R}_{\mathbf{b}^1}}{\|\mathbf{b}^2\mathrm{R}_{\mathbf{b}^1}\|},\ldots, \mathbf{b}^{k+1}\mathrm{R}_{\{\mathbf{b}^1,\ldots,\mathbf{b}^k\}}\}$. Furthermore, $(\mathbf{b}^{k+1}\mathrm{R}_{\{\mathbf{b}^1,\ldots,\mathbf{b}^k\}})^{\mathrm{T}}\,\mathbf{b}^1 = (\mathbf{b}^{k+1}\mathrm{R}_{\{\mathbf{b}^1,\ldots,\mathbf{b}^{k-1}\}} - \frac{(\mathbf{b}^{k+1}\mathrm{R}_{\{\mathbf{b}^1,\ldots,\mathbf{b}^{k-1}\}})^{\mathrm{T}}\mathbf{b}^k\mathrm{R}_{\{\mathbf{b}^1,\ldots,\mathbf{b}^{k-1}\}}}{\|\mathbf{b}^k\mathrm{R}_{\{\mathbf{b}^1,\ldots,\mathbf{b}^{k-1}\}}\|}\frac{\mathbf{b}^k\mathrm{R}_{\{\mathbf{b}^1,\ldots,\mathbf{b}^{k-1}\}}}{\|\mathbf{b}^k\mathrm{R}_{\{\mathbf{b}^1,\ldots,\mathbf{b}^{k-1}\}}\|})^{\mathrm{T}}\,\mathbf{b}^1 = 0$; and, for any $2 \le j \le$ k-1, $(\mathbf{b}^{k+1}\mathrm{R}_{\{\mathbf{b}^1,\ldots,\mathbf{b}^k\}})^{\mathrm{T}}\,\mathbf{b}^j\mathrm{R}_{\{\mathbf{b}^1,\ldots,\mathbf{b}^{j-1}\}} = (\mathbf{b}^{k+1}\mathrm{R}_{\{\mathbf{b}^1,\ldots,\mathbf{b}^{k-1}\}} - \frac{(\mathbf{b}^{k+1}\mathrm{R}_{\{\mathbf{b}^1,\ldots,\mathbf{b}^{k-1}\}})^{\mathrm{T}}\mathbf{b}^k\mathrm{R}_{\{\mathbf{b}^1,\ldots,\mathbf{b}^{k-1}\}}}{\|\mathbf{b}^k\mathrm{R}_{\{\mathbf{b}^1,\ldots,\mathbf{b}^{k-1}\}}\|}\frac{\mathbf{b}^k\mathrm{R}_{\{\mathbf{b}^1,\ldots,\mathbf{b}^{k-1}\}}}{\|\mathbf{b}^k\mathrm{R}_{\{\mathbf{b}^1,\ldots,\mathbf{b}^{k-1}\}}\|})^{\mathrm{T}}\,\mathbf{b}^j\mathrm{R}_{\{\mathbf{b}^1,\ldots,\mathbf{b}^{j-1}\}} = 0$; and $(\mathbf{b}^{k+1}\mathrm{R}_{\{\mathbf{b}^1,\ldots,\mathbf{b}^k\}})^{\mathrm{T}}\,\mathbf{b}^k\mathrm{R}_{\{\mathbf{b}^1,\ldots,\mathbf{b}^{k-1}\}} = (\mathbf{b}^{k+1}\mathrm{R}_{\{\mathbf{b}^1,\ldots,\mathbf{b}^{k-1}\}} - \frac{(\mathbf{b}^{k+1}\mathrm{R}_{\{\mathbf{b}^1,\ldots,\mathbf{b}^{k-1}\}})^{\mathrm{T}}\mathbf{b}^k\mathrm{R}_{\{\mathbf{b}^1,\ldots,\mathbf{b}^{k-1}\}}}{\|\mathbf{b}^k\mathrm{R}_{\{\mathbf{b}^1,\ldots,\mathbf{b}^{k-1}\}}\|}\frac{\mathbf{b}^k\mathrm{R}_{\{\mathbf{b}^1,\ldots,\mathbf{b}^{k-1}\}}}{\|\mathbf{b}^k\mathrm{R}_{\{\mathbf{b}^1,\ldots,\mathbf{b}^{k-1}\}}\|})^{\mathrm{T}}\,\mathbf{b}^k\mathrm{R}_{\{\mathbf{b}^1,\ldots,\mathbf{b}^{k-1}\}} = 0$. Q.E.D.

Let $\mathbf{u}^0 \equiv \frac{\mathbf{1}}{\|\mathbf{1}\|}$, $\mathbf{u}^1 \equiv \frac{\mathbf{v}(w_{10}^2,\mathbf{w}_1^2)\mathrm{R}_{\mathbf{1}}}{\|\mathbf{v}(w_{10}^2,\mathbf{w}_1^2)\mathrm{R}_{\mathbf{1}}\|}$, and $\mathbf{u}^j \equiv \frac{\mathbf{v}(w_{j0}^2,\mathbf{w}_j^2)\mathrm{R}_{\{\mathbf{1},\mathbf{v}(w_{10}^2,\mathbf{w}_1^2),\ldots,\mathbf{v}(w_{j-1,0}^2,\mathbf{w}_{j-1}^2)\}}}{\|\mathbf{v}(w_{j0}^2,\mathbf{w}_j^2)\mathrm{R}_{\{\mathbf{1},\mathbf{v}(w_{10}^2,\mathbf{w}_1^2),\ldots,\mathbf{v}(w_{j-1,0}^2,\mathbf{w}_{j-1}^2)\}}\|}$ for all $2 \le j \le p$. Thus $\mathbf{T}^{\mathrm{T}}\,\mathbf{u}^k = (w_0^3\,\mathbf{1} + \sum_{j=1}^{p} w_j^3\,\mathbf{v}(w_{j0}^2,\mathbf{w}_j^2))^{\mathrm{T}}\,\mathbf{u}^k = \sum_{j=k}^{p} w_j^3\,(\mathbf{v}(w_{j0}^2,\mathbf{w}_j^2))^{\mathrm{T}}\,\mathbf{u}^k$ for all $1 \le k \le p$, since, from Lemma 7, $\mathbf{v}(w_{j0}^2,\mathbf{w}_j^2)$ is in the subspace spanned by $\{\mathbf{u}^0,\mathbf{u}^1,\ldots,\mathbf{u}^j\}$ and $(\mathbf{u}^j)^{\mathrm{T}}\,\mathbf{u}^k \ne 0\ \forall\ k \le j \le p$.

Thus $w_p^3 \equiv \frac{\mathbf{T}^{\mathrm{T}}\mathbf{u}^p}{(\mathbf{v}(w_{p0}^2,\mathbf{w}_p^2))^{\mathrm{T}}\mathbf{u}^p}$, $w_j^3 \equiv \frac{(\mathbf{T} - \sum_{l=j+1}^{p} w_l^3\mathbf{v}(w_{l0}^2,\mathbf{w}_l^2))^{\mathrm{T}}\mathbf{u}^j}{(\mathbf{v}(w_{j0}^2,\mathbf{w}_j^2))^{\mathrm{T}}\mathbf{u}^j}$ for all p-1 $\ge j \ge$ 1, and $w_0^3 \equiv \frac{(\mathbf{T} - \sum_{l=1}^{p} w_l^3\mathbf{v}(w_{l0}^2,\mathbf{w}_l^2))^{\mathrm{T}}\mathbf{u}^0}{\mathbf{1}^{\mathrm{T}}\mathbf{u}^0}$.

Let $\mathbf{A}^{-1} \equiv \mathbf{I}$ and $\mathbf{A}^k \equiv \mathbf{A}^{k-1} - \mathbf{u}^k(\mathbf{u}^k)^{\mathrm{T}}$ for all $0 \le k \le N$-1. Thus, $\mathbf{A}^k = \mathbf{I} - \sum_{j=0}^{k}\mathbf{u}^j(\mathbf{u}^j)^{\mathrm{T}}$ for all $0 \le k \le N$-1. Since all vectors in the set $\{\mathbf{u}^0,\mathbf{u}^1,\ldots,\mathbf{u}^k\}$ are orthonormal, $(\mathbf{A}^k)^{\mathrm{T}} = \mathbf{A}^k$ and $\mathbf{A}^k\,\mathbf{A}^k = \mathbf{A}^k$. Furthermore, because $_{\mathbf{T}}\mathrm{R}_{\{\mathbf{u}^0,\mathbf{u}^1,\ldots,\mathbf{u}^p\}} = \mathbf{T} - \sum_{j=0}^{p}(\mathbf{u}^j)^{\mathrm{T}}\,\mathbf{T}\,\mathbf{u}^j$, $_{\mathbf{T}}\mathrm{R}_{\{\mathbf{u}^0,\mathbf{u}^1,\ldots,\mathbf{u}^p\}} = \mathbf{A}^p\,\mathbf{T}$. Similarly, $_{\mathbf{v}(w_0,\mathbf{w})}\mathrm{R}_{\{\mathbf{u}^0,\mathbf{u}^1,\ldots,\mathbf{u}^p\}} = \mathbf{A}^p\,\mathbf{v}(w_0,\mathbf{w})$. Thus, $_{\mathbf{v}(w_0,\mathbf{w})}\mathrm{R}_{\{\mathbf{u}^0,\mathbf{u}^1,\ldots,\mathbf{u}^p\}}{}^{\mathrm{T}}\,\mathbf{T} = \mathbf{v}(w_0,\mathbf{w})^{\mathrm{T}}\,\mathbf{A}^p\,\mathbf{T}$, $\mathbf{v}(w_0,\mathbf{w})^{\mathrm{T}}\,_{\mathbf{T}}\mathrm{R}_{\{\mathbf{u}^0,\mathbf{u}^1,\ldots,\mathbf{u}^p\}} = \mathbf{v}(w_0,\mathbf{w})^{\mathrm{T}}\,\mathbf{A}^p\,\mathbf{T}$, $_{\mathbf{v}(w_0,\mathbf{w})}\mathrm{R}_{\{\mathbf{u}^0,\mathbf{u}^1,\ldots,\mathbf{u}^p\}}{}^{\mathrm{T}}\,_{\mathbf{T}}\mathrm{R}_{\{\mathbf{u}^0,\mathbf{u}^1,\ldots,\mathbf{u}^p\}} = \mathbf{v}(w_0,\mathbf{w})^{\mathrm{T}}\,\mathbf{A}^p\,\mathbf{T}$, $_{\mathbf{T}}\mathrm{R}_{\{\mathbf{u}^0,\mathbf{u}^1,\ldots,\mathbf{u}^p\}}{}^{\mathrm{T}}\,_{\mathbf{T}}\mathrm{R}_{\{\mathbf{u}^0,\mathbf{u}^1,\ldots,\mathbf{u}^p\}} = \mathbf{T}^{\mathrm{T}}\,\mathbf{A}^p\,\mathbf{T}$, and $\|_{\mathbf{v}(w_0,\mathbf{w})}\mathrm{R}_{\{\mathbf{u}^0,\mathbf{u}^1,\ldots,\mathbf{u}^p\}}\|^2 = \mathbf{v}(w_0,\mathbf{w})^{\mathrm{T}}\,\mathbf{A}^p\,\mathbf{v}(w_0,\mathbf{w})$. thus, $\|_{\mathbf{T}}\mathrm{R}_{\mathbf{v}(w_0,\mathbf{w})}\|^2 = \mathbf{T}^{\mathrm{T}}\,\mathbf{T} - \frac{(\mathbf{v}(w_0,\mathbf{w})^{\mathrm{T}}\mathbf{T})^2}{\mathbf{v}(w_0,\mathbf{w})^{\mathrm{T}}\mathbf{v}(w_0,\mathbf{w})}$ and $\|_{\mathbf{T}}\mathrm{R}_{\{\mathbf{u}^0,\mathbf{u}^1,\ldots,\mathbf{u}^p,\mathbf{v}(w_0,\mathbf{w})\}}\|^2 = \mathbf{T}^{\mathrm{T}}\,\mathbf{A}^p\,\mathbf{T} - \frac{(\mathbf{v}(w_0,\mathbf{w})^{\mathrm{T}}\mathbf{A}^p\mathbf{T})^2}{\mathbf{v}(w_0,\mathbf{w})^{\mathrm{T}}\mathbf{A}^p\mathbf{v}(w_0,\mathbf{w})} = \frac{\mathbf{T}^{\mathrm{T}}\mathbf{A}^p\mathbf{T}\mathbf{v}(w_0,\mathbf{w})^{\mathrm{T}}\mathbf{A}^p\mathbf{v}(w_0,\mathbf{w}) - (\mathbf{v}(w_0,\mathbf{w})^{\mathrm{T}}\mathbf{A}^p\mathbf{T})^2}{\mathbf{v}(w_0,\mathbf{w})^{\mathrm{T}}\mathbf{A}^p\mathbf{v}(w_0,\mathbf{w})}$.

Lemma 8: If $_{\mathbf{T}}\mathrm{R}_{\{\mathbf{1},\mathbf{v}(w_{10}^2,\mathbf{w}_1^2),\ldots,\mathbf{v}(w_{p0}^2,\mathbf{w}_p^2)\}} \ne \mathbf{0}$ and there exists a vector $\mathbf{v}(\bar{w}_0,\bar{r}\mathbf{w}_1^2)$ such that $_{\mathbf{T}}\mathrm{R}_{\{\mathbf{1},\mathbf{v}(w_{10}^2,\mathbf{w}_1^2),\ldots,\mathbf{v}(w_{p0}^2,\mathbf{w}_p^2),\mathbf{v}(\bar{w}_0,\bar{r}\mathbf{w}_1^2)\}} = \mathbf{0}$, then $\min_{w_0,\gamma} \|_{\mathbf{T}}\mathrm{R}_{\{\mathbf{1},\mathbf{v}(w_{10}^2,\mathbf{w}_1^2),\ldots,\mathbf{v}(w_{p0}^2,\mathbf{w}_p^2),\mathbf{v}(w_0,\gamma\mathbf{w}_1^2)\}}\|^2$ leads to a non-zero

$_{\mathbf{v}(w_0^*,\gamma^*\mathbf{w}_1^2)}\mathrm{R}_{\{\mathbf{1},\mathbf{v}(w_{10}^2,\mathbf{w}_1^2),\ldots,\mathbf{v}(w_{p0}^2,\mathbf{w}_p^2)\}}$ such that $\frac{{}_{\mathbf{v}(w_0^*,\gamma^*\mathbf{w}_1^2)}\mathrm{R}_{\{\mathbf{1},\mathbf{v}(w_{10}^2,\mathbf{w}_1^2),\ldots,\mathbf{v}(w_{p0}^2,\mathbf{w}_p^2)\}}}{\|{}_{\mathbf{v}(w_0^*,\gamma^*\mathbf{w}_1^2)}\mathrm{R}_{\{\mathbf{1},\mathbf{v}(w_{10}^2,\mathbf{w}_1^2),\ldots,\mathbf{v}(w_{p0}^2,\mathbf{w}_p^2)\}}\|}$ and $\frac{{}_{\mathbf{v}(\bar{w}_0,\bar{\mathbf{w}}_1^2)}\mathrm{R}_{\{\mathbf{1},\mathbf{v}(w_{10}^2,\mathbf{w}_1^2),\ldots,\mathbf{v}(w_{p0}^2,\mathbf{w}_p^2)\}}}{\|{}_{\mathbf{v}(\bar{w}_0,\bar{\mathbf{w}}_1^2)}\mathrm{R}_{\{\mathbf{1},\mathbf{v}(w_{10}^2,\mathbf{w}_1^2),\ldots,\mathbf{v}(w_{p0}^2,\mathbf{w}_p^2)\}}\|}$ are parallel and $\|{}_{\mathbf{T}}\mathrm{R}_{\{\mathbf{1},\mathbf{v}(w_{10}^2,\mathbf{w}_1^2),\ldots,\mathbf{v}(w_{p0}^2,\mathbf{w}_p^2),\mathbf{v}(w_0^*,\gamma^*\mathbf{w}_1^2)\}}\|^2 = 0$.

Proof: Let $\bar{\mathbf{u}}^{p+1} \equiv \frac{{}_{\mathbf{v}(\bar{w}_0,\bar{\mathbf{w}})}\mathrm{R}_{\{\mathbf{u}^1,\ldots,\mathbf{u}^p\}}}{\|{}_{\mathbf{v}(\bar{w}_0,\bar{\mathbf{w}})}\mathrm{R}_{\{\mathbf{u}^1,\ldots,\mathbf{u}^p\}}\|}$. Thus, facts of ${}_{\mathbf{T}}\mathrm{R}_{\{\mathbf{1},\mathbf{v}(w_{10}^2,\mathbf{w}_1^2),\ldots,\mathbf{v}(w_{p0}^2,\mathbf{w}_p^2)\}} \neq \mathbf{0}$ and ${}_{\mathbf{T}}\mathrm{R}_{\{\mathbf{1},\mathbf{v}(w_{10}^2,\mathbf{w}_1^2),\ldots,\mathbf{v}(w_{p0}^2,\mathbf{w}_p^2),\mathbf{v}(\bar{w}_0,\bar{\mathbf{w}}_1^2)\}} = \mathbf{0}$ imply that $\{\mathbf{u}^0, \mathbf{u}^1, \ldots, \mathbf{u}^p, \bar{\mathbf{u}}^{p+1}\}$ is an orthonormal set, $\mathbf{T} = \sum_{j=0}^{p} (\mathbf{u}^j)^{\mathrm{T}} \mathbf{T}\, \mathbf{u}^j + (\bar{\mathbf{u}}^{p+1})^{\mathrm{T}} \mathbf{T}\, \bar{\mathbf{u}}^{p+1}$, and $(\bar{\mathbf{u}}^{p+1})^{\mathrm{T}} \mathbf{T} \neq 0$.

Suppose $\mathbf{v}(w_0,\gamma\mathbf{w}_1^2) \equiv \sum_{j=0}^{p} (\mathbf{u}^j)^{\mathrm{T}}\, \mathbf{v}(w_0,\gamma\mathbf{w}_1^2)\, \mathbf{u}^j + (\bar{\mathbf{u}}^{p+1})^{\mathrm{T}}\, \mathbf{v}(w_0,\gamma\mathbf{w}_1^2)\, \bar{\mathbf{u}}^{p+1} + {}_{\mathbf{v}(w_0,\gamma\mathbf{w}_1^2)}\mathrm{R}_{\{\mathbf{u}^0,\mathbf{u}^1,\ldots,\mathbf{u}^p,\bar{\mathbf{u}}^{p+1}\}}$.

Thus, $\|{}_{\mathbf{T}}\mathrm{R}_{\{\mathbf{u}^0,\mathbf{u}^1,\ldots,\mathbf{u}^p,\mathbf{v}(w_0,\gamma\mathbf{w}_1^2)\}}\|^2 = \frac{\mathbf{T}^{\mathrm{T}}\mathbf{A}^p\mathbf{T}(\mathbf{v}(w_0,\gamma\mathbf{w}_1^2))^{\mathrm{T}}\mathbf{A}^p\mathbf{v}(w_0,\gamma\mathbf{w}_1^2) - ((\mathbf{v}(w_0,\gamma\mathbf{w}_1^2))^{\mathrm{T}}\mathbf{A}^p\mathbf{T})^2}{(\mathbf{v}(w_0,\gamma\mathbf{w}_1^2))^{\mathrm{T}}\mathbf{A}^p\mathbf{v}(w_0,\gamma\mathbf{w}_1^2)} =$

$\frac{1}{(\mathbf{v}(w_0,\gamma\mathbf{w}_1^2))^{\mathrm{T}}\mathbf{A}^p\mathbf{v}(w_0,\gamma\mathbf{w}_1^2)}$ $[(\mathbf{T}^{\mathrm{T}}\bar{\mathbf{u}}^{p+1})^2\, (\mathbf{v}(w_0,\gamma\mathbf{w}_1^2))^{\mathrm{T}}\, ((\bar{\mathbf{u}}^{p+1})^{\mathrm{T}}\, \mathbf{v}(w_0,\gamma\mathbf{w}_1^2)\, \bar{\mathbf{u}}^{p+1} + {}_{\mathbf{v}(w_0,\gamma\mathbf{w}_1^2)}\mathrm{R}_{\{\mathbf{u}^0,\mathbf{u}^1,\ldots,\mathbf{u}^p,\bar{\mathbf{u}}^{p+1}\}})$

$- ((\bar{\mathbf{u}}^{p+1})^{\mathrm{T}}\, \mathbf{T}\, (\bar{\mathbf{u}}^{p+1})^{\mathrm{T}}\, \mathbf{v}(w_0,\gamma\mathbf{w}_1^2))^2] = \frac{(\mathbf{T}^{\mathrm{T}}\bar{\mathbf{u}}^{p+1})^2}{(\mathbf{v}(w_0,\gamma\mathbf{w}_1^2))^{\mathrm{T}}\mathbf{A}^p\mathbf{v}(w_0,\gamma\mathbf{w}_1^2)}$ $[((((\bar{\mathbf{u}}^{p+1})^{\mathrm{T}}\, \mathbf{v}(w_0,\gamma\mathbf{w}_1^2))^2 +$

$\|{}_{\mathbf{v}(w_0,\gamma\mathbf{w}_1^2)}\mathrm{R}_{\{\mathbf{u}^0,\mathbf{u}^1,\ldots,\mathbf{u}^p,\bar{\mathbf{u}}^{p+1}\}}\|^2) - ((\bar{\mathbf{u}}^{p+1})^{\mathrm{T}}\, \mathbf{v}(w_0,\gamma\mathbf{w}_1^2))^2\,] = \frac{(\mathbf{T}^{\mathrm{T}}\bar{\mathbf{u}}^{p+1})^2}{(\mathbf{v}(w_0,\gamma\mathbf{w}_1^2))^{\mathrm{T}}\mathbf{A}^p\mathbf{v}(w_0,\gamma\mathbf{w}_1^2)}\|{}_{\mathbf{v}(w_0,\gamma\mathbf{w}_1^2)}\mathrm{R}_{\{\mathbf{u}^0,\mathbf{u}^1,\ldots,\mathbf{u}^p,\bar{\mathbf{u}}^{p+1}\}}\|^2$.

If $\|{}_{\mathbf{v}(w_0,\gamma\mathbf{w}_1^2)}\mathrm{R}_{\{\mathbf{u}^0,\mathbf{u}^1,\ldots,\mathbf{u}^p,\bar{\mathbf{u}}^{p+1}\}}\|^2 > 0$, then $\|{}_{\mathbf{T}}\mathrm{R}_{\{\mathbf{u}^0,\mathbf{u}^1,\ldots,\mathbf{u}^p,\mathbf{v}(w_0,\gamma\mathbf{w}_1^2)\}}\|^2 > \left\|{}_{\mathbf{T}}\mathrm{R}_{\{\mathbf{u}^0,\mathbf{u}^1,\ldots,\mathbf{u}^p,\mathbf{v}(\bar{w}_0,\bar{\mathbf{w}})\}}\right\|^2$ because of a non-zero $(\bar{\mathbf{u}}^{p+1})^{\mathrm{T}}\, \mathbf{T}$. Therefore, the optimization of $\min_{w_0,\gamma} \|{}_{\mathbf{T}}\mathrm{R}_{\{\mathbf{1},\mathbf{v}(w_{10}^2,\mathbf{w}_1^2),\ldots,\mathbf{v}(w_{p0}^2,\mathbf{w}_p^2),\mathbf{v}(w_0,\gamma\mathbf{w}_1^2)\}}\|^2$ leads to a non-zero ${}_{\mathbf{v}(w_0^*,\gamma^*\mathbf{w}_1^2)}\mathrm{R}_{\{\mathbf{1},\mathbf{v}(w_{10}^2,\mathbf{w}_1^2),\ldots,\mathbf{v}(w_{p0}^2,\mathbf{w}_p^2)\}}$ such that $\|{}_{\mathbf{v}(w_0^*,\gamma^*\mathbf{w}_1^2)}\mathrm{R}_{\{\mathbf{u}^0,\mathbf{u}^1,\ldots,\mathbf{u}^p,\bar{\mathbf{u}}^{p+1}\}}\|^2 = 0$. From **Lemma 6**, $\frac{{}_{\mathbf{v}(w_0^*,\gamma^*\mathbf{w}_1^2)}\mathrm{R}_{\{\mathbf{u}^0,\mathbf{u}^1,\ldots,\mathbf{u}^p\}}}{\|{}_{\mathbf{v}(w_0^*,\gamma^*\mathbf{w}_1^2)}\mathrm{R}_{\{\mathbf{u}^0,\mathbf{u}^1,\ldots,\mathbf{u}^p\}}\|}$ and $\frac{{}_{\mathbf{v}(\bar{w}_0,\bar{\mathbf{w}})}\mathrm{R}_{\{\mathbf{u}^0,\mathbf{u}^1,\ldots,\mathbf{u}^p\}}}{\|{}_{\mathbf{v}(\bar{w}_0,\bar{\mathbf{w}})}\mathrm{R}_{\{\mathbf{u}^0,\mathbf{u}^1,\ldots,\mathbf{u}^p\}}\|}$ are parallel and thus, $\|{}_{\mathbf{T}}\mathrm{R}_{\{\mathbf{1},\mathbf{v}(w_{10}^2,\mathbf{w}_1^2),\ldots,\mathbf{v}(w_{p0}^2,\mathbf{w}_p^2),\mathbf{v}(w_0^*,\gamma^*\mathbf{w}_1^2)\}}\|^2 = 0$. **Q.E.D.**

References

1. M. Me'zard and J. Nadal, *Learning in feedforward layered networks: The tiling algorithm. Journal of Physics.* **A 22**, 2191– 2204 (1989).
2. S. Fahlman and C. Lebiere, *The Cascade-Correlation Learning Architecture. In Touretzky, D. (Eds.), Advances in Neural Information Processing Systems II (Denver, 1989) Morgan Kaufmann, San Mateo* (1990).
3. M. Frean, *The Upstart Algorithm: A Method for Constructing and Training Feedforward Neural Networks. Neural Computation*. **2**, 198– 209(1990).
4. G. Huang and H. Babri, *Upper bounds on the number of hidden neurons in feedforward networks with arbitrary bounded nonlinear activation functions. IEEE Transactions on Neural Networks*. **9**, 224– 229(1998).
5. D. Rumelhart, G. Hinton, and R. Williams, "Modeling Internal Representations By Error Propagation," in *Parallel Distributed Processing: Explorations in the Microstructure of*

Cognition, vol. 1, D. Rumelhart and J. McClelland, Eds. Cambridge, MA: MIT Press, 1986, pp. 318-362.
6. P. J. Rousseeuw and A. M. Leroy, *Robust Regression and Outlier Detection*. New York: Wiley, 1987.
7. R. D. Cook and S. Weisberg, *Residuals and Influence in Regression*, London: Chapman and Hall, 1982.
8. A. C. Atkinson, *Plots, Transformations and Regression*, Oxford: Oxford University Press, 1985.
9. F. R. Hampel, E. M. Ronchetti, P. J. Rousseeuw and W. A. Stahel, *Robust Statistics: The Approach Based on Influence Functions*. New York: John Wiley, 1986.
10. A. Zaman, P. J. Rousseeuw and M. Orhan, **"**Econometric Applications of High-Breakdown Robust Regression Techniques," *Econometrics Letters*, vol. 71, pp. 1-8, Apr. 2001.
11. A. C. Atkinson and T. C. Cheng, "Computing Least Trimmed Squares Regression with the Forward Search," *Statistics and Computing*, vol. 9, pp. 251-263, Nov. 1999.
12. A. C. Atkinson, "Fast Very Robust Methods for the Detection of Multiple Outliers," *Journal of the American Statistical Association*, vol. 89, pp. 1329-1339, 1994.
13. P. J. Rousseeuw and K. Van Driessen, "Computing LTS Regression for Large Data Sets," *Estadistica*, vol. 54, 163-190, 2002.
14. A. J. Stromberg, "Computation of High Breakdown Nonlinear Regression Parameters," Journal of the American Statistical Association, vol. 88, pp. 237-244, Mar. 1993.

計畫成果自評：

此研究計畫成果豐碩，已被送到相關研討會發表。其延伸之研究亦在進行中。

出席國際學術會議心得報告

計畫編號	95-2416-H-004-049
計畫名稱	相關性隱藏節點與學習演算法
出國人員姓名 服務機關及職稱	蔡瑞煌 國立政治大學資訊管理學系,教授
會議時間地點	July 18-24, 2007, Salt Lake City, U.S.A.
會議名稱	Joint Conference on Information Sciences
發表論文題目	A Constructive Learning Procedure

一、參加會議經過

我於 7/19 晚上到達 Salt Lake City。我於 7/20 主持 section CIEF-III 並於其間發表論文。附件一是相關議程。

我也於 7/20~7/21 聆聽此次會議裡的 Keynote Speeches 和論文發表。

二、與會心得

Keynote Speeches 邀請了不少的資訊科學學術界裡之知名學者來演講，我受益不少。

A Constructive Learning Procedure*

RAY TSAIH

Department of Management Information Systems, National Chengchi University, No.64, Sec. 2, Jhihnan Rd., Wunshan District,Taipei City 116, Taiwan

This study explores a deterministic learning procedure for the realization of a real-valued single-hidden layer feed-forward neural network (SLFN) with tanh activation functions of the hidden-layer nodes for arbitrary mapping problems.

1. The Constructive Learning Procedure

The proposed learning procedure will use none hidden node initially and recruit (add) more hidden nodes during the learning process. The goal of the proposed constructive learning procedure is to create a SLFN for fitting perfectly all given mapping samples. In the literature, there are some similar procedures; for instance, the tiling algorithm for binary-valued layered feed-forward neural networks (cf. [1]), the cascade-correlation algorithm (cf. [2]), and the upstart algorithm for binary-valued layered feed-forward neural networks (cf. [3]). In contrast to these researches, this study copes with the real-valued SLFN.

2. The Mapping Requirement and Notations

An SLFN provides a nonlinear mapping between $\mathbf{x}$ and y, whose form is $y = f(\mathbf{x})$. f is a nonlinear function whose parameters (i.e., weights and biases) are derived from a given data set of mapping samples $\{(\mathbf{x}^1, t^1),\ldots, (\mathbf{x}^N, t^N)\}$ with $\mathbf{x}^{c1} \neq \mathbf{x}^{c2}$, $c_1 \neq c_2$, and with t^c the target value of y corresponding to $\mathbf{x}^c$. $\mathbf{x} \equiv (x_1, x_2,\ldots, x_m)^T \in R^m$ where x_j is the j^{th} input component, with j from 1 to m.

Hereafter, m and p denote the numbers of adopted input and hidden nodes, repsectively; w_{j0}^2 stands for the bias of the j^{th} hidden node; $\mathbf{w}_j^2 \equiv (w_{j1}^2, w_{j2}^2,\ldots, w_{jm}^2)^T$ for the weights between the j^{th} hidden node and input layer; w_0^3 for the bias of the output node; and $\mathbf{w}^3 \equiv (w_1^3, w_2^3,\ldots, w_p^3)^T$ for the weights between the output node and all hidden nodes. Characters in bold represent column vectors; the superscript T indicates transposition.

Let the $tanh(t)$ activation function be used by all hidden nodes and a linear activation function be used by the output node. Thus, given the c^{th} sample $\mathbf{x}^c$, the activation value of the j^{th} hidden node $a^c(w_{j0}^2, \mathbf{w}_j^2)$ and the output value y^c are as follows:

$$a^c(w_{j0}^2, \mathbf{w}_j^2) \equiv tanh(w_{j0}^2 + \sum_{i=1}^{m} w_{ji}^2 x_i^c); \quad (1)$$

$$y^c \equiv w_0^3 + \sum_{j=1}^{p} w_j^3 a^c(w_{j0}^2, \mathbf{w}_j^2). \quad (2)$$

Given the N mapping samples, let $\mathbf{v}(w_{j0}^2, \mathbf{w}_j^2) \equiv (a^1(w_{j0}^2, \mathbf{w}_j^2), a^2(w_{j0}^2, \mathbf{w}_j^2),\ldots, a^N(w_{j0}^2, \mathbf{w}_j^2))^T \in (-1,1)^N$ be the responding vector of the j^{th} hidden node with the c^{th} component being $a^c(w_{j0}^2, \mathbf{w}_j^2)$. Furthermore, let $\mathbf{1}$ be a $N \times 1$ vector with all components 1 and $\mathbf{T} \equiv (t^1, t^2, \ldots, t^N)^T \in R^N$. Thus, the set of simultaneous equations $w_0^3 + \sum_{j=1}^{p} w_j^3 tanh(w_{j0}^2 + \sum_{i=1}^{m} w_{ji}^2 x_i^c) = t^c \ \forall\ c = 1, \ldots, N$ is equivalent to system (3), which states that $\mathbf{T}$ is in the (sub-)space spanned by $\mathbf{1}$ and p responding vectors,

* This study is supported by National Science Council of R.O.C. under Grants No. NSC 92-2416-H-004-004, NSC 93-2416-H-004-015, and NSC 43028F.

$\{\mathbf{v}(w_{j0}^2, \mathbf{w}_j^2), j = 1, \ldots, p\}$.

$$w_0^3 \mathbf{1} + \sum_{j=1}^{p} w_j^3 \mathbf{v}(w_{j0}^2, \mathbf{w}_j^2) = \mathbf{T}. \quad (3)$$

Hereafter, ${}_{\mathbf{a}}\mathrm{R}_{\mathbf{b}} \equiv \mathbf{a} - \frac{\mathbf{b}^{\mathrm{T}}\mathbf{a}}{\|\mathbf{b}\|} \frac{\mathbf{b}}{\|\mathbf{b}\|}$ denotes the residual of vector **a** regarding **b** after the part parallel with the vector **b** has been taken away. Furthermore, ${}_{\mathbf{a}}\mathrm{R}_{\{\mathbf{b}^1,\ldots,\mathbf{b}^k\}} \equiv {}_{\mathbf{a}}\mathrm{R}_{\{\mathbf{b}^1,\ldots,\mathbf{b}^{k-1}\}} - \frac{\mathbf{a}^{\mathrm{T}}{}_{\mathbf{b}^k}\mathrm{R}_{\{\mathbf{b}^1,\ldots,\mathbf{b}^{k-1}\}}}{\|{}_{\mathbf{b}^k}\mathrm{R}_{\{\mathbf{b}^1,\ldots,\mathbf{b}^{k-1}\}}\|} \frac{{}_{\mathbf{b}^k}\mathrm{R}_{\{\mathbf{b}^1,\ldots,\mathbf{b}^{k-1}\}}}{\|{}_{\mathbf{b}^k}\mathrm{R}_{\{\mathbf{b}^1,\ldots,\mathbf{b}^{k-1}\}}\|}$ denotes the residual of vector **a** regarding the ordered sequence of vectors $\{\mathbf{b}^1,\ldots, \mathbf{b}^k\}$ after the parts parallel with vectors $\mathbf{b}^j$s have been taken away.

Note that ${}_{\mathbf{a}}\mathrm{R}_{\{\mathbf{b}^1,\ldots,\mathbf{b}^k\}} = \mathbf{a} - \frac{\mathbf{a}^{\mathrm{T}}\mathbf{b}^1}{\|\mathbf{b}^1\|} \frac{\mathbf{b}^1}{\|\mathbf{b}^1\|} - \sum_{j=2}^{k} \frac{\mathbf{a}^{\mathrm{T}}{}_{\mathbf{b}^j}\mathrm{R}_{\{\mathbf{b}^1,\ldots,\mathbf{b}^{j-1}\}}}{\|{}_{\mathbf{b}^j}\mathrm{R}_{\{\mathbf{b}^1,\ldots,\mathbf{b}^{j-1}\}}\|} \frac{{}_{\mathbf{b}^j}\mathrm{R}_{\{\mathbf{b}^0,\mathbf{b}^1,\ldots,\mathbf{b}^{j-1}\}}}{\|{}_{\mathbf{b}^j}\mathrm{R}_{\{\mathbf{b}^0,\mathbf{b}^1,\ldots,\mathbf{b}^{j-1}\}}\|}$ and thus $\mathbf{a} = \frac{\mathbf{a}^{\mathrm{T}}\mathbf{b}^1}{\|\mathbf{b}^1\|} \frac{\mathbf{b}^1}{\|\mathbf{b}^1\|} + \sum_{j=2}^{k} \frac{\mathbf{a}^{\mathrm{T}}{}_{\mathbf{b}^j}\mathrm{R}_{\{\mathbf{b}^1,\ldots,\mathbf{b}^{j-1}\}}}{\|{}_{\mathbf{b}^j}\mathrm{R}_{\{\mathbf{b}^1,\ldots,\mathbf{b}^{j-1}\}}\|} \frac{{}_{\mathbf{b}^j}\mathrm{R}_{\{\mathbf{b}^0,\mathbf{b}^1,\ldots,\mathbf{b}^{j-1}\}}}{\|{}_{\mathbf{b}^j}\mathrm{R}_{\{\mathbf{b}^0,\mathbf{b}^1,\ldots,\mathbf{b}^{j-1}\}}\|} + {}_{\mathbf{a}}\mathrm{R}_{\{\mathbf{b}^1,\ldots,\mathbf{b}^k\}}$. Furthermore, we have **Lemma 1** below, which lists some properties associated with the above proposed residual vector.

Lemma 1: (i) If **a** is linearly dependent with **b**, then ${}_{\mathbf{a}}\mathrm{R}_{\mathbf{b}} = \mathbf{0}$. Similarly, if **a** can be linearly represented by the set of vectors $\{\mathbf{b}^1,\ldots, \mathbf{b}^k\}$,[1] then ${}_{\mathbf{a}}\mathrm{R}_{\{\mathbf{b}^1,\ldots,\mathbf{b}^k\}} = \mathbf{0}$. (ii) ${}_{\mathbf{a}}\mathrm{R}_{\mathbf{b}}$ is orthogonal to **b**. Similarly, ${}_{\mathbf{a}}\mathrm{R}_{\{\mathbf{b}^1,\ldots,\mathbf{b}^k\}}$ is orthogonal to all vectors in the set $\{\mathbf{b}^1,\ldots, \mathbf{b}^k\}$. (iii) If ${}_{\mathbf{a}}\mathrm{R}_{\{\mathbf{b}^1,\ldots,\mathbf{b}^k\}} = \mathbf{0}$, then ${}_{\mathbf{a}}\mathrm{R}_{\{\mathbf{b}^1,\ldots,\mathbf{b}^{k+1}\}} = \mathbf{0}$.

3. The Proposed Constructive Procedure

Table 1 presents the proposed procedure for constructing a SLFN appropriate for fitting the mapping embedded in $\{(\mathbf{x}^1, t^1), \ldots, (\mathbf{x}^N, t^N)\}$.

Table 1. The proposed deterministic procedure for constructing an appropriate SLFN for the mapping requirement of $\{(\mathbf{x}^1, t^1), \ldots, (\mathbf{x}^N, t^N)\}$. $\mathbf{T} \equiv (t^1, t^2, \ldots, t^N)^{\mathrm{T}}$ and $\mathbf{v}(w_0,\mathbf{w}) \equiv (a^1(w_0,\mathbf{w}), a^2(w_0,\mathbf{w}), \ldots, a^N(w_0,\mathbf{w}))^{\mathrm{T}}$ is a $N \times 1$ vector with the c^{th} component being $a^c(w_0,\mathbf{w}) \equiv tanh(\mathbf{w}^{\mathrm{T}} \mathbf{x}^c + w_0)$, in which $\mathbf{w} \equiv (w_1, w_2, \ldots, w_m)^{\mathrm{T}}$.

Step 1: Calculate $\mathbf{u}^0 = \frac{\mathbf{1}}{\|\mathbf{1}\|}$ and ${}_{\mathbf{T}}\mathrm{R}_{\mathbf{u}^0} \equiv \mathbf{T} - (\mathbf{u}^0)^{\mathrm{T}} \mathbf{T} \mathbf{u}^0$.

Step 2: If ${}_{\mathbf{T}}\mathrm{R}_{\mathbf{u}^0} = \mathbf{0}$, then (i) claim that the fitting job requests no hidden node; (ii) set the bias of output node as $\frac{\mathbf{T}^{\mathrm{T}}\mathbf{u}^0}{\mathbf{1}^{\mathrm{T}}\mathbf{u}^0}$ and the weight vector between the output node and the input layer as **0**; and (iii) stop.

Step 3: If ${}_{\mathbf{T}}\mathrm{R}_{\mathbf{u}^0} \neq \mathbf{0}$, then solve $\min_{w_0,\mathbf{w}} \|{}_{\mathbf{T}}\mathrm{R}_{\{\mathbf{v}(w_0,\mathbf{w})\}}\|^2$ and let w_0^* and $\mathbf{w}^*$ be the obtained optimal solution.

[1] Namely, **a** is in the (sub-)space spanned by $\{\mathbf{b}^1,\ldots, \mathbf{b}^k\}$.

Then set $w_{10}^2 = w_0^*$, $\mathbf{w}_1^2 = \mathbf{w}^*$, $\mathbf{u}^1 \equiv \dfrac{\mathbf{v}(w_{10}^2, \mathbf{w}_1^2)}{\| \mathbf{v}(w_{10}^2, \mathbf{w}_1^2) \|}$, and $p = 1$.

Step 4: Calculate ${}_{\mathbf{T}}\mathrm{R}_{\{\mathbf{u}^1,\ldots,\mathbf{u}^p\}}$.

Step 5: If ${}_{\mathbf{T}}\mathrm{R}_{\{\mathbf{u}^1,\ldots,\mathbf{u}^p\}} = \mathbf{0}$, then (i) claim that the fitting job requests p hidden nodes with the bias and weights of the j^{th} hidden node being the above w_{j0}^2 and $\mathbf{w}_j^2$, $\forall\, j = 1, \ldots, p$; (ii) set $w_0^3 \equiv 0$,

$$w_p^3 \equiv \frac{\mathbf{T}^{\mathrm{T}}\mathbf{u}^p}{\mathbf{v}(w_{p0}^2, \mathbf{w}_p^2)^{\mathrm{T}}\mathbf{u}^p}, \text{ and } w_j^3 \equiv \frac{(\mathbf{T} - \sum_{l=j+1}^{p} w_l^3 \mathbf{v}(w_{l0}^2, \mathbf{w}_l^2))^{\mathrm{T}}\mathbf{u}^j}{\mathbf{v}(w_{j0}^2, \mathbf{w}_j^2)^{\mathrm{T}}\mathbf{u}^j} \text{ for all } 1 \le j \le p\text{-}1; \text{ and}$$

(iii) stop.

Step 6: If ${}_{\mathbf{T}}\mathrm{R}_{\{\mathbf{u}^1,\ldots,\mathbf{u}^p\}} \neq \mathbf{0}$, then calculate $\mathbf{u}^0 = \dfrac{{}_{\mathbf{1}}\mathrm{R}_{\{\mathbf{u}^1,\ldots,\mathbf{u}^p\}}}{\| {}_{\mathbf{1}}\mathrm{R}_{\{\mathbf{u}^1,\ldots,\mathbf{u}^p\}} \|}$ and ${}_{\mathbf{T}}\mathrm{R}_{\{\mathbf{u}^1,\ldots,\mathbf{u}^p,\mathbf{u}^0\}} \equiv {}_{\mathbf{T}}\mathrm{R}_{\{\mathbf{u}^1,\ldots,\mathbf{u}^p\}} - (\mathbf{u}^0)^{\mathrm{T}}\,\mathbf{T}\,\mathbf{u}^0$.

Step 7: If ${}_{\mathbf{T}}\mathrm{R}_{\{\mathbf{u}^1,\ldots,\mathbf{u}^p,\mathbf{u}^0\}} = \mathbf{0}$, then (i) claim that the fitting job requests p hidden nodes with the bias and weights of the j^{th} hidden node being the above w_{j0}^2 and $\mathbf{w}_j^2$, $\forall\, j = 1, \ldots, p$; (ii) set

$$w_0^3 \equiv \frac{\mathbf{T}^{\mathrm{T}}\mathbf{u}^0}{\mathbf{1}^{\mathrm{T}}\mathbf{u}^0}, \quad w_p^3 \equiv \frac{(\mathbf{T} - w_0^3\mathbf{1})^{\mathrm{T}}\mathbf{u}^p}{\mathbf{v}(w_{p0}^2, \mathbf{w}_p^2)^{\mathrm{T}}\mathbf{u}^p}, \text{ and } w_j^3 \equiv \frac{(\mathbf{T} - w_0^3\mathbf{1} - \sum_{l=j+1}^{p} w_l^3 \mathbf{v}(w_{l0}^2, \mathbf{w}_l^2))^{\mathrm{T}}\mathbf{u}^j}{\mathbf{v}(w_{j0}^2, \mathbf{w}_j^2)^{\mathrm{T}}\mathbf{u}^j}$$

for all $1 \le j \le p$-1; and (iii) stop.

Step 8: If ${}_{\mathbf{T}}\mathrm{R}_{\{\mathbf{u}^1,\ldots,\mathbf{u}^p,\mathbf{u}^0\}} \neq \mathbf{0}$, then solve $\min_{w_0,\mathbf{w}} \| {}_{\mathbf{T}}\mathrm{R}_{\{\mathbf{u}^1,\ldots,\mathbf{u}^p,\mathbf{v}(w_0,\mathbf{w})\}} \|^2$ and let w_0^* and $\mathbf{w}^*$ be the obtained optimal solution. Then (i) set $w_{p+1,0}^2 = w_0^*$, $\mathbf{w}_{p+1}^2 = \mathbf{w}^*$, $\mathbf{u}^{p+1} \equiv \dfrac{{}_{\mathbf{v}(w_{p+1,0}^2, \mathbf{w}_{p+1}^2)}\mathrm{R}_{\{\mathbf{u}^1,\ldots,\mathbf{u}^p\}}}{\| {}_{\mathbf{v}(w_{p+1,0}^2, \mathbf{w}_{p+1}^2)}\mathrm{R}_{\{\mathbf{u}^1,\ldots,\mathbf{u}^p\}} \|}$, set $p+1 \rightarrow p$; and (ii) go to Step 4.

4. The Correctness of the Proposed Procedure

We now prove that the procedure stated in Table 1 creates an appropriate SLFN that meets the mapping requirement of $\{(\mathbf{x}^1, t^1), \ldots, (\mathbf{x}^N, t^N)\}$ without error.

Lemma 2 below is obvious from system (3) and the definition of the residual vector. **Lemma 2** states the necessary condition of a SLFN with p hidden nodes appropriate for the mapping requirement of $\{(\mathbf{x}^1, t^1), \ldots, (\mathbf{x}^N, t^N)\}$. Note that, from Lemma 1 (iii), ${}_{\mathbf{T}}\mathrm{R}_{\{\mathbf{v}(w_{10}^2, \mathbf{w}_1^2),\ldots,\mathbf{v}(w_{p0}^2, \mathbf{w}_p^2)\}} = \mathbf{0}$ results in ${}_{\mathbf{T}}\mathrm{R}_{\{\mathbf{v}(w_{10}^2, \mathbf{w}_1^2),\ldots,\mathbf{v}(w_{p0}^2, \mathbf{w}_p^2),\mathbf{1}\}} = \mathbf{0}$. Thus the condition stated in Steps 2, 5, and 7 are suitable stopping criteria.

Lemma 2: Regarding the mapping samples of $\{(\mathbf{x}^1, t^1), \ldots, (\mathbf{x}^N, t^N)\}$, the SLFN with p hidden nodes is appropriate for the mapping requirement if the associated ${}_{\mathbf{T}}\mathrm{R}_{\{\mathbf{v}(w_{10}^2,\mathbf{w}_1^2),\ldots,\mathbf{v}(w_{p0}^2,\mathbf{w}_p^2),\mathbf{1}\}}$ equals $\mathbf{0}$.

Namely, if ${}_{\mathbf{T}}\mathrm{R}_{\mathbf{u}^0} = \mathbf{0}$ at Step 2 of Table 1, in which $\mathbf{u}^0 = \frac{\mathbf{1}}{\|\mathbf{1}\|}$, then $\mathbf{T} = \frac{\mathbf{T}^{\mathrm{T}}\mathbf{u}^0}{\mathbf{1}^{\mathrm{T}}\mathbf{u}^0}$ and the mapping requirement asks for no hidden node. Furthermore, because $\mathbf{u}^1 \equiv \frac{\mathbf{v}(w_{10}^2, \mathbf{w}_1^2)}{\|\mathbf{v}(w_{10}^2, \mathbf{w}_1^2)\|}$, $\mathbf{u}^j \equiv \frac{\mathbf{v}(w_{j0}^2,\mathbf{w}_j^2)\mathrm{R}_{\{\mathbf{u}^1,\ldots,\mathbf{u}^{j-1}\}}}{\|_{\mathbf{v}(w_{j0}^2,\mathbf{w}_j^2)}\mathrm{R}_{\{\mathbf{u}^1,\ldots,\mathbf{u}^{j-1}\}}\|}$ for all $j = 2, \ldots, p$, and $\mathbf{u}^0 \equiv \frac{{}_{\mathbf{1}}\mathrm{R}_{\{\mathbf{u}^1,\ldots,\mathbf{u}^p\}}}{\|{}_{\mathbf{1}}\mathrm{R}_{\{\mathbf{u}^1,\ldots,\mathbf{u}^p\}}\|}$, ${}_{\mathbf{T}}\mathrm{R}_{\{\mathbf{v}(w_{10}^2,\mathbf{w}_1^2),\ldots,\mathbf{v}(w_{p0}^2,\mathbf{w}_p^2)\}} = {}_{\mathbf{T}}\mathrm{R}_{\{\mathbf{u}1,\ldots,\mathbf{u}p\}}$ and ${}_{\mathbf{T}}\mathrm{R}_{\{\mathbf{v}(w_{10}^2,\mathbf{w}_1^2),\ldots,\mathbf{v}(w_{p0}^2,\mathbf{w}_p^2),\mathbf{1}\}} = {}_{\mathbf{T}}\mathrm{R}_{\{\mathbf{u}1,\ldots,\mathbf{u}p,\mathbf{u}0\}}$. Thus, if ${}_{\mathbf{T}}\mathrm{R}_{\{\mathbf{u}1,\ldots,\mathbf{u}p\}} = \mathbf{0}$ at Step 5 of Table 1, then $\mathbf{T} = \sum_{j=1}^{p} w_j^3\,\mathbf{v}(w_{j0}^2, \mathbf{w}_j^2)$ and the mapping requirement asks for p hidden nodes. If ${}_{\mathbf{T}}\mathrm{R}_{\{\mathbf{u}1,\ldots,\mathbf{u}p,\mathbf{u}0\}} = \mathbf{0}$ at Step 7 of Table 1, then $\mathbf{T} = w_0^3\,\mathbf{1} + \sum_{j=1}^{p} w_j^3\,\mathbf{v}(w_{j0}^2, \mathbf{w}_j^2)$ and the mapping requirement asks for p hidden nodes.

The following **Lemmas 3** and **4** show that the proposed procedure will generate a sequence of orthonormal vectors.

Lemma 3: If $p < N$-1, then there always exist w_0 and $\mathbf{w}$ such that ${}_{\mathbf{v}(w_0,\mathbf{w})}\mathrm{R}_{\{\mathbf{u}1,\ldots,\mathbf{u}p,\mathbf{u}0\}} \neq \mathbf{0}$.

Proof of Lemma 3: Suppose $p < N$-1 and there are no w_0 and $\mathbf{w}$ such that ${}_{\mathbf{v}(w_0,\mathbf{w})}\mathrm{R}_{\{\mathbf{u}1,\ldots,\mathbf{u}p,\mathbf{u}0\}} \neq \mathbf{0}$. In other words, $p < N$-1 and there are no w_0 and $\mathbf{w}$ such that $\mathbf{v}(w_0, \mathbf{w})$ is linearly independent with $\{\mathbf{u}^1, \ldots, \mathbf{u}^p, \mathbf{u}^0\}$. This contradicts with the statement of [4] that, for any given $\{\mathbf{x}^c\ \forall\ c = 1, \ldots, N\}$, there exists a set of $\{(w_{j0}^2, \mathbf{w}_j^2),\ j = 1, \ldots, N\text{-}1\}$ such that the associated square matrix $(\mathbf{1}, \mathbf{v}(w_{10}^2, \mathbf{w}_1^2), \ldots, \mathbf{v}(w_{N-1,0}^2, \mathbf{w}_{N-1}^2))$ is invertible. **Q.E.D.**

Lemma 4: When the proposed procedure stops at some p, $1 \le p \le N$-1, all vectors in the set of $\{\mathbf{u}^1, \ldots, \mathbf{u}^p\}$ or $\{\mathbf{u}^1, \ldots, \mathbf{u}^p, \mathbf{u}^0\}$ are orthonormal.

Proof of Lemma 4: It is trivial for the case of $p = 1$ and $\{\mathbf{u}^1\}$. As for the case of $p = 1$ and $\{\mathbf{u}^1, \mathbf{u}^0\}$, from steps 6 and 7 of Table 1, ${}_{\mathbf{T}}\mathrm{R}_{\mathbf{u}1} \neq \mathbf{0}$, $\mathbf{u}^0 \equiv \frac{{}_{\mathbf{1}}\mathrm{R}_{\mathbf{u}^1}}{\|{}_{\mathbf{1}}\mathrm{R}_{\mathbf{u}^1}\|}$, and ${}_{\mathbf{T}}\mathrm{R}_{\{\mathbf{u}^1,\mathbf{u}^0\}} = \mathbf{0}$. Thus $\mathbf{u}^0$ is non-zero and, from **Lemma 1** (ii), orthogonal to $\mathbf{u}^1$. Namely, $\{\mathbf{u}^1, \mathbf{u}^0\}$ is an orthonormal set.

Now consider the case of any p with $2 \le p \le N$-1. Then, for each $1 \le j \le p$-1, from **Lemma 3**, there exists w_0 and $\mathbf{w}$ such that ${}_{\mathbf{v}(w_0,\mathbf{w})}\mathrm{R}_{\{\mathbf{u}1,\ldots,\mathbf{u}j,\,\mathbf{u}0\}}$ is not zero and thus, from **Lemma 1** (iii), ${}_{\mathbf{v}(w_0,\mathbf{w})}\mathrm{R}_{\{\mathbf{u}1,\ldots,\mathbf{u}j\}}$ is not zero. Furthermore, as stated in Step 8 of Table 1, ${}_{\mathbf{T}}\mathrm{R}_{\{\mathbf{u}1,\ldots,\mathbf{u}j,\mathbf{u}0\}} \neq \mathbf{0}$. Thus ${}_{\mathbf{T}}\mathrm{R}_{\{\mathbf{u}1,\ldots,\mathbf{u}j\}} \neq \mathbf{0}$, there exists a non-zero ${}_{\mathbf{v}(w_0,\mathbf{w})}\mathrm{R}_{\{\mathbf{u}1,\ldots,\mathbf{u}j\}}$ that ${}_{\mathbf{T}}\mathrm{R}_{\{\mathbf{u}1,\ldots,\mathbf{u}j\}} = {}_{\mathbf{T}}\mathrm{R}_{\{\mathbf{u}1,\ldots,\mathbf{u}j,\mathbf{v}(w_0,\mathbf{w})\}} + \frac{\mathbf{T}^{\mathrm{T}}{}_{\mathbf{v}(w_0,\mathbf{w})}\mathrm{R}_{\{\mathbf{u}^1,\ldots,\mathbf{u}^j\}}}{\|_{\mathbf{v}(w_0,\mathbf{w})}\mathrm{R}_{\{\mathbf{u}^1,\ldots,\mathbf{u}^j\}}\|^2}\ {}_{\mathbf{v}(w_0,\mathbf{w})}\mathrm{R}_{\{\mathbf{u}1,\ldots,\mathbf{u}j\}}$, and, from **Lemma 1** (ii), ${}_{\mathbf{v}(w_0,\mathbf{w})}\mathrm{R}_{\{\mathbf{u}1,\ldots,\mathbf{u}j\}}$ and ${}_{\mathbf{T}}\mathrm{R}_{\{\mathbf{u}1,\ldots,\mathbf{u}j,\mathbf{v}(w_0,\mathbf{w})\}}$ are orthogonal to each other. Namely, $\|{}_{\mathbf{T}}\mathrm{R}_{\{\mathbf{u}1,\ldots,\mathbf{u}j\}}\|^2 \ge \|{}_{\mathbf{T}}\mathrm{R}_{\{\mathbf{u}1,\ldots,\mathbf{u}j,\mathbf{v}(w_0,\mathbf{w})\}}\|^2$. Therefore, the optimization of $\min_{w_0,\mathbf{w}} \|{}_{\mathbf{T}}\mathrm{R}_{\{\mathbf{u}1,\ldots,\mathbf{u}j,\mathbf{v}(w_0,\mathbf{w})\}}\|^2$ leads to a non-zero ${}_{\mathbf{v}(w_0^*,\mathbf{w}^*)}\mathrm{R}_{\{\mathbf{u}1,\ldots,\mathbf{u}j\}}$, in which $(w_0^*, \mathbf{w}^*) \equiv \arg(\min_{w_0,\mathbf{w}} \|{}_{\mathbf{T}}\mathrm{R}_{\{\mathbf{u}1,\ldots,\mathbf{u}j,\,\mathbf{v}(w_0,\mathbf{w})\}}\|^2)$. Thus $\mathbf{u}^{j+1} \equiv \frac{{}_{\mathbf{v}(w_0^*,\mathbf{w}^*)}\mathrm{R}_{\{\mathbf{u}^1,\ldots,\mathbf{u}^j\}}}{\|_{\mathbf{v}(w_0^*,\mathbf{w}^*)}\mathrm{R}_{\{\mathbf{u}^1,\ldots,\mathbf{u}^j\}}\|}$ is a unit vector and, from **Lemma 1** (ii), orthogonal to the set of vectors $\{\mathbf{u}^1, \ldots, \mathbf{u}^j\}$. So, all vectors in the generated set $\{\mathbf{u}^1, \ldots, \mathbf{u}^p\}$ at Step 8 are orthonormal.

From Steps 6 and 7 of Table 1, if ${}_{\mathbf{T}}\mathrm{R}_{\{\mathbf{u}^1,\ldots,\mathbf{u}^p\}} \neq \mathbf{0}$, $\mathbf{u}^0 = \frac{{}_{\mathbf{I}}\mathrm{R}_{\{\mathbf{u}^1,\ldots,\mathbf{u}^p\}}}{\| {}_{\mathbf{I}}\mathrm{R}_{\{\mathbf{u}^1,\ldots,\mathbf{u}^p\}} \|}$ and ${}_{\mathbf{T}}\mathrm{R}_{\{\mathbf{u}^1,\ldots,\mathbf{u}^p,\mathbf{u}^0\}} = 0$, then $\mathbf{u}^0$ is non-zero and, from **Lemma 1** (ii), orthogonal to all vectors in the set $\{\mathbf{u}^1, \ldots, \mathbf{u}^p\}$. Namely, $\{\mathbf{u}^1, \ldots, \mathbf{u}^p, \mathbf{u}^0\}$ is an orthonormal set.

Q.E.D.

If $\mathbf{T} = \sum_{j=1}^{p} w_j^3 \mathbf{v}(w_{j0}^2, \mathbf{w}_j^2)$, then $\mathbf{T}^{\mathrm{T}} \mathbf{u}^k = (\sum_{j=1}^{p} w_j^3 \mathbf{v}(w_{j0}^2, \mathbf{w}_j^2))^{\mathrm{T}} \mathbf{u}^k = \sum_{j=k}^{p} w_j^3 \mathbf{v}(w_{j0}^2, \mathbf{w}_j^2)^{\mathrm{T}} \mathbf{u}^k$, since $\mathbf{v}(w_{j0}^2, \mathbf{w}_j^2)$ is in the subspace spanned by $\{\mathbf{u}^1, \ldots, \mathbf{u}^j\}$ for all $j = 1, \ldots, p$ and $(\mathbf{u}^i)^{\mathrm{T}} \mathbf{u}^j \neq 0 \ \forall \ i \neq j$. Thus, as stated in Step 5 of Table 1, $w_p^3 \equiv \frac{\mathbf{T}^{\mathrm{T}} \mathbf{u}^p}{\mathbf{v}(w_{p0}^2, \mathbf{w}_p^2)^{\mathrm{T}} \mathbf{u}^p}$, $w_j^3 \equiv \frac{(\mathbf{T} - \sum_{l=j+1}^{p} w_l^3 \mathbf{v}(w_{l0}^2, \mathbf{w}_l^2))^{\mathrm{T}} \mathbf{u}^j}{\mathbf{v}(w_{j0}^2, \mathbf{w}_j^2)^{\mathrm{T}} \mathbf{u}^j}$ for all $1 \leq j \leq p$-1, and $w_0^3 \equiv 0$. Similarly, if $\mathbf{T} = w_0^3 \mathbf{1} + \sum_{j=1}^{p} w_j^3 \mathbf{v}(w_{j0}^2, \mathbf{w}_j^2)$, then $\mathbf{T}^{\mathrm{T}} \mathbf{u}^k = (w_0^3 \mathbf{1} + \sum_{j=1}^{p} w_j^3 \mathbf{v}(w_{j0}^2, \mathbf{w}_j^2))^{\mathrm{T}} \mathbf{u}^k = (w_0^3 \mathbf{1} + \sum_{j=k}^{p} w_j^3 \mathbf{v}(w_{j0}^2, \mathbf{w}_j^2))^{\mathrm{T}} \mathbf{u}^k$. Therefore, as stated in Step 7 of Table 1, $w_0^3 \equiv \frac{\mathbf{T}^{\mathrm{T}} \mathbf{u}^0}{\mathbf{1}^{\mathrm{T}} \mathbf{u}^0}$, $w_p^3 \equiv \frac{(\mathbf{T} - w_0^3 \mathbf{1})^{\mathrm{T}} \mathbf{u}^p}{\mathbf{v}(w_{p0}^2, \mathbf{w}_p^2)^{\mathrm{T}} \mathbf{u}^p}$, and $w_j^3 \equiv \frac{(\mathbf{T} - w_0^3 \mathbf{1} - \sum_{l=j+1}^{p} w_l^3 \mathbf{v}(w_{l0}^2, \mathbf{w}_l^2))^{\mathrm{T}} \mathbf{u}^j}{\mathbf{v}(w_{j0}^2, \mathbf{w}_j^2)^{\mathrm{T}} \mathbf{u}^j}$ for all $1 \leq j \leq p$-1.

References

1. M. Me'zard and J. Nadal, *Learning in feedforward layered networks: The tiling algorithm. Journal of Physics.* **A 22**, 2191– 2204(1989).
2. S. Fahlman and C. Lebiere, *The Cascade-Correlation Learning Architecture. In Touretzky, D. (Eds.), Advances in Neural Information Processing Systems II (Denver, 1989) Morgan Kaufmann, San Mateo* (1990).
3. M. Frean, *The Upstart Algorithm: A Method for Constructing and Training Feedforward Neural Networks. Neural Computation*. **2**, 198– 209(1990).
4. G. Huang and H. Babri, *Upper bounds on the number of hidden neurons in feedforward networks with arbitrary bounded nonlinear activation functions. IEEE Transactions on Neural Networks*. **9**, 224– 229(1998).