1998 年情報論的学習理論ワークショップ
1998 Workshop on Information-Based Induction Sciences (IBIS'98)
Hakone, Japan, July 11-12, 1998.

多層ニューラルネットワークのバッチ学習における過学習の存在

Existence of Overtraining in Batch Learning of Multilayer Neural Networks

福水 健次*
Kenji Fukumizu

Abstract: This paper discusses batch gradient descent learning in multilayer networks with a large number of statistical training data. We emphasize on the difference between regular cases, where the prepared model has the same size as the true function, and overrealizable cases, where the model has surplus hidden units to realize the true function. First, experimental study on multilayer perceptrons and linear neural networks (LNN) shows that batch learning induces strong overtraining on both models in overrealizable cases. We theoretically analyze the dynamics in LNN, and show that this overtraining is caused by shrinkage of the parameters corresponding to surplus units.

1 はじめに

本論文は、多層ニューラルネットがどのような「多層固有の」性質を持つかを議論する。関数近似の観点からは、ある関数空間内では、固定の基底関数を持つ関数系より３層ニューラルネットのほうが高い近似精度を持つことが示されているが、統計的な観点からの本質的な違いは未だ不明な点が多い。部分的な結果として、多層モデルでは、正解がモデルより小さいサイズのネットで実現できる場合には、正解を与えるパラメータが高次元多様体をなすという構造的性質があり ([Fuk96])、通常の統計理論では説明できない現象が生じ得る。例えば、簡単な３層モデルで最尤推定量の汎化誤差が通常の場合より大きい例がある ([Fuk97])。これは、ある意味では多層構造の欠点であり、利点が示されているわけではない。

そこで多層ネットワークの統計モデルとしての利点を考察することが重要となる。本論文では「バッチ学習時における過学習」に焦点を当てて、特別な例を通じて多層モデルの利点を示す。ニューラルネットの学習では、誤差逆伝播法など学習誤差を最小にするようパラメータを逐次更新する手法が用いられるが、学習の本来の目的は理想的な正解と現在の予測との誤差（汎化誤差）であり、学習誤差最小化が汎化誤差を最小にするとは限らない。学習途上で汎化誤差がいったん最小値をとり、その後漸増していく（学習し過ぎ）ことを**過学習**と呼ぶ。

過学習の存在に関しては様々な議論があり、多くの実際的な応用において過学習が報告されている一方、Amari らはその効果が理論的には非常に小さいことを示している ([AMM96])。しかし Amari らの解析は通常の漸近理論に基づいており、正解が多様体をなす場合の過学習に関しては未だ議論されていない。本論文は、３層モデルの利点として、このような場合の過学習を論じる。

2 統計的な学習の枠組

一般に３層ニューラルネットとは、パラメータを持った $\mathbb{R}^L$ から $\mathbb{R}^M$ への関数の族 $\{\boldsymbol{f}(\boldsymbol{x};\boldsymbol{\theta})\}$

$$f_i(\boldsymbol{x};\boldsymbol{\theta}) = \sum_{j=1}^{H} w_{ij}\, s(\sum_{k=1}^{L} u_{jk}x_k + \zeta_j) + \eta_i \qquad (1)$$

として定義される。ここで、$\boldsymbol{\theta} = (w_{ij}, \eta_i, u_{jk}, \zeta_j)$ はパラメータ、H は中間素子の個数をあらわす。１変数関数 $s(t)$ は活性化関数と呼ばれ、多層パーセプトロン（MLP）ではシグモイド関数 $s(t) = \frac{1}{1+e^{-t}}$ が用いられる。

本論文では、ノイズを含んだ有限個のデータから真の関数を推定する、統計的推定の問題を考察する。真のシステムから与えられるデータは次式に従うと仮定する。

$$\boldsymbol{y} = \boldsymbol{f}(\boldsymbol{x}) + \boldsymbol{v}. \qquad (2)$$

ここで $\boldsymbol{f}(\boldsymbol{x})$ は正解の関数であり、観測ノイズ $\boldsymbol{v}$ は平均

*理化学研究所 脳科学総合研究センター〒 351-0198 埼玉県和光市広沢 2-1 tel. 048-467-9664, e-mail fuku@brain.riken.go.jp,
RIKEN Brain Science Institute, Hirosawa 2-1, Wako, Saitama 351-0198, Japan

0 共分散行列 $\sigma^2 I$（スカラー行列）の正規雑音、入力 $\boldsymbol{x}$ は確率 Q に従う。学習データ $\{(\boldsymbol{x}^{(\nu)}, \boldsymbol{y}^{(\nu)})\}_{\nu=1}^{N}$ はこの機構から独立に発生した標本である。本論文では、正解 $\boldsymbol{f}(\boldsymbol{x})$ はモデルによって実現可能であるとし、真のパラメータを $\boldsymbol{\theta}_0$ として $\boldsymbol{f}(\boldsymbol{x};\boldsymbol{\theta}_0) = \boldsymbol{f}(\boldsymbol{x})$ を仮定する。正解 $\boldsymbol{f}(\boldsymbol{x})$ が H 個より中間素子数の小さいネットワークで実現可能なときに、この正解を**過実現可能**と呼ぶ。

ネットワークの学習は、学習誤差

$$\mathrm{E}_{tr} = \sum_{\nu=1}^{N} \|\boldsymbol{y}^{(\nu)} - \boldsymbol{f}(\boldsymbol{x}^{(\nu)};\boldsymbol{\theta})\|^2 \tag{3}$$

を最小にするように行なう。この学習は、条件付確率のモデル $p(\boldsymbol{y}|\boldsymbol{x};\boldsymbol{\theta}) = \frac{1}{(2\pi\sigma^2)^{M/2}} \exp\left(-\frac{1}{2\sigma^2}\|\boldsymbol{y} - \boldsymbol{f}(\boldsymbol{x};\boldsymbol{\theta})\|^2\right)$ に関する最尤推定量（MLE）を求める学習に一致する。一般にモデルが線形でないと、MLE は解析的に求められず数値的手法が用いられる。本論文では最急降下法

$$\boldsymbol{\theta}(t+1) = \boldsymbol{\theta}(t) - \beta \frac{\partial \mathrm{E}_{tr}}{\partial \boldsymbol{\theta}} \tag{4}$$

を考察する。この最急降下法は特に、予め用意された学習データを一度に用いるので、**バッチ学習**と呼ばれ、常に新しいデータを用いるオンライン学習と区別される。

ネットワークの性能は次式の汎化誤差で評価される。

$$\mathrm{E}_{gen} = \int \|\boldsymbol{f}(\boldsymbol{x};\boldsymbol{\theta}) - \boldsymbol{f}(\boldsymbol{x})\|^2 dQ(\boldsymbol{x}). \tag{5}$$

学習誤差は汎化誤差を近似してはいるが一致してはいないため、学習誤差を減少させる最急降下学習が汎化誤差を単調に減少させるとは限らない。そこで、学習途上における汎化誤差の解析が非常に重要となる。

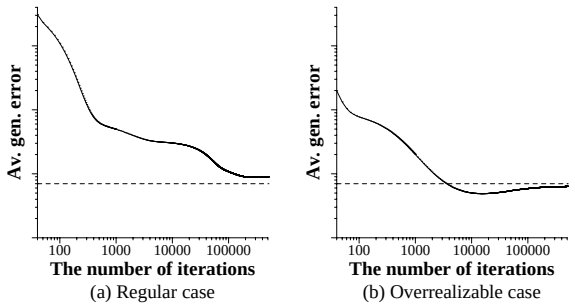
3 線形ニューラルネットワーク

MLP を用いた実験には局所解など様々な問題がある。特に過実現可能な正解を用いると、誤差曲面は MLE の周りに殆んど平坦な部分多様体を持つため ([Fuk96]) 学習速度は極端に遅く、停止条件を決めるのが難しい。これらは実験結果から有意義な結論を導くのを困難にする。

そこで、理論解析が可能な最も簡単な 3 層モデルとして、3 層線形ニューラルネット（LNN）を導入する。LNN は $H \times L$ 行列 A と $M \times H$ 行列 B に対し、

$$\boldsymbol{f}(\boldsymbol{x}; A, B) = BA\boldsymbol{x} \tag{6}$$

により定義される。以降 $H \leq L$, $H \leq M$ を仮定する。実現される写像は線形写像に過ぎないが、パラメータは 2 次の非線形性を持つ。このモデルは $\boldsymbol{f}(\boldsymbol{x}; C) = C\boldsymbol{x}$ という線形写像全体ではなく、ランクが H 以下に制限されているため、通常の線形モデルとは異なる。文献 [Fuk97] では、LNN で正解が過実現可能な場合に、MLE の汎化誤差が通常の理論とは異なることが示されている。



(a) Regular case (b) Overrealizable case

図 1: Learning curve of MLP. L = M = 1 and H = 2. N=100. The overrealizable target is $\boldsymbol{f}(\boldsymbol{x}) = 0$.

4 計算機実験 — 過学習の観測 —

ここでは MLP と LNN のバッチ学習時における汎化誤差の推移を実験的に調べる。

MLP の最急降下法は誤差逆伝播法を導く。3 章で議論した問題をなるべく避けるため、学習データ 1 セットに対し 30 種の初期値で学習を行ない、最終的に学習誤差を最小にしたものを正しい学習として採用した。図 1 に 30 セットの学習データに対する汎化誤差の平均を示す。正解が過実現可能な場合に顕著な過学習が見られる。

一方、LNN の MLE は解析的に解けることが知られており、最急降下法を用いる実際的な必要はない。しかし、ここでの興味はその動的な挙動であるため故意に最急降下学習を行なう。$X = (\boldsymbol{x}^{(1)}, \ldots, \boldsymbol{x}^{(N)})^T$, $Y = (\boldsymbol{y}^{(1)}, \ldots, \boldsymbol{y}^{(N)})^T$ とおくと、LNN の学習方程式は

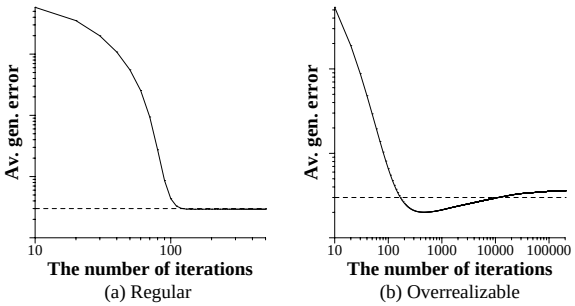
$$\begin{cases} A(t+1) &= A(t) + \beta B^T (Y^T X - BAX^T X) \\ B(t+1) &= B(t) + \beta (Y^T X - BAX^T X) A^T \end{cases} \tag{7}$$

で表される。図 2 に異なる 100 セットの学習データに対する汎化誤差の平均を示す。この場合も過実現可能な正解に対してのみ顕著な過学習が起こっている。

これらの結果から、一般に 3 層モデルでは通常の場合と正解が過実現可能な場合とで学習の挙動が本質的に違い、後者にのみ顕著な過学習が見られるという予想がたつ。もしこれが真実ならば、学習停止時間の最適化により、余分な中間素子の悪影響が改善されるという点で、3 層モデルは利点を持つことになる。

5 バッチ学習のダイナミクス

ここでは 過実現可能な正解に対して LNN が過学習を示すことを理論的に導く。



(a) Regular (b) Overrealizable

図 2: Learning curves of LNN. L=M=2 and H=1.

5.1 学習方程式の解

以降簡単のため以下の４つを仮定する。
(a) $H \leq L = M$, (b) $\boldsymbol{f}(\boldsymbol{x}) = B_0A_0\boldsymbol{x}$ で $\mathrm{rk}[B_0A_0] = r$, (c) $\mathrm{E}[\boldsymbol{x}\boldsymbol{x}^T] = \tau^2 I_L$, (d) $A(0)A(0)^T = B(0)^TB(0)$ かつこれらのランクが H.
本論文では解析を可能にするため、離散時間の更新則ではなく連続時間の微分方程式を考える。行列 Y をノイズ成分 V を用いて $Y = XA_0^TB_0^T + V$ と分解すると、最急降下法を表す微分方程式は次式で与えられる。

$$\begin{cases} \dot{A} = \beta B^T(B_0A_0X^TX + V^TX - BAX^TX) \\ \dot{B} = \beta(B_0A_0X^TX + V^TX - BAX^TX)A^T \end{cases} \tag{8}$$

$Z_O = \frac{1}{\sigma}V^TX(X^TX)^{-1/2}$ とおき、$X^TX = \tau^2NI_L + \tau^2\sqrt{N}Z_I$ と分解する。$Z = B_0A_0Z_I + \frac{\sigma}{\tau}Z_O$ とおき、

$$F = B_0A_0 + \varepsilon Z, \qquad \varepsilon = \frac{1}{\sqrt{N}} \tag{9}$$

と書くことにすると、N が十分大きいとき Z は定数オーダーを持ち ε よりも十分大きい。このとき (8) 式は

$$\begin{cases} \dot{A} &= \beta\tau^2NB^TF - \beta\tau^2NB^TBA \\ \dot{B} &= \beta\tau^2NFA^T - \beta\tau^2NBAA^T \end{cases} \tag{10}$$

により近似される。この近似は摂動と考えることもでき、$N \to \infty$ のときにはよい近似となる。
仮定 (d) と $\frac{d}{dt}(AA^T) = \frac{d}{dt}(B^TB)$ により、常に $AA^T = B^TB$ が満たされる。そこで

$$R = \begin{pmatrix} A^T \\ B \end{pmatrix} \tag{11}$$

という行列を導入すると、これは $S = \begin{pmatrix} 0 & F^T \\ F & 0 \end{pmatrix}$ に対し、

$$\frac{dR}{dt} = \beta\tau^2NSR - \frac{\beta\tau^2N}{2}RR^TR \tag{12}$$

を満足する。(12) 式は３次の非線形性を持つが、Oja の学習方程式 ([YHM94]) と同様にして解析できる。(12) 式より Riccati の行列微分方程式

$$\frac{d}{dt}(RR^T) = \beta\tau^2N\{SRR^T + RR^TS - (RR^T)^2\} \tag{13}$$

を得ることにより、次の定理が導かれる。

Theorem 1. $R(0)$ のランクが H のとき、*(13)* 式はすべての $t \geq 0$ に対し次式で表される唯一の解を持つ。

$$\begin{aligned} &R(t)R^T(t) = e^{\beta\tau^2NSt}R(0) \\ &\times \Big[I_H + \frac{1}{2}R(0)^T\{e^{\beta\tau^2NSt}S^{-1}e^{\beta\tau^2NSt} - S^{-1}\}R(0)\Big]^{-1} \\ &\times R(0)^Te^{\beta\tau^2NSt}. \end{aligned} \tag{14}$$

5.2 汎化誤差の挙動

過学習の証明に移る。仮定 (c) より $\mathrm{E}_{gen} = \mathrm{Tr}[(BA - B_0A_0)(BA - B_0A_0)^T]$ なので、$BA - B_0A_0$ の行列ノルムを考える。これに左右から異なる直交行列をかけても E_{gen} は不変なので、特異値分解により、はじめから

$$B_0A_0 = \begin{pmatrix} \Lambda^{(0)} & 0 \\ 0 & 0 \end{pmatrix}, \quad \Lambda^{(0)} = \mathrm{diag}(\lambda_1^{(0)}, \ldots, \lambda_r^{(0)}) \tag{15}$$

とする。正解が過実現可能なのは $r < H$ の場合である。B_0A_0 の特異値は定数オーダーであり ε よりも十分大きいと仮定する。F の特異値分解を

$$F = W\Lambda U^T, \quad \Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_L) \tag{16}$$

とすると、F はノイズを含むので、確率１で $\lambda_1 > \lambda_2 > \cdots > \lambda_L > 0$ と仮定してよい。よく知られているように、行列の成分の摂動は特異値に同じオーダーの摂動をもたらすので、対角行列 Λ を

$$\Lambda = \begin{pmatrix} \Lambda_1 & & 0 \\ & \varepsilon\tilde{\Lambda}_2 & \\ 0 & & \varepsilon\tilde{\Lambda}_3 \end{pmatrix} \tag{17}$$

と分解すると Λ_1, $\tilde{\Lambda}_2$, $\tilde{\Lambda}_3$ はそれぞれ r, $H-r$, $L-H$ 次元の定数オーダーの行列であり、

$$\begin{aligned} \lambda_i &= \lambda_i^{(0)} + O(\varepsilon), \qquad (1 \leq i \leq r), \\ \lambda_{r+j} &= \varepsilon\tilde{\lambda}_{r+j}, \qquad (1 \leq j \leq L-r) \end{aligned} \tag{18}$$

を満たす。以下では次の事実を示す。

> $r < H$ （正解が過実現可能）のとき、区間
>
> $$\frac{1}{\beta\tau^2\sqrt{N}(\tilde{\lambda}_H - \tilde{\lambda}_{H+1})} \ll t \ll \frac{\log\sqrt{N}}{\beta\tau^2\sqrt{N}(\tilde{\lambda}_H - \tilde{\lambda}_{H+1})} \tag{19}$$
>
> において $\mathrm{E}_{gen}(t) < \mathrm{E}_{gen}(\infty)$ が成立する。

区間の条件は $\varepsilon \ll \exp\{-\beta\tau^2\sqrt{N}(\tilde{\lambda}_H - \tilde{\lambda}_{H+1})t\} \ll 1$ と同値である。
(14) 式の $[\cdot]^{-1}$ 内の主要部は $e^{\beta\tau^2NSt}S^{-1}e^{\beta\tau^2NSt}$ なので、この解は $S^{-\frac{1}{2}}e^{\beta\tau^2NSt}R(0)$ の列ベクトルの張る H 次元空間への S の直交射影と解釈できる。この部分

空間は S の大きい H 個の固有値に対応する固有空間へ収束するが、この収束の様子をさらに調べる。

まず S の対角化は

$$S = \Phi \begin{pmatrix} \Lambda & 0 \\ 0 & -\Lambda \end{pmatrix} \Phi^T, \quad \Phi = \frac{1}{\sqrt{2}} \begin{pmatrix} U & U \\ W & -W \end{pmatrix} \tag{20}$$

で与えられる。また、$A(0)A(0)^T = B(0)^T B(0)$ より $R(0)$ の特異値分解は次のような形で書ける。

$$R(0) = \Theta J \Gamma G^T \tag{21}$$

($\Theta = \begin{pmatrix} P & 0 \\ 0 & Q \end{pmatrix}$, G は直交行列、$J = (I_H \ 0 \ I_H \ 0)^T$, Γ は特異値行列)。すると、直交射影の像空間は

$$\begin{aligned} & S^{-\frac{1}{2}} e^{\beta\tau^2 NSt} R(0) \\ &= \Phi \begin{pmatrix} \Lambda^{-\frac{1}{2}} e^{\beta\tau^2 N\Lambda t} & 0 \\ 0 & \sqrt{-1}\Lambda^{-\frac{1}{2}} e^{-\beta\tau^2 N\Lambda t} \end{pmatrix} \Phi^T R(0) \\ &= \Phi K \times \Lambda_H^{-\frac{1}{2}} e^{\beta\tau^2 N\Lambda_H t} C_H \Gamma G^T \end{aligned} \tag{22}$$

となる。ここで $C = \frac{1}{\sqrt{2}}(U^T P + W^T Q) = \begin{pmatrix} C_H & * \\ C_3 & * \end{pmatrix}$, $D = \frac{1}{\sqrt{2}}(U^T P - W^T Q) = \begin{pmatrix} D_H & * \\ D_3 & * \end{pmatrix}$, $\Lambda_H = \begin{pmatrix} \Lambda_1 & 0 \\ 0 & \Lambda_2 \end{pmatrix}$,

$$K = \begin{pmatrix} I_H \\ \Lambda_3^{-\frac{1}{2}} e^{\beta\tau^2\sqrt{N}\tilde{\Lambda}_3 t} C_3 C_H^{-1} e^{-\beta\tau^2 N\Lambda_H t} \Lambda_H^{\frac{1}{2}} \\ \sqrt{-1}\Lambda_H^{-\frac{1}{2}} e^{-\beta\tau^2 N\Lambda_H t} D_H C_H^{-1} e^{-\beta\tau^2 N\Lambda_H t} \Lambda_H^{\frac{1}{2}} \\ \sqrt{-1}\Lambda_3^{-\frac{1}{2}} e^{-\beta\tau^2\sqrt{N}\tilde{\Lambda}_3 t} D_3 C_H^{-1} e^{-\beta\tau^2 N\Lambda_H t} \Lambda_H^{\frac{1}{2}} \end{pmatrix}. \tag{23}$$

K の列ベクトルの張る部分空間への直交射影を $P_K = K(K^T K)^{-1} K^T$ と書くことにすると、解は

$$RR^T \sim 2\Phi \begin{pmatrix} \Lambda^{\frac{1}{2}} & 0 \\ 0 & \sqrt{-1}\Lambda^{\frac{1}{2}} \end{pmatrix} P_K \begin{pmatrix} \Lambda^{\frac{1}{2}} & 0 \\ 0 & \sqrt{-1}\Lambda^{\frac{1}{2}} \end{pmatrix} \Phi^T \tag{24}$$

となる。K は指数オーダーで $(I_H \, 0)^T$ に収束する。

行列 K の中で最も遅い収束は第2ブロックの後半 $H-r$ 列内のオーダー$\exp\{-\beta\tau^2\sqrt{N}(\tilde{\lambda}_{r+j} - \tilde{\lambda}_{H+k})t\}$ の項である。(19) 式は、このオーダーがε よりも十分大きいことを要求している。K の主要部だけで近似して

$$K \sim \begin{pmatrix} I_r & 0 \\ 0 & I_{H-r} \\ 0 & K_{22} \\ & 0 \end{pmatrix} \tag{25}$$

とすると、BA は次式で近似できる。

$$BA \sim W\Lambda^{\frac{1}{2}} \begin{pmatrix} I_r & 0 & 0 \\ 0 & I_{H-r} - K_{22}^T K_{22} & K_{22}^T \\ 0 & K_{22} & K_{22} K_{22}^T \end{pmatrix} \Lambda^{\frac{1}{2}} U^T. \tag{26}$$

$K_{22} = 0$ に対応する MLE（$\hat{B}\hat{A}$ と書く）と比べて、上の BA では第2、第3ブロックに一種の shrinkage が生じており、結果的にノイズの影響を押えている。実際、

$$B_0 A_0 = W\Lambda^{\frac{1}{2}} \left(\begin{pmatrix} I_r & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} + O(\varepsilon) \right) \Lambda^{\frac{1}{2}} U^T \tag{27}$$

であることを用いると、

$$\begin{aligned} BA - B_0 A_0 \sim W\Lambda^{\frac{1}{2}} \Bigg\{ & \begin{pmatrix} I_r & 0 & 0 \\ 0 & I_{H-r} - K_{22}^T K_{22} & K_{22}^T \\ 0 & K_{22} & K_{22} K_{22}^T \end{pmatrix} \\ & - \left(\begin{pmatrix} I_r & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} + O(\varepsilon) \right) \Bigg\} \Lambda^{\frac{1}{2}} U^T, \end{aligned} \tag{28}$$

$$\begin{aligned} & \hat{B}\hat{A} - B_0 A_0 \sim W\Lambda^{\frac{1}{2}} \\ & \left\{ \begin{pmatrix} I_r & 0 & 0 \\ 0 & I_{H-r} & 0 \\ 0 & 0 & 0 \end{pmatrix} - \left(\begin{pmatrix} I_r & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} + O(\varepsilon) \right) \right\} \Lambda^{\frac{1}{2}} U^T \end{aligned} \tag{29}$$

となり、ε が K_{22} に比べて無視できれば、$BA - B_0 A_0$ の行列ノルムは $\hat{B}\hat{A} - B_0 A_0$ のそれよりも小さい。従って汎化誤差も小さくなる。

学習の初期値が十分大きく ε が無視できる場合には、汎化誤差が減少関数であることは容易に示せるので、汎化誤差は、学習の当初は減少し続け、ある時間帯で最終状態よりも小さい値を取り、徐々に増加して収束に向かう。これは過学習の存在を証明している。

6 おわりに

本論文では、3層ネットワークのバッチ学習において、正解が過実現可能な場合に過学習が見られることを示した。ここでの理論解析は LNN という単純な場合に限られたが、実験から、この現象は3層モデル一般の性質である可能性がある。もしそうであれば、学習停止時間を最適化することにより、3層モデル固有の利点となり得る。他のモデルに関する研究が今後期待される。

参考文献

[Fuk96] Fukumizu, K. (1996). A regularity condition of the information matrix of a multilayer perceptron network. *Neural Networks.* Vol. 9. pp.871–879.

[Fuk97] Fukumizu, K. (1997). Special statistical properties of neural network learning. *Proc. NOLTA'97.* pp.747–750.

[AMM96] Amari, S., Murata, N., & Müller, K. (1996). Statistical theory of overtraining – is cross-validation asymptotically effective?. *Advances in NIPS 8.* pp.176–182.

[YHM94] Yan, W., Helmke, U. & Moore, J. (1994). Global analysis of Oja's flow for neural networks. *IEEE Trans. on NN.* Vol. 5. No. 5, pp.674–683.