

局所錐型モデルの漸近理論とそのニューラルネットへの応用

Asymptotic Theory of Locally Conic Models and its Applications to Neural Networks

福水健次*
 Kenji Fukumizu

Abstract: Multilayer neural networks have a problem of unidentifiability in its parameterization. If a network has surplus hidden units to realize a target function, the parameters to give the function consist of a high dimensional subset. Many of usual statistical views fail in such cases. This paper discusses the likelihood ratio of the maximum likelihood estimation in unidentifiable cases, using the framework of locally conic models. We derive a sufficient condition that the likelihood ratio has a larger order than usual $O_p(1)$. The exact order of the likelihood ratio of multilayer perceptrons is derived, and a new regularization scheme is proposed to overcome the strong overfitting.

Keywords: Neural network, Maximum likelihood estimation, Unidentifiability, Locally conic model, Regularization

1 Introduction

In a multilayer neural network model with H hidden units, if a function can be realized by a network with $H - 1$ hidden units, the parameter to give the function is not unique, but a high-dimensional continuous subset ([10], [3], [7]). This problem is known as *unidentifiability* of parameters, which is seen in many important statistical models such as mixture models and ARMA ([4]). For example, consider the following three-layer network with two hidden units:

$$\varphi(x; \theta) = \sum_{j=1}^2 b_j s(x; w_j) + d, \quad (1)$$

where $s(x; w)$ is a nonlinear function with a parameter vector w . This includes multilayer perceptrons and RBF. Suppose the true input-output relation $\varphi_0(x)$ can be given by a network with 1 hidden unit, that is, $\varphi_0(x) = b_0 s(x; w_0) + d_0$. We can easily see that the set of true parameters to give $\varphi_0(x)$ includes high-dimensional submanifolds $\{b_2 = 0, b_1 = b_0, w_1 = w_0, d = d_0, w_2 : \text{free}\}$ and $\{b_1 + b_2 = b_0, w_1 = w_2 = w_0, d = d_0\}$ in the parameter space (Fig.1). This is quite different from a model with linear parameterization like polynomial regression, in which the parameter is always determined uniquely even if the size of the model is surplus to give a target function.

Unidentifiability influences strongly on the statistical behaviors and learning dynamics of multilayer neural networks. Many statistical techniques, including

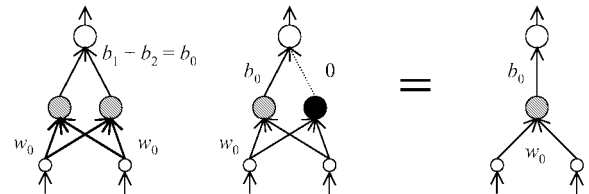


Figure 1: Unidentifiability in multilayer networks.

model selection criteria (AIC and MDL), which assume the uniqueness of the optimum parameter, are not directly applicable, if the true parameter is unidentifiable.

This paper discusses the likelihood ratio of the maximum likelihood estimator (MLE) in unidentifiable cases, using the framework of a locally conic model ([4]), and applied the results to multilayer neural networks. In particular, we focus on the asymptotic order of the likelihood ratio in unidentifiable cases of multilayer neural networks, and show that the order is larger than the usual constant order, which means strong overfitting with given data. Based on the locally conic parameterization, we will introduce a new regularization scheme for multilayer perceptrons to overcome such strong overfitting.

*統計数理研究所, 〒 106-8569 港区南麻布 4-6-7 tel. 03-5421-8730, e-mail fukumizu@ism.ac.jp,
 The Institute of Statistical Mathematics, 4-6-7 Minami-Azabu, Minato-ku, Tokyo 106-8569, Japan

2 Unidentifiability and Locally Conic Models

The general theory of a locally conic model is explained in this section. Let $\{p(z; \theta) \mid \theta \in \Theta\}$ be a family of probability densities. A parameter $\theta_0 \in \Theta$ is called *unidentifiable* if there exists a submanifold $\Theta_0 \subset \Theta$ such that θ_0 is included in Θ_0 , $\dim \Theta_0 \geq 1$, and $p(z; \theta) = p(z; \theta_0)$ for all $\theta \in \Theta_0$.

In the case of a function $\varphi(x; \theta)$ from x to y , by introducing a noise model $r(y|s)$ and an input density $q(x)$, we can regard it as a family of densities:

$$p(x, y; \theta) = r(y|\varphi(x; \theta))q(x). \quad (2)$$

The Gaussian distribution $\frac{1}{\sqrt{2\pi}\sigma} \exp\{-\frac{1}{2\sigma^2}(y-s)^2\}$ and cross-entropy $\frac{e^{ys}}{1+e^s}$ for $y \in \{0, 1\}$ are popular choice for a noise model. As the example in Section 1 illustrates, in three-layer neural networks, if a parameter defines a function which can be realized by a network with $H-1$ hidden units, the parameter is unidentifiable.

If the true probability p_0 , which generates i.i.d. training data, is given by an unidentifiable parameter in a model, the statistical analysis of estimation is difficult. We cannot use the asymptotic theory or Gaussian approximation for the distribution of the estimator. Different approaches are needed in such cases ([8],[11]).

Locally conic models have been introduced by Dacunha-Castelle and Gassiat ([4]) to discuss the unidentifiability. We use their formulation with some modifications. Let $S = \{p(z; \theta) \mid \theta \in \Theta\}$ be a family of probability density functions. We assume that the parameter space Θ is an open subset of $A_0 \times \mathbb{R}$, where A_0 is a $(d-1)$ dimensional space, and write $\theta = (\alpha, \beta)$ according to this decomposition. The model S is called *locally conic* at $p_0 \in S$ if the following four conditions are satisfied;

1. $p(z; (\alpha, \beta))$ is differentiable with respect to β for p_0 -almost every z .
2. The parameter space Θ contains $\Theta_0 := A_0 \times \{0\}$.
3. The set of parameters to give p_0 is Θ_0 , that is

$$p(z; (\alpha, \beta)) = p_0(z) \iff \beta = 0.$$

4. For all $\alpha \in A_0$, the Fisher information of the one-dimensional submodel $S_\alpha = \{p(z; \alpha, \beta) \mid (\alpha, \beta) \in \Theta\}$ at $\beta = 0$ is one; that is, $\|\frac{\partial \log p(z; \alpha, 0)}{\partial \beta}\|_{L^2(p_0)} = 1$.

If $\dim A_0 \geq 1$, the parameter to give p_0 is unidentifiable. In the set of true parameter Θ_0 , α can not be identified. Intuitively, the locally conic model is a d -dimensional subset in the space of all the probability density functions, while the point corresponding to p_0 is a conic singularity in the model (Fig.2). For

each $\alpha \in A_0$, the one-dimensional submodel S_α is an identifiable model, which gives p_0 only by $\beta = 0$. A locally conic model is a union of such one-dimensional submodels. The derivative of the log likelihood

$$v_\alpha(z) = \frac{\partial}{\partial \beta} \log p(z; (\alpha, 0)) \quad (3)$$

can be regarded as a tangent vector along S_α with unit $L^2(p_0)$ -norm. We call the set of such unit tangent vectors $C = \{v_\alpha \mid \alpha \in A_0\}$ *the basis of the tangent cone*, because C defines the tangent cone at the singularity.

3 Likelihood Ratio of a Locally Conic Model

Given an i.i.d. sample Z_1, \dots, Z_n following p_0 , the *likelihood ratio* is defined by

$$L_n(\theta) = \sum_{i=1}^n \log \frac{p(Z_i; \theta)}{p_0(Z_i)}. \quad (4)$$

The MLE $\hat{\theta}$ is the maximizer of $L_n(\theta)$. We focus on the likelihood ratio of MLE,

$$\sup_{\theta \in \Theta} L_n(\theta), \quad (5)$$

which essentially expresses the negative *training error* of a learning machine. In fact, for neural networks with the Gaussian noise model, the likelihood ratio is

$$L_n(\theta) = -\frac{1}{2\sigma^2} \sum_{i=1}^n (y_i - \varphi(x_i; \theta))^2 + \frac{1}{2\sigma^2} (y_i - \varphi_0(x_i))^2. \quad (6)$$

This measures *overfitting* of a learning machine. The more a machine fits with given data, the larger the likelihood ratio is. The likelihood ratio is also used in a statistical test, since, under the regularity conditions of asymptotic theory, we have

$$2L_n(\hat{\theta}) \longrightarrow \chi_d^2 \quad (n \rightarrow \infty) \quad \text{in law}, \quad (7)$$

where χ_d^2 is the chi-square distribution with freedom d . However, if the true parameter is unidentifiable, the asymptotic distribution of likelihood ratio is not chi-square, or may not be even $O_p(1)$ as we see later.

Locally conic parameterization is useful for the analysis of the likelihood ratio in unidentifiable cases. Let $\hat{\beta}_\alpha$ be MLE in the submodel S_α . As S_α is identifiable, a standard argument using Taylor expansion leads

$$L_n(\alpha, \hat{\beta}_\alpha) = \frac{1}{2} U_n(\alpha)^2 + o_p(1) \quad (8)$$

for each fixed α , where $U_n(\alpha)$ is defined by

$$U_n(\alpha) = \frac{1}{\sqrt{n}} \sum_{i=1}^n \frac{\partial \log p(Z_i; \alpha, 0)}{\partial \beta} = \frac{1}{\sqrt{n}} \sum_{i=1}^n v_\alpha(Z_i). \quad (9)$$

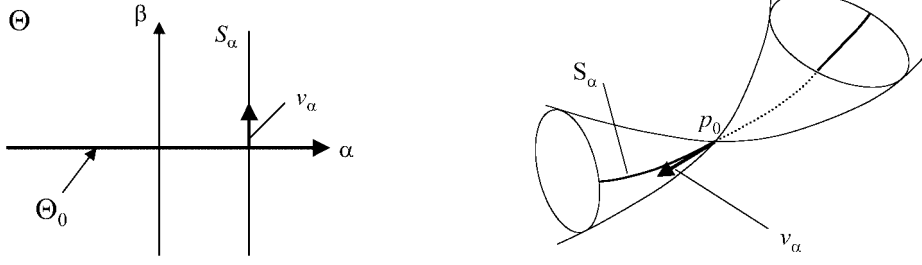


Figure 2: Locally conic model: parameter space (left) and the model in the space of density functions (right).

Note that the Fisher information of S_α at $\beta = 0$ is equal to one by the definition. The likelihood ratio of a locally conic model is, then, given by

$$\sup_{\theta \in \Theta} L_n(\theta) = \sup_{\alpha \in A_0} L_n(\alpha, \hat{\beta}_\alpha) = \sup_{\alpha \in A_0} \left(\frac{1}{2} U_n(\alpha)^2 + o_p(1) \right). \quad (10)$$

The MLE $\hat{\alpha}$, the maximizer of eq.(10), does not necessarily converge to a point in A_0 , but it distributes along Θ_0 .

The random variable $U_n(\alpha)$ converges in law to the standard normal distribution $N(0, 1)$ for each α . Considering all α , the random element U_n is a random process over α or the basis of the tangent cone. While the process marginally converges to $N(0, 1)$ for each α , it does not necessarily converge as a random process. Indeed, the process may not converge uniformly, and the likelihood ratio can diverge to infinity. Hartzigan ([9]) shows such divergence in a special case of the normal mixture model with two components. His argument is as follows. The marginal distribution of U_n over finite points $v_{\alpha_1}, \dots, v_{\alpha_m}$ in the basis of the tangent cone C converges to an m dimensional normal distribution with the covariance $E_{p_0}[v_{\alpha_i} v_{\alpha_j}]$. If we can find m "almost uncorrelated" elements in C for arbitrary $m \in \mathbb{N}$, the maximum of the $U_n(\alpha_j)$ ($1 \leq j \leq m$) is very close to the maximum of m i.i.d. samples from $N(0, 1)$, which is $\sqrt{2 \log m}$ for large m . As m is arbitrary, this maximum is not bounded. As an generalization of this fact we have the following theorem.

Theorem 1. *Let $S = \{p(z; (\alpha, \beta))\}$ be locally conic at $p_0 \in S$, and C be its basis of the tangent cone. Assume for each $\alpha \in A_0$ the submodel $S_\alpha = \{p(z; \alpha, \beta) \mid \beta\}$ satisfies the regularity conditions of asymptotic normality. If there is a sequence in C , which converges to zero in p_0 -probability, then, for arbitrary $M > 0$ we have*

$$\lim_{n \rightarrow \infty} \text{Prob} \left(\sup_{(\alpha, \beta)} L_n(\alpha, \beta) \leq M \right) = 0. \quad (11)$$

Proof. See [6]. \square

Eq.(11) shows that $L_n(\hat{\theta})$ is larger than the constant order $O_p(1)$. The above theorem gives a simple sufficient condition of divergence of the likelihood ratio.

4 Strong Overfitting of Multilayer Neural Networks

We consider the three-layer perceptron model with H hidden units:

$$\varphi(x; \theta) = \sum_{j=1}^H b_j s(a_j^T x + c_j) + d, \quad (12)$$

where $s(t) = \tanh t$, and the parameter space is denoted by Θ_H . We assume the output is one-dimensional for simplicity. Suppose that the true function is realized by a network with K hidden units

$$\varphi_0(x) = \sum_{k=1}^K b_k^0 s(a_k^0 x + c_k^0) + d^0, \quad (13)$$

for $0 \leq K \leq H$. If $K \leq H - 1$, the true parameter is unidentifiable, as we see in Section 1.

We can introduce a locally conic parameterization to formulate this unidentifiability. A slightly restricted parameter space Θ_H^* is defined by $\Theta_H^* = \{\theta \in \Theta_H \mid a_j \neq 0, b_j \neq 0 (1 \leq j \leq H), (a_j, c_j) \neq \pm(a_h, c_h) (1 \leq j < h \leq H), (a_j, c_j) \neq \pm(a_k^0, c_k^0) (1 \leq k \leq K, K + 1 \leq j \leq H)\}$. Note that Θ_H^* eliminates the parameters of functions realizable by a smaller-sized network (see [10]). This reduction does not matter in discussing MLE, since it lies in Θ_H^* with probability one. We introduce a new parameterization by

$$\begin{aligned} \beta &= \text{sgn}(b_{K+1}) \sqrt{b_{K+1}^2 + \dots + b_H^2}, & \delta &= \frac{d - d^0}{\beta}, \\ \xi_k &= \frac{a_k - a_k^0}{\beta}, \eta_k = \frac{b_k - b_k^0}{\beta}, & \zeta_k &= \frac{c_k - c_k^0}{\beta}, \\ \xi_j &= a_j, \eta_j = \frac{b_j}{\beta}, & \zeta_j &= c_j, \end{aligned} \quad (14)$$

for $1 \leq k \leq K$ and $K + 1 \leq j \leq H$. The three-layer perceptron is rewritten using this parameterization:

$$\begin{aligned} \psi(x; \omega) &= \sum_{k=1}^K (b_k^0 + \beta \eta_k) s((a_k^0 + \beta \xi_k) x + (c_k^0 + \beta \zeta_k)) \\ &+ \sum_{j=K+1}^H \beta \eta_j s(\xi_j x + \zeta_j) + \beta \delta. \end{aligned} \quad (15)$$

The new parameter spaces Π_H and Π_H^* are defined by $\Pi_H = \{\omega = (\xi_\ell, \eta_\ell, \zeta_\ell, \delta, \beta) \mid a_k^0 + \beta\xi_k \neq 0, b_k^0 + \beta\eta_k \neq 0, (a_k^0 + \beta\xi_k, c_k^0 + \beta\zeta_k) \neq \pm(a_h^0 + \beta\xi_h, c_h^0 + \beta\zeta_h), (a_k^0 + \beta\xi_k, c_k^0 + \beta\zeta_k) \neq \pm(\xi_j, \zeta_j), (\xi_j, \zeta_j) \neq \pm(a_k^0, c_k^0), \eta_j \neq 0, \xi_j \neq 0, (\xi_j, \zeta_j) \neq \pm(\xi_i, \zeta_i), (1 \leq k < h \leq K, K+1 \leq j < i \leq H), \sum_{j=K+1}^H \eta_j^2 = 1, \eta_{K+1} > 0, \beta \in \mathbb{R}\}$ and $\Pi_H^* = \{\omega \in \Pi_H \mid \beta \neq 0\}$, respectively. It is easy to see that $\varphi(x; \theta) = \psi(x; \omega)$ for corresponding $\theta \in \Theta_H^*$ and $\omega \in \Pi_H^*$, and that $\psi(x; \omega) = \varphi_0(x)$ if and only if $\beta = 0$. Thus, it suffices to consider $\{\psi(x; \omega) \mid \omega \in \Pi_H\}$, when MLE is discussed. We define the statistical model of three-layer perceptron $S_H = \{p(x, y; \omega) \mid \omega \in \Pi_H\}$ by

$$p(x, y; \omega) = r(y|\psi(x; \omega))q(x), \quad (16)$$

for some noise model $r(y|s)$ and input density $q(x)$. The model S_H consists of a probability density $p_0(x, y)$ corresponding to $\varphi_0(x)$ and densities given by $\varphi(x; \theta)$ for $\theta \in \Theta_H^*$. If we summarize $(\xi_1, \dots, \zeta_H, \delta)$ by α , $p(x, y; \alpha, \beta)$ gives a locally conic parameterization;

Theorem 2. *Under some regularity conditions on $r(y|s)$ and $q(x)$, the multilayer perceptron model S_H is locally conic at p_0 .*

Proof. It is easy to see that the conditions 1–3 are satisfied. For the condition 4, taking $N(\alpha) = \|\frac{\partial}{\partial \beta} \log p(x, y; (\alpha, 0))\|_{L^2(p_0)}$, the parameter $\tilde{\beta} = \frac{\beta}{N(\alpha)}$ instead of β makes the Fisher information one. \square

This locally conic model satisfies the assumptions of Theorem 1, and we have

Theorem 3. *Suppose that the model is given by eq.(12) and the true function by eq.(13). If $K \leq H - 1$, then, under some regularity conditions on $r(y|s)$ and $q(x)$, we have for arbitrary $M > 0$*

$$\lim_{n \rightarrow \infty} \text{Prob}(\sup_{\theta} L_n(\theta) \leq M) = 0. \quad (17)$$

Remark. This theorem means that the likelihood ratio has a larger order than the usual constant order $O_p(1)$, if a network has redundant hidden unit to realize the target. It indicates very strong overfitting.

Outline of the proof. We will prove the theorem only for $K \leq H - 2$, while it can be proved also for $K = H - 1$ (see [6]). We have only to consider a submodel $g(x, y; \xi, h, t) = r(y|\varphi_0(x) + \beta w(x; \xi, h, t))q(x)$, where $w(x; \xi, h, t)$ is in a function class $\mathcal{W} = \{w(x; \xi, h, t) = \frac{1}{\sqrt{A(\xi, h, t)}} \frac{1}{2} \{\tanh(\xi(x-t+h)) - \tanh(\xi(x-t-h))\}\}$. The constant $A(\xi, h, t)$ is a normalization of L^2 norm of the tangent vector $v(z; \xi, h, t)$ defined below. Note that the shape of the function w is bell-shaped (Fig.3). For this submodel, the basis of the tangent cone consists of the functions of the form:

$$v(z; \xi, h, t) = \frac{\partial \log r(y|\varphi_0(x))}{\partial s} w(x; \xi, h, t). \quad (18)$$

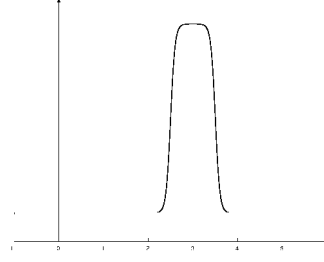


Figure 3: A function in the subclass \mathcal{W} ($\xi = 10, t = 3, h = 0.5$).

We can easily take a sequence (ξ_n, h_n, t_n) such that $\xi_n \rightarrow \infty, h_n \rightarrow 0$, and $v(z; \xi_n, h_n, t_n)$ converges to zero. \square

Using the subclass \mathcal{W} in the above proof, we can derive also a lower bound of the likelihood ratio, if $K \leq H - 2$;

Theorem 4. *Suppose that the model is given by eq.(12), and the true function by eq.(13). If $K \leq H - 2$, under some regularity conditions on $r(y|s)$ and $q(x)$, there exists $\delta > 0$ such that*

$$\liminf_{n \rightarrow \infty} \text{Prob}(\sup_{\theta} L_n(\theta) \geq \delta \log n) > 0. \quad (19)$$

Outline of the proof. We will show that we can find $m = n^\gamma$ ($\gamma > 0$) almost uncorrelated elements in the basis of the tangent cone C for the sample size n . For a closed interval $I \subset \mathbb{R}$, we define $M(I) = E_{p_0} [(\frac{\partial \log r(y|\varphi_0(x))}{\partial s})^2 \chi_I(x)]$, where $\chi_I(x)$ is the indicator function of I . For m disjoint intervals I_k ($1 \leq k \leq m$) in \mathbb{R} , we define one-dimensional models $r(y|\varphi_0(x) + \beta \frac{1}{\sqrt{M(I_k)}} \chi_{I_k}(x))q(x)$. The unit tangent vectors at the origin are $u_k(z) = \frac{1}{\sqrt{M(I_k)}} \frac{\partial \log r(y|\varphi_0(x))}{\partial u} \chi_{I_k}(x)$, which are uncorrelated. Then, under some regularity conditions, we can show that the distribution of the m -dimensional random vector $V_n = (\frac{1}{\sqrt{n}} \sum_{i=1}^n u_1(Z_i), \dots, \frac{1}{\sqrt{n}} \sum_{i=1}^n u_m(Z_i))$ can be approximated by the m -dimensional standard normal distribution for an appropriate choice of $\{I_k\}$ and a sufficiently small γ . The maximum of $|V_n|$ is arbitrarily close to $\sqrt{2 \log m} = \sqrt{2\gamma \log n}$. On the other hand, $\frac{1}{\sqrt{M(I)}} \chi_I(x)$ can be arbitrarily approximated by a function in \mathcal{W} , which means there exist n^γ functions in \mathcal{W} such that eq.(10) has the order of $\log n$. The rigorous proof needs delicate discussion. See Fukumizu ([6]) for the details. \square

The above theorem shows the order of the likelihood ratio is at least $\log n$, if the model has two redundant hidden units. This order is formerly obtained by Hagiwara et al. ([8]) under stronger assumptions of Gaussian noise and $\varphi_0(x) \equiv 0$.

We can derive an upper bound of the likelihood ratio for wider class of learning machines and the Gaussian noise model.

Theorem 5. Let $\mathcal{F} = \{\varphi(x; \theta)\}$ be a family of functions, and $\varphi_0(x)$ be a bounded function in \mathcal{F} . If the VC dimension of \mathcal{F} is finite, and the noise model is the Gaussian distribution $r(y|s) = \frac{1}{\sqrt{2\pi}\sigma} \exp\{-\frac{1}{2\sigma^2}(y-s)^2\}$, then, we have

$$\sup_{\theta} L_n(\theta) = O_p(\log n). \quad (20)$$

Outline of the proof. Since the probability of the event $\{\max_{1 \leq i \leq n} Y_i > 2\sqrt{\log n}\}$ converges to 0, we can consider the case in which $|Y_i - \varphi_0(X_i)| \leq 2\sqrt{\log n}$ and $|\varphi(x_i; \theta)| \leq 2\sqrt{\log n}$ hold. For a given $X = (X_1, \dots, X_n)$, the conditional probability of $L_n(\theta) - E_Y[L_n(\theta)|X] = \frac{1}{\sigma^2} \sum_{i=1}^n (Y_i - \varphi_0(X_i))(\varphi(X_i; \theta) - \varphi_0(X_i))$ is the normal distribution with mean zero and variance $V_{\theta, X} = \sum_{i=1}^n (\varphi(X_i; \theta) - \varphi_0(X_i))^2$. By the exponential inequality for the tail probability of a normal distribution, for arbitrary $\lambda > 0$ we have $\text{Prob}(L_n(\theta) \geq \lambda - \frac{1}{2\sigma^2} V_{\theta, X} | X) \leq \frac{\sqrt{V_{\theta, X}}}{\sqrt{2\pi}\lambda} e^{-\frac{\lambda^2}{2V_{\theta, X}}}$. Setting $\lambda = M \log n + \frac{1}{2\sigma^2} V_{\theta, X}$ for $M > 0$, we obtain

$$\text{Prob}(L_n(\theta) \geq M \log n) \leq \frac{\sigma}{\sqrt{2\pi}} \frac{1}{\sqrt{2M \log n}} n^{-M/\sigma^2}. \quad (21)$$

Since the VC dimension of \mathcal{F} is finite, we can find n^γ ($\gamma > 0$) parameters $\{\theta_k\}$ such that for arbitrary θ there exists θ_k satisfying

$$|L_n(\theta_k) - L_n(\theta)| \leq \sqrt{2 \log n}. \quad (22)$$

From eqs.(21) and (22), we obtain the theorem. \square

From theorems 4 and 5, we see that under the assumptions of Theorem 4 and of additive Gaussian noise, the order of $\sup_{\theta} L_n(\theta)$ is exactly $\log n$. In contrast to the order $O_p(1)$ in regular cases, a redundant neural network strongly overfits with the training data.

5 Regularization in Learning of Multilayer Perceptrons

The very strong overfitting shown in the previous section explains the heuristics that regularization is particularly important in neural networks. The proof of the previous theorems suggests that large parameter values in making a delta function can be a cause of the strong overfitting, which agrees with the Bartlett's statement ([2]) that large weight values worsen the generalization.

The conventional regularization terms like weight decay are not necessarily reasonable in multilayer neural networks. The weight decay, which adds the term

$$\lambda_n \frac{1}{2} \|\theta\|^2 \quad (23)$$

to the loss function, assumes that the ℓ_2 norm of θ represents the local distance of learning machines. This is not true about a locally conic model.

We propose the following regularization method in a locally conic model:

$$\begin{aligned} \tilde{L}_n(\alpha, \beta) &= L(\alpha, \beta) - \lambda_n \Phi(\alpha), \\ \Phi(\alpha) &= \frac{1}{2} \|\alpha - \alpha^0\|^2, \end{aligned} \quad (24)$$

where α^0 is a point in A_0 , and the regularization coefficient λ_n satisfies $\sup_{\theta} L_n(\theta) \ll \lambda_n \ll \sqrt{n}$. We can see that the maximizer $\tilde{\alpha}$ of \tilde{L}_n converges to α^0 in probability. In fact, if $\|\alpha - \alpha^0\| \geq \delta$ for some $\delta > 0$, the term $\lambda_n \Phi(\alpha)$ is asymptotically larger than $L_n(\alpha, \beta)$, and $\tilde{L}_n(\alpha, \beta)$ becomes negative. Such α cannot be the maximizer, because $\sup_{\beta} \tilde{L}_n(\alpha^0, \beta)$ is asymptotically positive. When we concentrate on a small compact neighborhood K of α_0 in maximizing $\tilde{L}_n(\alpha, \beta)$, we have

$$\begin{aligned} \sup_{\alpha, \beta} \tilde{L}_n(\alpha, \beta) &= \sup_{\alpha \in K} \frac{1}{2} \\ &\times \left\{ \frac{\left(\frac{1}{\sqrt{n}} \sum_{i=1}^n v_{\alpha}(X_i, Y_i) - \frac{\lambda_n}{\sqrt{n}} (\alpha - \alpha^0) \right)^2}{-\frac{1}{n} \sum_{i=1}^n \frac{\partial^2 \log p_0(Y_i|X_i; \alpha, 0)}{\partial \beta^2} + \frac{\lambda_n}{n}} \right\} + o_p(1). \end{aligned} \quad (25)$$

As the $o_p(1)$ term is uniform on the compact set K , from the fact $\lambda_n \ll \sqrt{n}$, the leading term of the right hand side is $\sup_{\alpha \in K} \frac{1}{2} \left(\frac{1}{\sqrt{n}} \sum_{i=1}^n v_{\alpha}(X_i, Y_i) \right)^2$, which is of the order $O_p(1)$. Therefore, the extent of overfitting is very improved so that the likelihood ratio does not diverge to infinity.

In three-layer perceptrons, the locally conic parameterization depends on the unknown true function. We utilize the parameterization for the constant-zero target $\varphi_0(x) \equiv 0$ to construct a regularization term. Taking $(\eta_1^0, \dots, \eta_H^0) = (\frac{1}{\sqrt{H}}, \dots, \frac{1}{\sqrt{H}})$, $\xi_j^0 = 1$, $\zeta_j^0 = 0$, and $\delta^0 = 0$ for α^0 , the regularization term of the three-layer perceptron is given by

$$\begin{aligned} \Phi(\alpha) &= \frac{1}{2} \left\{ -\frac{1}{H} \left(\sum_{j=1}^H \eta_j \right)^2 + \sum_{j=1}^H (\xi_j - 1)^2 + \sum_{j=1}^H \zeta_j^2 + \delta^2 \right\} \\ &= \frac{1}{2} \left\{ -\frac{(\sum_{j=1}^H b_j)^2}{\sum_{j=1}^H b_j^2} + \sum_{j=1}^H (a_j - 1)^2 + \sum_{j=1}^H c_j^2 + \frac{d^2}{\sum_j b_j^2} \right\}. \end{aligned} \quad (26)$$

From Theorem 4, the coefficient λ_n is chosen so that $\log n \ll \lambda_n \ll \sqrt{n}$. We use the cosine of the angle between η and η_0 in regularizing η , since η is restricted on the unit sphere. A clear difference from the weight decay is that $\Phi(\alpha)$ does not shrink a_j and b_j to zero, but suppress the large fluctuation of $\hat{\alpha}$. Also, it is known ([1]) that in regular cases the best generalization is attained by the constant order of λ_n for weight decay, while the order of the coefficient λ_n for $\Phi(\alpha)$ should not be smaller than $\log n$.

	Average	Std. Deviation
No regul.	7.06×10^{-4}	4.51×10^{-4}
$\Phi(\alpha)$	5.55×10^{-4}	3.41×10^{-4}
Weight decay	6.10×10^{-4}	3.73×10^{-4}

Table 1: Experimental results on regularization terms.

	Average	Std. Deviation
No regul.	5.33×10^{-4}	1.18×10^{-4}
$\Phi(\alpha)$	5.09×10^{-4}	1.05×10^{-4}
Weight decay	5.18×10^{-4}	1.07×10^{-4}

Table 2: Experimental results on regularization terms in the color conversion problem.

We made two experiments to see the effectiveness of this regularization method. One is an artificial problem, in which the true input-output relation is given by a three-layer perceptron with one hidden unit with additive Gaussian noise, and the model is a three-layer perceptron with four hidden units. The number of training data is 100. We evaluate the mean square error between the true function and trained networks. Table 1 shows the average and standard deviation over different 100 data sets with the same distribution. The second experiment is to learn a color conversion problem. Networks with 10 hidden units are trained to simulate a specific color reproduction system, in which the color ink is supplied by CMY (cyan, magenta, yellow), but the produced print is measured by the physical color system RGB (red, green, blue). The conversion table from RGB to CMY is needed to produce a desired color print ([5]). Instead of measuring a real reproduction system, we prepare 300 data per a simulation by inverting the theoretical Neugebauer equation from CMY to RGB using numerical optimization. Gaussian noise with variance 0.25×10^{-2} is added to the CMY output as observation noise. Table 2 shows the results over 50 simulations. In both experiments, the coefficients λ_n are decided by preliminary experiments. Both of the results show that the proposed regularization method improves the generalization.

6 Discussion

We have analyzed unidentifiable cases, assuming the true parameter is in the special location. In the case of multilayer networks, this means that the true function is completely realized by a smaller-sized network. Although one might think this assumption unnatural, we need such analysis to consider the model selection problem. Also, in real world applications of neural networks, a large number of hidden units are often used, and it is realistic to think there are many almost re-

dundant hidden units in the model. The analysis which assumes the uniqueness of the best parameters does not give a good insight in such situations.

The large order of likelihood ratio is not special to multilayer perceptrons. In many models, such as RBF, normal mixture models, and ARMA, the divergence of the likelihood ratio can be seen. Theorem 1 explains the reason of such divergence in a unified viewpoint.

The analysis of generalization error will be also very important, while we have discussed only training error in this paper. The theoretical analysis of generalization is very difficult in unidentifiable cases, because the estimator $\hat{\beta}_\alpha$ in a locally conic model does not converge uniformly over α . A new approach would be needed to discuss generalization.

References

- [1] S. Amari and N. Murata. Statistical analysis of regularization constant - from Bayes, MDL and NIC points of view. In *International Work-Conference on Artificial and Natural Neural Networks*, 1997.
- [2] P. L. Bartlett. For valid generalization, the size of the weights is more important than the size of the network. In *Advances in Neural Information Processing Systems 9*, pages 134–140. MIT Press, 1997.
- [3] A. M. Chen, H. Lu, and R. Hecht-Nielsen. On the geometry of feedforward neural network error surfaces. *Neural Computation*, 5:910–927, 1993.
- [4] D. Dacunha-Castelle and E. Gassiat. Testing in locally conic models and application to mixture models. *ESAIM Probability and Statistics*, 1:285–317, 1997.
- [5] K. Fukumizu. Statistical active learning in multilayer perceptrons. *IEEE Transactions on Neural Networks*, 11(1):17–26, 2000.
- [6] K. Fukumizu. Likelihood ratio of unidentifiable models and multilayer neural networks. Research Memorandum 780, The Institute of Statistical Mathematics, 2001.
- [7] K. Fukumizu and S. Amari. Local minima and plateaus in hierarchical structures of multilayer perceptrons. *Neural Networks*, 13(3):317–327, 2000.
- [8] K. Hagiwara, K. Kuno, and S. Usui. On the problem in model selection of neural network regression in overrealizable scenario. In *Proc. of Intern. Joint Conf. on Neural Networks*, volume VI, pages 461–466, 2000.
- [9] J. A. Hartigan. A failure of likelihood asymptotics for normal mixtures. In *Proceedings of Berkeley Conference in Honor of Jerzy Neyman and Jack Kiefer*, pages 807–810, 1985.
- [10] H. J. Sussmann. Uniqueness of the weights for minimal feedforward nets with a given input-output map. *Neural Networks*, 5:589–593, 1992.
- [11] S. Watanabe. Algebraic analysis for non-identifiable learning machines. *Neural Computation*, 13(4):899–933, 2001.