

Special Statistical Properties of Neural Network Learning

Kenji Fukumizu

Ricoh Co., Ltd., fuku@src.ricoh.co.jp

Abstract— We elucidate essential differences between feed-forward neural network models and conventional linear statistical models. When the target is overrealizable, the MLE of the former shows worse generalization, while experimental results reveals that iterative learning of a neural network shows eminent overtraining and better generalization in the middle.

I. INTRODUCTION

It is well-known that learning in feed-forward neural networks can be described as the parametric estimation from the viewpoint of statistics. It is an interesting problem whether there is any difference between neural network models and conventional linear models like the polynomials in their statistical properties.

We elucidate essential differences, focusing on the generalization error in overrealizable cases, where the target function is realized by a smaller-sized network than the model in use. The standard asymptotic theory tells that if we use the *maximum likelihood estimator* (MLE), the generalization error is proportional to the number of parameters. In overrealizable cases, however, the Fisher information matrix is singular, which makes the standard theory inapplicable.

This paper discusses the linear neural network (LNN) model. We give a rigorous calculation of the generalization error of a LNN in an overrealizable problem. Experimental results are shown on the generalization error in regular but almost overrealizable cases, which are often seen in practical problems. Moreover, we experimentally investigate the learning curve of steepest descent method and compare the results of overrealizable and regular cases.

II. STATISTICAL PRELIMINARIES

A. Linear neural networks (LNN)

A LNN has the identity function as its activation function. The i th output ($1 \leq i \leq M$) is given by

$$f_i(\mathbf{x}; A, B) = \sum_{j=1}^H B_{ij} \sum_{k=1}^L A_{jk} x_k. \quad (1)$$

The function $\mathbf{f}(\mathbf{x}; A, B)$ is a linear map from R^L to R^M . We assume $H \leq M \leq L$ throughout this paper.

An output of the target system is observed with a measurement noise. A pair of data (\mathbf{x}, \mathbf{y}) satisfies

$$\mathbf{y} = \mathbf{f}(\mathbf{x}) + Z, \quad (2)$$

where $\mathbf{f}(\mathbf{x})$ is the *target function* and Z is a random vector whose distribution is $N(0, \sigma^2 I_M)$, a normal distribution with 0 as its mean and $\sigma^2 I_M$ as its variance-covariance matrix. An input vector \mathbf{x} is generated randomly with its probability density function $q(\mathbf{x})$. Training data $\{(\mathbf{x}^{(\nu)}, \mathbf{y}^{(\nu)})\}_{\nu=1}^N$ are independent samples from the joint distribution of $q(\mathbf{x})$ and eq.(2).

We use the least square error estimator (LSEE):

$$(\hat{A}, \hat{B}) = \arg \min_{A, B} \sum_{\nu=1}^N \|\mathbf{y}^{(\nu)} - \mathbf{f}(\mathbf{x}^{(\nu)}; A, B)\|^2. \quad (3)$$

If we assume a conditional probability

$$p(\mathbf{y}|\mathbf{x}; A, B) = (2\pi\sigma^2)^{-M/2} e^{-\frac{1}{2\sigma^2} \|\mathbf{y} - \mathbf{f}(\mathbf{x}; A, B)\|^2}, \quad (4)$$

the LSEE is equal to the MLE, whose statistical behavior for a large number of training data is given by the statistical asymptotic theory ([1]).

We assume that $\mathbf{f}(\mathbf{x})$ is perfectly realized by the prepared model, and $\mathbf{f}(\mathbf{x}; A_0, B_0) = \mathbf{f}(\mathbf{x})$.

B. Generalization error – regular cases –

The generalization error of the MLE $\mathbf{f}(\mathbf{x}; \hat{\theta})$ is, in general, defined by

$$E_{gen} \equiv E_{X, Y} \left[\int \|\mathbf{f}(\mathbf{x}; \hat{\theta}) - \mathbf{f}(\mathbf{x})\|^2 q(\mathbf{x}) d\mathbf{x} \right]. \quad (5)$$

The statistical asymptotic theory shows

$$E_{gen} \approx \frac{\sigma^2}{N} \times S \quad (N \rightarrow \infty), \quad (6)$$

where S is the dimension of the parameter θ . Eq.(6) plays an essential role also in the derivation of Akaike information criterion (AIC).

The derivation of eq.(6) requires the regularity of the *Fisher information matrix* at the true parameter. The Fisher information matrix is given by

$$I_{ab}(\theta) = \frac{1}{\sigma^2} \int \frac{\partial \mathbf{f}^T(\mathbf{x}; \theta)}{\partial \theta_a} \frac{\partial \mathbf{f}(\mathbf{x}; \theta)}{\partial \theta_b} q(\mathbf{x}) d\mathbf{x}. \quad (7)$$

Eq.(6) is not applicable to a LNN, because of the singularity of its Fisher information matrix. A transform $A \mapsto CA, B \mapsto BC^{-1}$ does not change the function $\mathbf{f}(\mathbf{x}; A, B)$ for any regular matrix C . There is a subspace around each (A, B) such that the parameters in

it gives the same function as $f(\mathbf{x}; A, B)$. Therefore, the directional derivative tangent to the subspace is zero, which causes the singularity.

To avoid this redundancy, we consider the reduced model with the first H rows of B fixed as I_H . A LNN can always be converted to a network in the reduced model. It is easy to see that a Fisher information matrix of the reduced model is regular if and only if the linear map of the network is of full rank. The E_{gen} of a LNN of full rank is, then, given by

$$E_{gen} = \frac{\sigma^2}{N} H(L + M - H) + O(N^{-3/2}). \quad (8)$$

The target is overrealizable iff the rank of the map is less than H . Note that the generalization error in an overrealizable case is not necessarily given by eq.(8).

C. The importance of overrealizable cases

We discuss here the significance of analysis of overrealizable cases, in which the true parameters consist of the subvariety of dimension more than 0 ([2]). In addition to theoretical interest, the analysis of such cases is important also for practical reasons.

Consider the case where the target function is located very close to the subvariety of redundant networks. The generalization error for a sufficient number of training data is still expected to be given by eq.(8). However, the required number of data for eq.(8) might be extremely large in such cases. We make experiments to verify this, preparing a LNN with 2 input, 1 hidden, and 2 output units. The true function is

$$\mathbf{f}(\mathbf{x}; \boldsymbol{\theta}_0) = \begin{pmatrix} \varepsilon \\ 0 \end{pmatrix} (\varepsilon \ 0) \mathbf{x}, \quad (9)$$

where ε is a small positive number. This is not overrealizable unless $\varepsilon = 0$.

Figure 1 shows the averaged mean square errors for 100 simulations using 1000 training data. Since the effective number of parameters is only three, 1000 training data would be sufficient to apply eq.(8) in usual cases. The generalization errors for ε smaller than 0.1 show eminent increase from the theoretical prediction. The behavior of a network with small weight values cannot be necessarily described by the asymptotic theory even for a considerably large number of data, but is approximated by the error for $\varepsilon = 0$ which we will derive later. Since some of parameters are often very small in practical applications, we should consider practical situations as overrealizable cases.

III. GENERALIZATION ERROR IN OVERREALIZABLE CASES

A. General results

We analyze the simplest overrealizable case where the target function is constant zero. We use the notations:

$$X = (\mathbf{x}^{(1)} \dots \mathbf{x}^{(N)})^T, \quad Y = (\mathbf{y}^{(1)} \dots \mathbf{y}^{(N)})^T. \quad (10)$$

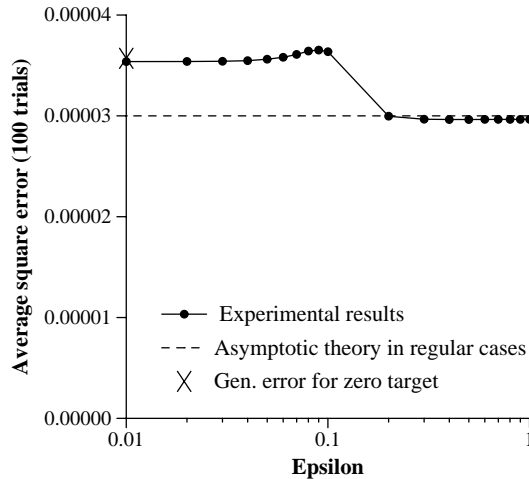


Figure 1: Almost overrealizable target

Proposition 1 ([3]). Let V_H be a $M \times H$ matrix whose i th column is the eigen vector corresponding to the i th largest eigenvalue of $Y^T X (X^T X)^{-1} X^T Y$. Then, the MLE of a linear neural network is given by

$$\hat{A} = V_H^T Y^T X (X^T X)^{-1}, \quad \hat{B} = V_H. \quad (11)$$

If the target function is constant zero, Y is independent of X . In this case, we have

Theorem 1. Assume $f(\mathbf{x}) = 0$. Let $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_H \geq 0$ be the eigen values of $Y^T X (X^T X)^{-1} X^T Y$. Then, the generalization error is given by

$$E_{gen} = \frac{1}{N} E[\lambda_1 + \dots + \lambda_H] + O(N^{-3/2}). \quad (12)$$

B. Networks with two output units

It is very difficult to calculate $E[\lambda_i]$ in general. If the output is two dimensional, $E[\lambda_1]$ can be obtained ([4]).

Theorem 2. Assume that $M = 2$. Then, the generalization error for the zero target function is

$$E_{gen} = \begin{cases} \frac{\sigma^2}{N} \left\{ L + \sqrt{\pi} \frac{\Gamma(\frac{L+1}{2})}{\Gamma(\frac{L}{2})} \right\} & \text{if } H = 1, \\ \frac{\sigma^2}{N} 2L & \text{if } H = 2. \end{cases} \quad (13)$$

Note that the effective number of parameters for a regular target function is $(L + 1)$ if $H = 1$. The generalization error in Theorem 2 is worse than the generalization error in regular cases. If the number of input units is very large, the Stirling's formula gives

$$E_{gen} \sim \frac{\sigma^2}{N} \left\{ L + \sqrt{\pi L/2} \right\}. \quad (14)$$

This is much worse than the error for a regular target.

We make computer simulations to confirm the theoretical results. The uniform distribution on $[-1, 1]^L$

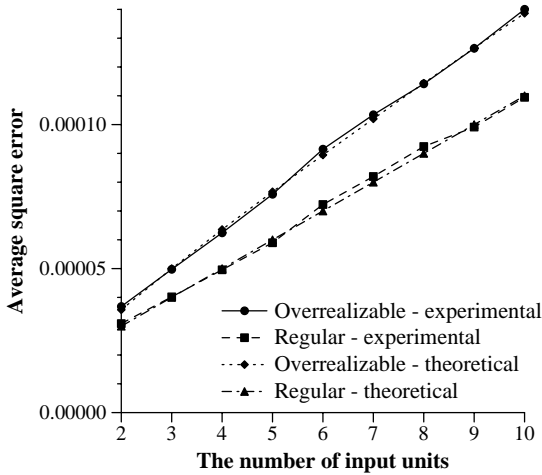


Figure 2: Generalization errors (2 output units)

is used for $q(\mathbf{x})$, and $\sigma = 0.1$. We prepare two target functions: one is the constant zero for an overrealizable case, and the other is a regular target defined by $A = (I_H \ O)$ and $B = (I_H \ O)^T$. The MLEs are calculated for 1000 samples. The average square error

$$\int \|\mathbf{f}(\mathbf{x}; \hat{A}, \hat{B}) - \mathbf{f}(\mathbf{x}, A_0, B_0)\|^2 q(\mathbf{x}) d\mathbf{x}$$

is obtained for each MLE. We make 1000 simulations, changing the random numbers. Figure 2 shows the average of experimental results and the theoretical predictions. We see that the experimental results coincides with the theoretical predictions very much.

C. Large scale networks

We analyze the generalization error of a large scale LNN by calculating the density function of λ_i under the assumption that L and M are very large and $M/L = \alpha$. We assume $\sigma = 1$ w.l.o.g., since the general result is obtained by multiplication of σ^2 . Let $W = (X^T X)^{-1/2} X^T Y$. All the elements of W are subject to $N(0, 1)$ and mutually independent.

Let $u_1 \geq u_2 \geq \dots \geq u_M \geq 0$ be the eigenvalues of $\frac{1}{L} W^T W$. We define the limiting density of u_i as

$$\rho(u) = \lim_{M \rightarrow \infty} \left\langle \frac{1}{M} \sum_{i=1}^M \delta(u - u_i) \right\rangle, \quad (15)$$

where $\langle \cdot \rangle$ is the expectation with respect to W . Using the replica method, we can calculate $\rho(u)$ as ([5])

$$\rho(u) = \frac{1}{2\pi\alpha} \frac{\sqrt{(u - u_m)(u_M - u)}}{u}, \quad (16)$$

where $u_m = (\sqrt{\alpha} - 1)^2$ and $u_M = (\sqrt{\alpha} + 1)^2$. The density takes a positive value only on (u_m, u_M) .

Let $\beta = H/M$. We define u_β as the β -percentile point of $\rho(u)$; that is,

$$\int_{u_\beta}^{u_M} \rho(u) du = \beta. \quad (17)$$

If we use the transform $t = \{u - (u_m + u_M)/2\}/(2\sqrt{\alpha})$,

$$\frac{2}{\pi} \int_{t_\beta}^1 \frac{\sqrt{1-t^2}}{2\sqrt{\alpha}t+1+\alpha} dt = \beta, \quad (18)$$

where t_β is β -percentile point. Then, we have

$$\begin{aligned} \sum_{h=1}^H \mathbb{E}[\lambda_h] &= LM \int_{u_\beta}^{u_M} u \rho(u) du \\ &= LM \left\{ \cos^{-1}(t_\beta) - t_\beta \sqrt{1-t_\beta^2} \right\} \end{aligned} \quad (19)$$

Therefore,

$$E_{gen} \approx \frac{\sigma^2}{N} LM \left(\cos^{-1}(t_\beta) - t_\beta \sqrt{1-t_\beta^2} \right). \quad (20)$$

On the other hand, the generalization error in regular cases is approximated by

$$E_{gen} \approx \frac{\sigma^2}{N} LM \beta (1 + \alpha - \alpha\beta). \quad (21)$$

Since elemental calculus gives

$$\cos^{-1}(t_\beta) - t_\beta \sqrt{1-t_\beta^2} \geq \beta(1 + \alpha - \alpha\beta), \quad (22)$$

we can conclude that the generalization error of the overrealizable case is always inferior to that of regular cases.

We make experiments to see the applicability of the theoretical results. We show here only the results of networks with 20 input and 10 output units. The above calculation requires infinitely many units theoretically, and moreover, the validity of the replica method has not been ensured yet. Figure 3 show the experimental results. The target functions in both cases are the same as the previous ones. We make 100 simulations, in each of which the MLE is calculated for 10000 training data. Even in the networks with smallest number of units, the theoretical results coincides with the experimental ones very much. We can conclude that the above approximation gives a good theoretical prediction of the actual generalization error.

IV. DYNAMICS OF BATCH LEARNING

In a general network model like the multilayer perceptron, the MLE cannot be calculated directly. We must employ an iterative method like the steepest descent. We experimentally investigate learning curves of the steepest descent method in overrealizable/regular cases.

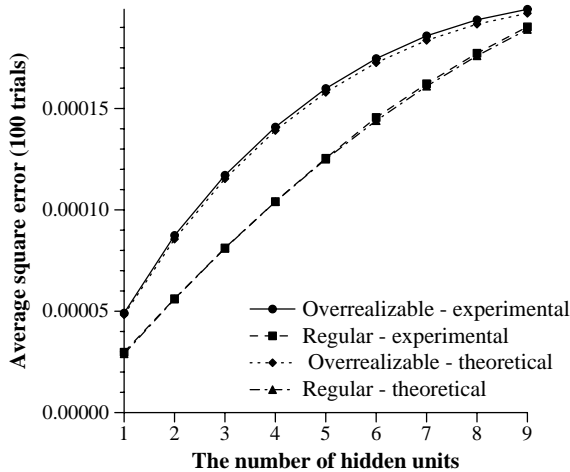


Figure 3: Comparison of generalization errors

The model and the target functions are the same as the experiments in III.B. We use a LNN with 1 hidden unit. We generate 1000 training data and train the network using the steepest descent method to minimize the sum of square errors. Figure 4 show the average generalization error during training for the overrealizable/regular target. The learning curve in the overrealizable case shows eminent overtraining in the middle of learning. It once becomes smaller than the expected generalization error for a regular target, increases gradually, and finally attains the error of the MLE. The learning curve in the regular case does not show such overtraining. This indicates that some early stopping criterion is effective in overrealizable cases, which corresponds with the practitioners' assertion that early stopping has remarkable effect of reducing the generalization error in many applications. It is reasonable to think that models in use often have almost redundant hidden units and the learning behavior is similar to our result of the overrealizable case.

V. CONCLUDING REMARKS

This paper elucidated essential differences between neural network models and usual linear models from the viewpoint of mathematical statistics. We presented an example in which the generalization error of the MLE changes according to the overrealizability of the target, and an experimental result that shows eminent overtraining only in the overrealizable case.

This paper is the first step to clarifying special statistical properties of multilayer network models. Although we discuss a very simple model of linear neural network, it is a very important and challenging problem to elucidate the static/dynamic behavior of network learning in more general situations.

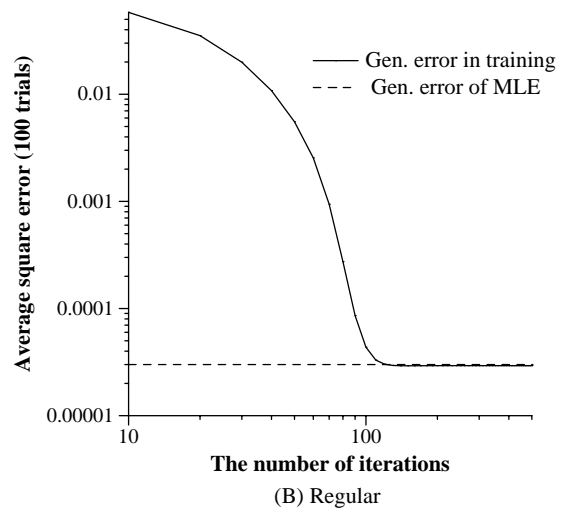
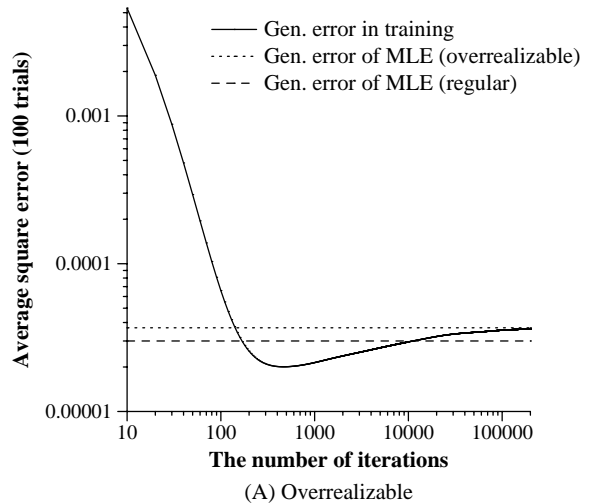


Figure 4: Learning curves on the generalization error

References

- [1] H. Cramér. *Mathematical method of statistics*, pages 497–506. Princeton University Press, 1946.
- [2] K. Fukumizu. A regularity condition of the information matrix of a multilayer perceptron network. *Neural Networks*, 9(5):871–879, 1996.
- [3] P. F. Baldi and K. Hornik. Learning in linear neural networks: a survey. *IEEE Trans. neural networks*, 6(4):837–858, 1995.
- [4] K. Fukumizu. Generalization error of a linear neural network with a singular Fisher information matrix. *IEICE technical report*, NC96-3, 1996.
- [5] M. Opper. Learning in neural networks : Solvable dynamics. *Europhys. Lett.*, 8(4):389–392, 1989.