# Over-fitting behavior of Gaussian unit under Gaussian noise

Katsuyuki Hagiwara
Faculty of Education, Mie University
1515 Kamihama, Tsu, 514-8507 Japan
E-mail: hagi@edu.mie-u.ac.jp

Kenji Fukumizu
Institute of Statistical Mathematics, ROIS
4-6-7 Minami-azabu, Minato-ku, Tokyo 106-8569 Japan
E-mail: fukumizu@ism.ac.jp

***Abstract*— In the training of neural networks and radial basis function networks under noisy environment, it is important to know how the network over-fits to the noise in the given data since it is directly related to the model selection and regularization problem. In this article, we firstly derive a probabilistic upper bound for the degree of over-fitting. By applying this result, we consider the over-fitting behavior of a Gaussian unit, which is trained under Gaussian noise, and we show that the probability that the width parameter of the Gaussian unit takes an extremely small value in training under Gaussian noise goes to one as the number of samples goes to infinity.**

## I. Introduction

In training neural networks and radial basis function networks under noisy environment, it is important to know how the network over-fits to noise in the given data since it is directly related to the model selection and regularization problem. This article mainly discusses the over-fitting property of neural network regression. For both of the model selection and the regularization problems, the well-known asymptotic theory is useful when the model satisfies some regularity conditions[15][16]. In a model selection criterion derived based on the standard asymptotic theory, the penalty term is given by the sum of the degree of over-fitting and the estimation error (see. e.g. [1][15]), where the latter represents the increase in the generalization error due to the over-fitting. Also in the case of applying the regularization method, the type of regularizer and a suitable value for the regularization parameter can be specified by knowing the over-fitting property (see e.g. [16]).

In general, the problem of over-fitting is serious when the number of nodes in a network is relatively large. Theoretically, this situation is considered to be the over-realizable case, in which the true function is realizable by a network with fewer hidden nodes than the assumed network. The over-realizable case is assumed in deriving a model selection criterion as in [1]. Unfortunately, it is known that the asymptotic expansion is not applicable for neural networks and radial basis function networks in over-realizable cases, thus, in considering their over-fitting properties[3][8][20]. This is caused by the lack of identifiability of connection weights, which breaks the regularity conditions assumed in the standard asymptotic theory. Under the over-realizable scenario, [7] and [10] gave a non-trivial upper bound of the training error, which is smaller than the theoretical value derived by the asymptotic expansion[15].

In this article, we discuss the over-fitting property of a Gaussian unit, which is trained under Gaussian noise. We first derive a non-trivial probabilistic upper bound of the degree of over-fitting under some restrictions on neural networks. In the statistical/computational learning theory, including the complexity regularization, distribution-free upper bounds for the estimation error are mainly discussed[4][19][12][5][13]. These nonparametric approaches for analyzing the estimation error has an advantage for neural networks because they do not suffer from the problem of the unidentifiability mentioned above. However, those bounds are conservative in general and too loose for our purpose. Here, by applying methods in the statistical/computational learning theory, we give a tight bound under the assumption of Gaussian noise. Based on this result, we analyze the over-fitting behavior of a Gaussian unit.

In the following, Section II gives a framework of regression estimation. In Section III, under some conditions on neural networks, we prove a key theorem that gives a probabilistic upper bound for the degree of over-fitting under Gaussian noise. The main theorem for a Gaussian unit is proved in Section IV. The conclusions with some discussions and future works are given in Section V.

## II. Framework of regression estimation

In this section, we first formulate the problem of regression estimation. While the discussion below might be regarded as that of a network with only one component, it is not restricted to the case of one component as we show in Section V. Notice that the following formulation is a special case of that given in [7] and [10].

Let $\{(X_i, Y_i) : 1 \leq i \leq n\}$ be i.i.d. pair of input-output samples according to a probability distribution on $\mathbf{R}^d \times \mathbf{R}$ and be used as training data. Throughout this article, let $G$ be a class of functions on $\mathbf{R}^d$, taking their values in $[-1, 1]$. We write $\mathbf{X} = (X_1, \ldots, X_n)$ and $\mathbf{Y} = (Y_1, \ldots, Y_n)$. Define $\mathbf{G}(\mathbf{X})$ by the restriction of $G$ to the set $\{X_1, \ldots, X_n\} \subseteq (\mathbf{R}^d)^n$, and write $\boldsymbol{g}(\mathbf{X}) = (g(X_1), \ldots, g(X_n))$, and $\mathbf{G}(\mathbf{X}) = \{\boldsymbol{g}(\mathbf{X}) : g \in G\}$. Consider the problem of fitting this training data by a function, which is defined by

$$f_{c,\boldsymbol{b}}(X) = c g_{\boldsymbol{b}}(X), \ c \in \mathbf{R}, \tag{1}$$

where $(c, \boldsymbol{b}) \in \mathbf{R} \times \mathbf{B}$ is the parameter vector. The function $g_{\boldsymbol{b}}$ is regarded as the output of one hidden unit in a layered

neural network or one radial basis function. Here, we assume that $|g_{\boldsymbol{b}}(X)| \leq 1$ for any $\boldsymbol{b} \in \mathbf{B}$ and $X \in \mathbf{R}^d$. By setting $G = \{g_{\boldsymbol{b}} : \boldsymbol{b} \in \mathbf{B}\}$ for (1), we can write $f_{c,\boldsymbol{b}} = f_{c,g}$ for $g \in G$. The empirical squared error of $f_{c,g}$ for the given training data is defined by

$$R_{\mathrm{emp}}(c, g) = \frac{1}{n}\|\mathbf{Y} - c\boldsymbol{g}(\mathbf{X})\|^2, \tag{2}$$

where $\boldsymbol{g}(\mathbf{X}) \in \mathbf{G}(\mathbf{X})$ and $\|\cdot\|$ denotes the Euclidean norm. Let $\widehat{c}(\boldsymbol{g}(\mathbf{X}))$ be an estimate of $c$ for a fixed $\boldsymbol{g}(\mathbf{X})$, which satisfies

$$R_{\mathrm{emp}}(\widehat{c}(\boldsymbol{g}(\mathbf{X})), \boldsymbol{g}(\mathbf{X})) = \min_{c \in \mathbf{R}} R_{\mathrm{emp}}(c, \boldsymbol{g}(\mathbf{X})). \tag{3}$$

Then, we have the minimization of $R_{\mathrm{emp}}(c, \boldsymbol{g}(\mathbf{X}))$ with respect to $(c, \boldsymbol{g}(\mathbf{X}))$ by

$$\inf_{\boldsymbol{g}(\mathbf{X}) \in \mathbf{G}(\mathbf{X})} R_{\mathrm{emp}}(\widehat{c}(\boldsymbol{g}(\mathbf{X})), \boldsymbol{g}(\mathbf{X})) := R_{\mathrm{emp}}(\widehat{c}, \widehat{\boldsymbol{g}}(\mathbf{X})). \tag{4}$$

Since $R_{\mathrm{emp}}(c, g(\mathbf{X}))$ is quadratic with respect to $c$, we obtain

$$\widehat{c}(g) = \frac{\langle \mathbf{Y}, \boldsymbol{g}(\mathbf{X}) \rangle}{\|\boldsymbol{g}(\mathbf{X})\|^2} \tag{5}$$

for each fixed $g$. Thus, it is easy to see

$$R_{\mathrm{emp}}(\widehat{c}(\boldsymbol{g}(\mathbf{X})), \boldsymbol{g}(\mathbf{X})) = \frac{1}{n}\|\mathbf{Y}\|^2 - \frac{1}{n} Z^2(\boldsymbol{g}(\mathbf{X})) \tag{6}$$

$$Z(\boldsymbol{g}(\mathbf{X})) = \widehat{c}(g)\|\boldsymbol{g}(\mathbf{X})\| = \frac{\langle \mathbf{Y}, \boldsymbol{g}(\mathbf{X}) \rangle}{\|\boldsymbol{g}(\mathbf{X})\|}. \tag{7}$$

Since the first term of the right hand side of (6) does not depend on $\boldsymbol{g}(\mathbf{X})$, we obtain

$$R_{\mathrm{emp}}(\widehat{c}, \widehat{g}) = \frac{1}{n}\|\mathbf{Y}\|^2 - \frac{1}{n} \sup_{\boldsymbol{g}(\mathbf{X}) \in \mathbf{G}(\mathbf{X})} Z^2(\boldsymbol{g}(\mathbf{X})). \tag{8}$$

From (8), we can see that the empirical squared error is small when $\sup_{\boldsymbol{g}(\mathbf{X})} Z^2(\boldsymbol{g}(\mathbf{X}))$ is large. Hence, $\sup_{\boldsymbol{g}(\mathbf{X})} Z^2(\boldsymbol{g}(\mathbf{X}))$ is viewed as a measure of the degree of fitting to the given data. To restrict our attention on the over-fitting property, we assume

$$Y_1, \ldots, Y_n \overset{\text{i.i.d.}}{\sim} N(0, 1), \tag{9}$$

which means the output data is an i.i.d. Gaussian noise sequence. Then, $\sup_{\boldsymbol{g}(\mathbf{X})} Z^2(\boldsymbol{g}(\mathbf{X}))$ represents the degree of fitting to noise, that is, in this case, the degree of the over-fitting under Gaussian noise.

## III. A PROBABILISTIC UPPER BOUND FOR $\sup Z^2(\boldsymbol{g}(\mathbf{X}))$

In this section, we give a probabilistic upper bound for the degree of over-fitting $\sup_{\boldsymbol{g}(\mathbf{X})} Z^2(\boldsymbol{g}(\mathbf{X}))$ under some conditions. Here, since both of $f_{c,\boldsymbol{b}}$ and $Z(\boldsymbol{g}(\mathbf{X}))$ are not uniformly bounded, we cannot directly apply the usual results in the computational learning theory[4][11]. Nevertheless, due to the limitation of the parametric treatment of neural networks, we apply some tools in the computational learning theory. In our case under the assumption of Gaussian noise, the problem of bounding $Z(\boldsymbol{g}(\mathbf{X}))$ is reduced to bounding the supremum of a Gaussian process on $g \in G$. Here, since the pseudo-dimension of $G$ is finite, this problem is reduced to bounding the maximum among Gaussian random variables on a finite $\epsilon$-cover of $G$ since $G$ is a class of bounded functions. Additionally, we can use the exponential bound for evaluating the upper probability of a Gaussian distribution, which is used as an alternative to the Hoeffding's inequality used commonly in the computational learning theory. This scenario is adopted by [9] under the same formulation in this section. However, a bound in [9] is somewhat weak for our purpose. To obtain a tight bound, here, we take account of the correlation structure of a Gaussian process induced by $G$.

Here, we consider the following restrictions on $g$.

*Assumption 1:* Under the fixed $\mathbf{X}$, there is a positive constant $\delta \in (0, 1)$ that satisfies $\|\boldsymbol{g}(\mathbf{X})\|^2/n \geq \delta$ for any $g \in G$.

*Assumption 2:* The pseudo-dimension of $G$ is finite and is $\gamma \geq 1$. Here, the pseudo-dimension is defined as the Vapnik-Chervonenkis(VC) dimension of the subgraphs of $G$ (see [4]).

Then, we have the following theorem under these assumptions.

*Theorem 1:* Let us assume (9);i.e. the output data is an i.i.d. Gaussian noise sequence. Under Assumption 1 and 2, if $n > 8\sqrt{2(4\gamma/\alpha + 1)/\delta}$ then

$$\begin{aligned} &\mathbf{P}\left( \sup_{\boldsymbol{g} \in \mathbf{G}} Z^2(\boldsymbol{g}(\mathbf{X})) > \alpha \log n + \beta \middle| \mathbf{X} \right) \\ \leq\ & C_1 n^{1/2} e^{-C_2 n} + \frac{C_3 n^{-\alpha}}{\sqrt{\log n}} + \frac{C_4 n^{-\alpha/8}}{\sqrt{\log n}} \end{aligned} \tag{10}$$

for any fixed $\alpha > 0$ and $\beta > 0$, where $C_m,\ m = 1, \ldots, 4$ are positive constants. □

We give the proof in Appendix.

Since the right hand side of (10) does not depend on $\mathbf{X}$ and goes to 0 as $n \to \infty$, the following corollary is immediately followed by Theorem 1.

*Corollary 1:* Let $G$ be a class of functions, taking their values in $[-1, 1]$. For the $G$, we assume that the pseudo-dimension is finite. Furthermore, we assume that the probability that $\|\boldsymbol{g}(\mathbf{X})\|^2 > \delta n$ holds for all $g \in G$ and for some $\delta \in (0, 1)$ is one under the assumed input probability distribution, where $\delta$ is a constant that does not depend on $n$. Then,

$$\mathbf{P}\left( \sup_{\boldsymbol{g}(\mathbf{X}) \in \mathbf{G}(\mathbf{X})} Z^2(\boldsymbol{g}(\mathbf{X})) > \alpha \log n + \beta \right) \to 0$$

as $n \to \infty$ holds for any $\alpha, \beta > 0$. □

As found in [1][2][15], the standard asymptotic theory says that the degree of over-fitting is bounded in probability since it converges to a $\chi^2$ distribution. This cannot be applied to our case because of the unidentifiability. Actually, for some types of neural networks including Gaussian radial basis function networks, [7] and [10] have shown that the probability that $\sup_{\boldsymbol{g}} Z^2(\boldsymbol{g})/\log n$ is larger than a constant goes to one as the number of samples goes to infinity. Since $g_{\boldsymbol{b}}$ in the above does not necessarily consist of one component as shown in the section V, our result says that the degree of over-fitting is suppressed by the restriction given in Assumption 1. In the context of the computational learning theory, [12] and [13] gave upper bounds for the estimation error for radial

basis function networks. Although these bounds are quite general, it is not known that a type of the restriction given in Assumption 1 controls the over-fitting property. In the following, we analyze the over-fitting behavior of a Gaussian unit by using this result.

## IV. Over-fitting behavior of Gaussian Unit

### A. Theoretical result

Let us assume that (9) holds for the output data and the input data $X_1, \ldots, X_n$ are i.i.d. according to a probability distribution defined on $[-K, K]^d$, where $K$ is a positive constant. We consider a Gaussian radial basis function, which is defined by

$$g_{\boldsymbol{b}}(X) = \exp\left\{-\|X - \boldsymbol{b}_1\|^2 / 2b_0\right\}, \tag{11}$$

$$G^{\mathrm{grbf}} = \left\{g_{\boldsymbol{b}} : \boldsymbol{b} = (b_0, \boldsymbol{b}_1) \in (0, \infty) \times [-M, M]^d\right\}, \tag{12}$$

where $M$ is a positive constant. For suitable fitting, we assume that $M > K$. Further, we define $G_\tau^{\mathrm{grbf}} = \{g \in G^{\mathrm{grbf}} : (b_0, \boldsymbol{b}_1) \in [\tau, \infty) \times [-M, M]^d\}$ for a positive constant $\tau$. Correspondingly, we define $\mathbf{G}^{\mathrm{grbf}}(\mathbf{X}) = \{\boldsymbol{g}(\mathbf{X}) : g \in G^{\mathrm{grbf}}\}$ and $\mathbf{G}_\tau^{\mathrm{grbf}}(\mathbf{X}) = \{\boldsymbol{g}(\mathbf{X}) : g \in G_\tau^{\mathrm{grbf}}\}$. We set $u_n := (2-\epsilon)\log n + \beta$, where $\epsilon \in (0, 2)$ and $\beta$ is a positive constant. Let us consider the probability $\mathbf{P}(\widehat{\boldsymbol{g}}(\mathbf{X}) \in \mathbf{G}_\tau^{\mathrm{grbf}}(\mathbf{X}))$. We define $\overline{\mathbf{G}}^{\mathrm{grbf}}(\mathbf{X}) = \mathbf{G}^{\mathrm{grbf}}(\mathbf{X}) - \mathbf{G}_\tau^{\mathrm{grbf}}(\mathbf{X})$. By (8), we obtain

$$\begin{aligned}
&\mathbf{P}\left(\widehat{\boldsymbol{g}}(\mathbf{X}) \in \mathbf{G}_\tau^{\mathrm{grbf}}(\mathbf{X})\right) \\
= \;& \mathbf{P}\left(\sup_{\boldsymbol{g} \in \mathbf{G}_\tau^{\mathrm{grbf}}(\mathbf{X})} Z^2(\boldsymbol{g}) > \sup_{\boldsymbol{g} \in \overline{\mathbf{G}}^{\mathrm{grbf}}(\mathbf{X})} Z^2(\boldsymbol{g})\right) \\
\le \;& \mathbf{P}\left(\sup_{\boldsymbol{g} \in \mathbf{G}_\tau^{\mathrm{grbf}}(\mathbf{X})} Z^2(\boldsymbol{g}) > u_n\right) \\
& + \mathbf{P}\left(\sup_{\boldsymbol{g} \in \overline{\mathbf{G}}^{\mathrm{grbf}}(\mathbf{X})} Z^2(\boldsymbol{g}) \le u_n\right) \\
:= \;& Q_1(n) + Q_2(n).
\end{aligned} \tag{13}$$

Let us first evaluate $Q_1(n)$. For all $g \in G_\tau^{\mathrm{grbf}}$, we have $|g(X)| \le 1$ for any $X \in \mathbf{R}^d$. For all $\boldsymbol{g}(\mathbf{X}) \in \mathbf{G}_\tau^{\mathrm{grbf}}(\mathbf{X})$, we obtain

$$\|\boldsymbol{g}(\mathbf{X})\|^2 \ge \delta(\tau) n \tag{14}$$

for any $\mathbf{X} \in [-K, K]^{dn}$, where we defined $\delta(\tau) := \exp\{-2d(K+M)^2/\tau\}$. Since $\delta(\tau) \in (0, 1)$ for $\tau \in (0, \infty)$, $\mathbf{G}_\tau^{\mathrm{grbf}}(\mathbf{X})$ satisfies Assumption 1 with $\delta = \delta(\tau)$ for any $\mathbf{X}$ and any probability distribution of $X_1$ on $[-K, K]^d$. Furthermore, since the pseudo-dimension of $G_\tau^{\mathrm{grbf}}$ is shown to be bounded above by $\gamma' = d^2 + d + 1$[12], Assumption 2 is satisfied. Thus, by Corollary 1, $Q_1(n)$ goes to 0 as $n \to \infty$. On the other hand, by following [7] and [10], it can be shown that

$$\limsup_{n \to \infty} Q_2(n) = 0 \tag{15}$$

holds here. This result is summarized into the following theorem.

*Theorem 2:* Under the above notations,

$$\limsup_{n \to \infty} \mathbf{P}\left(\widehat{\boldsymbol{g}}(\mathbf{X}) \in \mathbf{G}_\tau^{\mathrm{grbf}}(\mathbf{X})\right) = 0 \tag{16}$$

holds for any fixed $\tau > 0$, if the output data is an i.i.d. Gaussian noise sequence and the input data are i.i.d. samples from a probability distribution on $[-K, K]^d$ with $0 < K < M$. □

This theorem says that the probability to obtain the extremely small value for the width parameter in training under Gaussian noise goes to one as the number of samples goes to infinity. On the other hand, by (15), we obtain

$$\mathbf{P}\left(\sup_{\boldsymbol{g}(\mathbf{X}) \in \overline{\mathbf{G}}^{\mathrm{grbf}}(\mathbf{X})} Z^2(\boldsymbol{g}(\mathbf{X})) \le (2-\epsilon)\log n\right) \to 0$$

as $n \to \infty$ for any $\epsilon \in (0, 2)$. Hence, the degree of over-fitting of the Gaussian unit is larger than $2 \log n$, which is achieved by the fitting with extremely small value for the width parameter. In the context of computational learning theory, [21] gave a choice for the width parameter, by which the optimal convergence rate of the generalization error is obtained. However, in their analysis, the width parameter is fixed through the training and is allowed to vary with the sample size, while our article deals with a behavior of the estimate of the width parameter, which is obtained in training. Generally, it is empirically known that the output with high curvature is obtained in training neural networks with relatively large size. This is one reason for applying the regularization method. In the classification problem, [6] has given a theoretical support for this phenomenon based on an upper bound for the estimation error, in which the values of the connection weights are shown to be essential for the estimation error. However, his result does not give us a insight into the product of the training. Our result concerns with the degree of over-fitting and suggests us that the output with high curvature can be frequently chosen in training. Additionally, Theorem 1 together with Theorem 2 tell us that the restriction given in Assumption 1 is effective if we expect to suppress the over-fitting in the context of regression under Gaussian noise.

### B. Numerical Example

We consider a simple numerical experiment on one dimensional regression by a Gaussian unit whose output is given by

$$f_{c,b_0,b_1}(x) = c \exp\left\{-\frac{1}{b_0}(x - b_1)^2\right\}, \; x \in \mathbf{R} \tag{17}$$

where $c \in \mathbf{R}$, $b_0 \in \mathbf{B}_0$, $b_1 \in \{x_1, \ldots, x_n\}$, in which $x_1, \ldots, x_n$ are training inputs. The output data is a Gaussian noise sequence with mean zero and variance one. The input data are generated from the uniform distribution on $[-5, 5]$. Here, $\mathbf{B}_0$ is a set of 1000 points with log-scale in the range of $[10^{-5}, 100]$. The parameter space is $\mathbf{B}_0 \times \{x_1, \ldots, x_n\}$, for which the number of elements is finite. Thus, we can find the least squares estimator by searching every elements. We prepared 500 sets of $n$ samples as the training data and 500

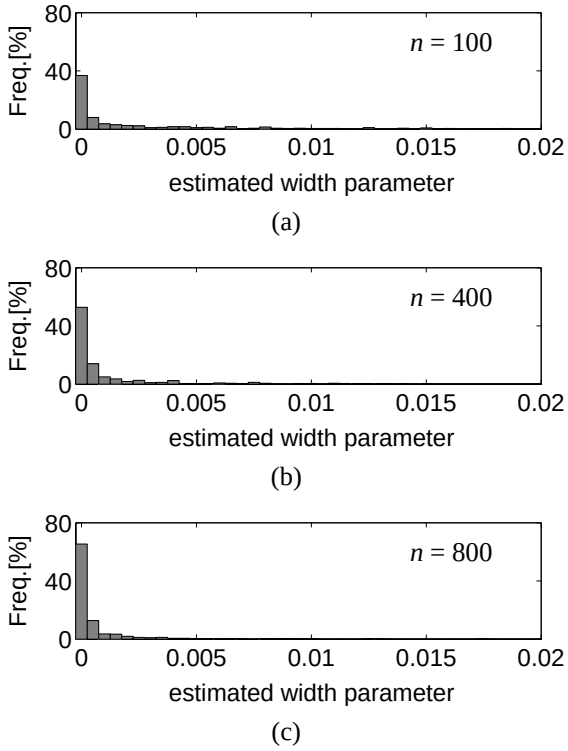


Fig. 1. The histograms of the estimates of the width parameter $b_0$ in the Gaussian unit at each $n$. The frequency of the histogram is written by the percentage for 500 trials.

runs of training are conducted. In each training, the estimate of $c$ that minimizes the squared error sum at each fixed $(b_0, b_1)$ is calculated. Then, the estimate of $(b_0, b_1)$ was chosen according to the squared error sums that have already been minimized according to $c$. Figure 1 shows the histograms of the estimated width parameters at $n = 100, 400, 800$. As we can see, the estimated width parameters frequently fall into the neighborhood of zero and the concentration is higher more and more as $n$ increases. This result is consistent with the above theoretical result.

## V. Conclusions and future works

In this article, we elucidated an over-fitting property of a Gaussian unit, which is trained under Gaussian noise. We showed that the probability that an extremely small value for the width parameter is obtained goes to one as the number of samples goes to infinity. Although we mainly focused on the case of one basis function, the function $g$ appeared in (1) does not necessarily have only one component. Consider the function $g$ defined by $g_{\boldsymbol{b}}(X) = \sum_{j=1}^{m} b_{0,j} h_{\boldsymbol{b}_{1,j}}(X)$, where $\boldsymbol{b} = (\boldsymbol{b}_0, \boldsymbol{b}_1)$, $\boldsymbol{b}_0 = (b_{0,1}, \ldots, b_{0,m})$ and $\boldsymbol{b}_1 = (\boldsymbol{b}_{1,1}, \ldots, \boldsymbol{b}_{1,m}) \in \mathbf{B}^m$. We assume that $\sum_{j=1}^{m} |b_{0,j}| \leq 1$ and $|h_{\boldsymbol{b}_{1,j}}(X)| \leq 1$ for any $\boldsymbol{b}_{1,j} \in \mathbf{B}$ and any $X \in \mathbf{R}^d$. Then, $|g_{\boldsymbol{b}}(X)| \leq 1$ for any $\boldsymbol{b}$ and $X$. Notice that the network output is generally given by $f_{\boldsymbol{v},\boldsymbol{b}_1}(X) = \sum_{j=1}^{m} v_j h_{\boldsymbol{b}_{1,j}}(X)$, where $(\boldsymbol{v}, \boldsymbol{b}_1)$ is a weight vector. If we define $\lambda := \sum_{j=1}^{m} |v_j|$, we have $f_{\boldsymbol{v},\boldsymbol{b}_1}(X) = \lambda \sum_{j=1}^{m} \frac{v_j}{\lambda} h_{\boldsymbol{b}_{1,j}}(X)$ for $\lambda \neq 0$. Hence, if we set $b_{0,j} = v_j/\lambda$, $c = \lambda$ then $f_{c,\boldsymbol{b}}(X) = f_{\boldsymbol{v},\boldsymbol{b}_1}(X)$ for any $X$. Since $\lambda = 0$ implies $f_{\boldsymbol{v},\boldsymbol{b}_1}(X) = 0$ for any $X$, the above representation gives a reparametrization for the general network form. Therefore, Theorem 1 is valid for general cases, and we can say that Assumption 1 controls the degree of over-fitting also in neural networks and radial basis function networks. In this case, the degree of over-fitting is asymptotically larger than $2k \log n$ if the network is not restricted[7][10], where $k$ is number of hidden nodes. Hence, the restricted network, which satisfies Assumption 1, is rarely chosen in training when the number of samples is large enough. According to the analysis on a Gaussian unit, one may understand that the role of Assumption 1 in this article is to exclude the extreme locality of the output on the input space. However, if two hidden nodes or radial basis functions are nearly linearly dependent in the above formulation, this assumption is also violated. Hence, we cannot specify the source of over-fitting in general cases while the output of the sum of the functions that are nearly linearly dependent can be sharpened on the input space. In this meaning, more deeper considerations is needed for the extension.

## Acknowledgment

This research was supported in part by Grants-in-Aid for Scientific Research 15700187 from the Ministry of Education, Science, Sports and Culture, Japan.

## References

[1] H. Akaike, "Information Theory and an Extension of the Maximum Likelihood Principle", In *2nd International Symposium on Information Theory*, B.N.Petrov and F.Csáki eds., Akadémia Kiado, Budapest, 1973, pp.267-281.

[2] S. Amari, and N. Murata, "Statistical theory of learning curves under entropic loss criterion", *Neural Computation*, **5**, pp.140-153, 1993.

[3] S. Amari, and T. Ozeki, "Differential and algebraic geometry of multilayer perceptrons", *IEICE Trans. Fundamentals*, **E84-A**, pp.31-38, 2001.

[4] M. Anthony, and P. L. Bartlett, *Neural network learning: Theoretical foundations*, Cambridge University Press, 1999.

[5] A. R. Barron, "Approximation and estimation bounds for artificial neural networks", *Machine Learning*, **14**, pp.115-133, 1994.

[6] P. L. Bartlett, "The sample complexity of pattern classification with neural networks: The size of the weights is more important than the size of the network", *IEEE Trans.on Information Theory*, **44**, 2, pp.525-536, 1998.

[7] K. Fukumizu, "Likelihood Ratio of Unidentifiable Models and Multilayer Neural Networks", *Annals of Statistics*, **31**, pp.833-851, 2003.

[8] K. Hagiwara, N. Toda, and S. Usui, "On the problem of applying AIC to determine the structure of a layered feedforward neural network", In *Proceedings. of IJCNN*, Nagoya, Japan, **3**, 1993, pp.2263-2266.

[9] K. Hagiwara, "On the training error and generalization error of neural network regression without identifiability", In *Proceedings. of KES'2001*, Nara, Japan, **2**, 2001, pp.1575-1579.

[10] K. Hagiwara, "On the problem in model selection of neural network regression in overrealizable scenario", *Neural Computation*, **14**, pp.1979-2002, 2002.

[11] D. Haussler, "Decision Theoretic Generalization of the PAC Model for Neural Net and Other Learning Applications", *Information and Computation*, **100**, pp.78-150, 1992.

[12] A. Krzyżak, T. Linder, and G. Lugosi, "Nonparametric estimation and classification using radial basis function nets and empirical risk minimization", *IEEE Trans. on Neural Networks*, **7**, pp.475-487, 1996.

[13] A. Krzyżak, and T. Linder, "Radial basis function networks and complexity regularization in function learning", *IEEE Trans. on Neural Networks*, **9**, 2, pp.247-256, 1998.

[14] M. R. Leadbetter, G. Lindgren, and H. Rootz'en, *Extremes, and related properties of random sequences and processes*, Springer-Verlag, 1983.
[15] N. Murata, S. Yoshizawa, and S. Amari, "Network information criterion – determining the number of hidden units for an artificial neural network model", *IEEE Trans. on Neural Networks*, *5*, 6, pp.865-872, 1994.
[16] N. Murata, "Bias of estimators and regularization terms", In *Proceedings of IBIS'98*, Hakone, Japan, 1998, pp.87-94.
[17] J. Pickands III, "An iterated logarithm law for the maximum in a stationary Gaussian sequence", *Z. Wahrscheinlichkeitstheorie verw. Geb.*, **12**, pp.344-353, 1969.
[18] D. Pollard, *Convergence of Stochastic Processes*, Springer-Verlag, 1984.
[19] V. Vapnik, "Statistical learning theory", John Wiley & Sons, 1998.
[20] H. White, "Learning in artificial neural networks : A statistical perspective", *Neural Computation*, **1**, pp.425-464, 1989.
[21] L. Xu, A. Krzyżak and A. Yuille, "On radial basis function nets and kernel regression: Statistical consistency, convergence rates, and receptive field size", *Neural Networks*, **7**, pp.609-628, 1994.

## APPENDIX

We fix $\mathbf{X}$ until the last part of this section and write $\mathbf{G}$ instead of $\mathbf{G}(\mathbf{X})$ and $\boldsymbol{g} = (g_1, \ldots, g_n)$ instead of $\boldsymbol{g}(\mathbf{X})$ for a while. For $\boldsymbol{f} = (f_1, \ldots, f_n) \in \mathbf{R}^n$ and $\boldsymbol{h} = (h_1, \ldots, h_n) \in \mathbf{R}^n$, we define the $l_1$-norm $\rho_1$ by $\rho_1(\boldsymbol{f}, \boldsymbol{h}) = \frac{1}{n}\sum_{i=1}^n |f_i - h_i|$, the Euclidean norm $\|\cdot\|$ by $\|\boldsymbol{f}\| = \left(\sum_{i=1}^n |f_i|^2\right)^{1/2}$ and the Euclidean inner product $\langle \cdot, \cdot \rangle$ by $\langle \boldsymbol{f}, \boldsymbol{h} \rangle = \sum_{i=1}^n f_i h_i$.

We need some lemmas to prove Theorem 1.

Let $\mathbf{T}_1$ and $\mathbf{T}_2$ be the minimal $\epsilon_1$-covering and $\epsilon_2$-covering of $\mathbf{G} \subseteq \mathbf{R}^n$ with respect to the $l_1$-norm, whose sizes are denoted by $n_l = N(\epsilon_l, \mathbf{G}, \rho_l)$, $l = 1, 2$ respectively. Since, for all $\boldsymbol{g} \in \mathbf{G}$, $|g(X)| \leq 1$ holds for any $X \in \mathbf{R}^d$ by the definition, we have the following fact.

*Lemma 1:* ([4][11][18]) Under Assumption 2,

$$n_l \leq C_\gamma (1/\epsilon_l)^{2\gamma} \tag{18}$$

holds for any $\epsilon_l > 0$, where $l = 1, 2$ and $C_\gamma$ is a constant that depends only on $\gamma$.

Since $|g_i| \leq 1$ for all $i$, the following is trivial.

*Lemma 2:* For any $\boldsymbol{t} \in \mathbf{T}_1$, $|t_i| \leq 1$ holds for all $1 \leq i \leq n$.

The following lemmas are used to prove the main claim in this section.

*Lemma 3:* If we set $\epsilon_1 < \delta/4$ then $\|\boldsymbol{t}\|^2 > n\delta/2$ holds for any $\boldsymbol{t} \in \mathbf{T}_1$ under Assumption 1.

*Lemma 4:* Under Assumption 2, there exists a partition $\{\mathbf{S}_1, \ldots, \mathbf{S}_{n_2} : \mathbf{S}_k \subseteq \mathbf{R}^n\}$ of $\mathbf{T}_1$ such that $\mathbf{T}_1 = \bigcup_{k=1}^{n_2} \mathbf{S}_k$ and $\mathbf{S}_j \bigcap \mathbf{S}_k = \phi$ for any $j \neq k$ hold. Furthermore, for any $\boldsymbol{t}, \boldsymbol{s} \in \mathbf{S}_k$, $\rho_1(\boldsymbol{t}, \boldsymbol{s}) < 2\epsilon_1 + 2\epsilon_2$ holds for all $k$.

We omit these proofs here.

*Lemma 5:* Let us choose $\epsilon_1$ so as to satisfy $\epsilon_1 < \delta/4$, where $\delta$ is in Assumption 1. Then, under Assumption 1 and 2, the conditional probability distribution of $Z(\boldsymbol{g})$ given $\mathbf{X}$ is a Gaussian distribution with mean zero and variance one for any $\boldsymbol{g} \in \mathbf{S}_k$ and for any $1 \leq k \leq n_2$. Furthermore,

$$\begin{aligned} \mathrm{Cov}(Z(\boldsymbol{g}_l), Z(\boldsymbol{g}_m)|\mathbf{X}) &:= \mathbf{E}(Z(\boldsymbol{g}_l) Z(\boldsymbol{g}_m)|\mathbf{X}) \\ &\geq 1 - \frac{8(\epsilon_1 + \epsilon_2)}{\delta} \end{aligned} \tag{19}$$

holds for any pair of $\boldsymbol{g}_l, \boldsymbol{g}_m \in \mathbf{S}_k$.

**Proof.** Let us fix a pair $\boldsymbol{g}_l, \boldsymbol{g}_m \in \mathbf{S}_k$ arbitrarily. By the definition of $\mathbf{Y}$, we obtain $\mathbf{E}(Z(\boldsymbol{g}_l)) = \mathbf{E}(Z(\boldsymbol{g}_m)) = 0$ and

$$\mathbf{E}(Z(\boldsymbol{g}_l) Z(\boldsymbol{g}_m)|\mathbf{X}) = \begin{cases} 1 & l = m \\ \dfrac{\langle \boldsymbol{g}_l, \boldsymbol{g}_m \rangle}{\|\boldsymbol{g}_l\| \|\boldsymbol{g}_m\|} & l \neq m \end{cases}. \tag{20}$$

By (9), it is easy to see that the first part of the assertion holds. Here, $\rho_1(\boldsymbol{g}_l, \boldsymbol{g}_m) < 2\epsilon_1 + 2\epsilon_2$ holds by lemma 4. Additionally, $\rho_1(\boldsymbol{g}_l, \boldsymbol{g}_m) \geq \|\boldsymbol{g}_l - \boldsymbol{g}_m\|^2/2n$ holds since all coordinates of $\boldsymbol{g}_l$ and $\boldsymbol{g}_m$ are in $[-1, 1]$ by Lemma 2. Thus, $\|\boldsymbol{g}_l - \boldsymbol{g}_m\|^2 < 4n(\epsilon_1 + \epsilon_2)$ holds. This and the trivial relation $2\|\boldsymbol{g}_l\|\|\boldsymbol{g}_m\| \leq \|\boldsymbol{g}_l\|^2 + \|\boldsymbol{g}_m\|^2$ yield $\mathrm{Cov}(Z(\boldsymbol{g}_l), Z(\boldsymbol{g}_m)|\mathbf{X}) \geq 1 - 4n(\epsilon_1 + \epsilon_2)/\|\boldsymbol{g}_l\|\|\boldsymbol{g}_m\|$. Since $\boldsymbol{g}_l, \boldsymbol{g}_m \in \mathbf{S}_k \subseteq \mathbf{T}_1$, $\|\boldsymbol{g}_l\|\|\boldsymbol{g}_m\| \geq n\delta/2$ by Lemma 3. Thus, by putting this into the previous inequality, we obtain (19). □

*Lemma 6:* Let us choose $\epsilon_1$ so as to satisfy $\epsilon_1 < \delta/4$, where $\delta$ is in Assumption 1. Then, under Assumption 1, for any $\boldsymbol{g} \in \mathbf{G}$, there exist some $\boldsymbol{t} \in \mathbf{T}_1$, which satisfies

$$Z^2(\boldsymbol{g}) - Z^2(\boldsymbol{t}) \leq \frac{8n\epsilon_1}{\delta^2} \max_{1 \leq i \leq n} Y_i^2. \tag{21}$$

**Proof.** First, $Z(\boldsymbol{g})$ and $Z(\boldsymbol{t})$ are well-defined for any $\boldsymbol{g} \in \mathbf{G}$ and $\boldsymbol{t} \in \mathbf{T}_1$ by Assumption 1 and the proof of Lemma 5 under the above choice of $\epsilon_1$. For any $\boldsymbol{g} \in \mathbf{G}$ and $\boldsymbol{t} \in \mathbf{T}_1$, according to the definition of $Z$, we obtain

$$\begin{aligned} & Z^2(\boldsymbol{g}) - Z^2(\boldsymbol{t}) \\ \leq\ & \frac{1}{\|\boldsymbol{g}\|^2} \left| \langle \mathbf{Y}, \boldsymbol{g} \rangle^2 - \langle \mathbf{Y}, \boldsymbol{t} \rangle^2 \right| \\ & + \frac{\langle \mathbf{Y}, \boldsymbol{t} \rangle^2}{\|\boldsymbol{g}\|^2 \|\boldsymbol{t}\|^2} \left| \|\boldsymbol{t}\|^2 - \|\boldsymbol{g}\|^2 \right|. \end{aligned} \tag{22}$$

Since $\boldsymbol{g} \in \mathbf{G}$, there are some $\boldsymbol{t} \in \mathbf{T}_1$ such that $\epsilon_1$-ball with the center $\boldsymbol{t}$ includes the given $\boldsymbol{g}$. Let us pick up one such $\boldsymbol{t} \in \mathbf{T}_1$. Let us denote $\boldsymbol{g} = (g_1, \ldots, g_n)$ and $\boldsymbol{t} = (t_1, \ldots, t_n)$. For the first term of the right hand side of (22), by Assumption 1, we obtain

$$\begin{aligned} & \frac{1}{\|\boldsymbol{g}\|^2} \left| \langle \mathbf{Y}, \boldsymbol{g} \rangle^2 - \langle \mathbf{Y}, \boldsymbol{t} \rangle^2 \right| \\ \leq\ & \frac{1}{\delta n} \left| \langle \mathbf{Y}, \boldsymbol{g} \rangle + \langle \mathbf{Y}, \boldsymbol{t} \rangle \right| \left| \langle \mathbf{Y}, \boldsymbol{g} \rangle - \langle \mathbf{Y}, \boldsymbol{t} \rangle \right| \\ \leq\ & \frac{2}{\delta} \max_{1 \leq i \leq n} Y_i^2 \sum_{i=1}^n |g_i - t_i| \leq \frac{2n\epsilon_1}{\delta} \max_{1 \leq i \leq n} Y_i^2, \end{aligned}$$

where we use $|g_i| \leq 1$ by the definition of $G$ and $|t_i| \leq 1$ by Lemma 2.

Through a similar calculation to the above one together with Lemma 3, for the second term of the right hand side of (22), we obtain

$$\begin{aligned} & \frac{\langle \mathbf{Y}, \boldsymbol{t} \rangle^2}{\|\boldsymbol{g}\|^2 \|\boldsymbol{t}\|^2} \left| \|\boldsymbol{t}\|^2 - \|\boldsymbol{g}\|^2 \right| \\ \leq\ & \frac{2n\epsilon_1}{\delta(\delta - 2\epsilon_1)} \max_{1 \leq i \leq n} Y_i^2 \\ \leq\ & \frac{4n\epsilon_1}{\delta^2} \max_{1 \leq i \leq n} Y_i^2, \end{aligned}$$

where the last inequality comes from the choice of $\epsilon_1$. Since $\delta \in (0,1)$ by Assumption 1, we obtain (21). □

**proof of Theorem 1** First, we obtain

$$\begin{aligned} & \mathbf{P}\left(\sup_{\boldsymbol{g}\in\mathbf{G}} Z^2(\boldsymbol{g}) > \alpha\log n + \beta \middle| \mathbf{X}\right) \\ \leq\ & \mathbf{P}\left(\sup_{\boldsymbol{g}\in\mathbf{G}} (Z^2(\boldsymbol{g}) - Z^2(\boldsymbol{t})) > \beta \middle| \mathbf{X}\right) \\ & + \mathbf{P}\left(\max_{\boldsymbol{t}\in\mathbf{T}_1} Z^2(\boldsymbol{t}) > \alpha\log n \middle| \mathbf{X}\right) \\ :=\ & P_1(n) + P_2(n). \end{aligned} \tag{23}$$

Let us firstly consider $P_1(n)$. By applying Lemma 6, $P_1(n)$ is bounded by $\mathbf{P}\left(\max_{1\leq i\leq n} Y_i^2 > \beta\delta^2/(8n\epsilon_1)\right)$ if $\epsilon_1 < \delta/4$. Let us set $\epsilon_1 = 1/n^2$. If $n > 2/\sqrt{\delta}$ then this is bounded above by $\mathbf{P}\left(\max_{1\leq i\leq n} Y_i^2 > n\beta\delta^2/8\right)$. It is easy to see that $\mathbf{P}(Y_1^2 > y) \leq \frac{2}{\sqrt{2\pi y}} e^{-y/2}$, since $Y_1$ is a Gaussian random variable. Furthermore, $\mathbf{P}(\max_{1\leq i\leq n} Y_i^2 > y) \leq n\mathbf{P}(Y_1^2 > y)$ since $Y_i$'s are i.i.d.. Hence, we obtain

$$P_1(n) \leq C_1'(\delta,\beta) n^{1/2} e^{-C_2'(\delta,\beta)n}, \tag{24}$$

where $C_1'(\delta,\beta) = 4/\sqrt{\pi\delta^2\beta}$ and $C_2'(\delta,\beta) = \delta^2\beta/16$.

Next, let us consider $P_2(n)$. By Lemma 4, we obtain

$$\begin{aligned} P_2 &= \mathbf{P}\left(\max_{1\leq k\leq n_2} \max_{\boldsymbol{t}\in\mathbf{S}_k} Z^2(\boldsymbol{t}) > \alpha\log n \middle| \mathbf{X}\right) \\ &\leq 2\sum_{k=1}^{n_2} \mathbf{P}\left(\max_{\boldsymbol{t}\in\mathbf{S}_k} Z(\boldsymbol{t}) > \sqrt{\alpha\log n} \middle| \mathbf{X}\right), \end{aligned} \tag{25}$$

where the last line holds due to the property of Gaussian random variables. Let us fix $k$ and define $m_k := |\mathbf{S}_k|$. Since we set $\epsilon_1 = 1/n^2$, by Lemma 5,

$$\mathrm{Cov}(Z(\boldsymbol{t})Z(\boldsymbol{s})|\mathbf{X}) \geq 1 - 16\epsilon_2/\delta$$

holds for any $\boldsymbol{t}, \boldsymbol{s} \in \mathbf{S}_k$ when $n > \sqrt{1/\epsilon_2}$. If we set $\epsilon_2 = \theta\delta/16$ for $\theta \in (0,1)$, then $\mathrm{Cov}(Z(\boldsymbol{t})Z(\boldsymbol{s})|\mathbf{X}) \geq 1-\theta$ holds if $n > \sqrt{1/\epsilon_2}$, where $\theta$ is specified later to simplifying the description below. Let $Z_1', \ldots, Z_{m_k}'$ be random variables, for which $Z_i' \sim N(0,1)$ for any $i$ and $\mathrm{Cov}(Z_i', Z_j') = 1-\theta$ for any $i \neq j$, and $Z_1', \ldots, Z_{m_k}'$ are independent. Then, we obtain

$$\begin{aligned} & \mathbf{P}\left(\max_{\boldsymbol{t}\in\mathbf{S}_k} Z(\boldsymbol{t}) > \sqrt{\alpha\log n} \middle| \mathbf{X}\right) \\ \leq\ & \mathbf{P}\left(\max_{1\leq i\leq m_k} Z_i' > \sqrt{\alpha\log n} \middle| \mathbf{X}\right) \\ :=\ & P_2'(n) \end{aligned} \tag{26}$$

by applying Corollary 4.2.3 in [14], p.84. Next, let us follow the idea in [14], p.138. Let $W, W_1, \ldots, W_{m_k}$ be i.i.d. random variables according to $N(0,1)$. We define $V_i = \sqrt{\theta}W_i + \sqrt{1-\theta}W$. Then, we can easily show that $V_i \sim N(0,1)$ and $\mathrm{Cov}(V_i, V_j) = 1-\theta$ for $i \neq j$. Therefore, the joint probability distribution of $V_1, \ldots, V_{m_k}$ is the same as that of $Z_1', \ldots, Z_{m_k}'$. Hence, $\max_{1\leq i\leq m_k} V_i = \sqrt{\theta}\max_{1\leq i\leq m_k} W_i + \sqrt{1-\theta}W$ holds[14]. By using this, we obtain

$$\begin{aligned} P_2'(n) &= \mathbf{P}\left(\max_{1\leq i\leq m_k} V_i > \sqrt{\alpha\log n} \middle| \mathbf{X}\right) \\ &\leq \mathbf{P}\left(\max_{1\leq i\leq m_k} W_i > \sqrt{\frac{\alpha\log n}{4\theta}} \middle| \mathbf{X}\right) \\ &\quad + \mathbf{P}\left(W > \sqrt{\frac{\alpha\log n}{4(1-\theta)}}\right). \end{aligned} \tag{27}$$

By Lemma 1, $n_1 \leq C_\gamma(1/\epsilon_1)^{2\gamma}$. As we set $\epsilon_1 = 1/n^2$, $n_1 \leq C_\gamma n^{4\gamma} := n_1'$. Since $\mathbf{S}_k \subseteq \mathbf{T}_1$, $m_k \leq n_1'$ holds for any $k$. Let $W_{m_k+1}, \ldots, W_{n_1'}$ be random variables, which satisfy that $W_1, \ldots, W_{m_k}, W_{m_k+1}, W_{n_1'}$ are i.i.d. random variables. Then, we obtain

$$\begin{aligned} & \mathbf{P}\left(\max_{1\leq i\leq m_k} W_i > \sqrt{\frac{\alpha\log n}{4\theta}} \middle| \mathbf{X}\right) \\ <\ & \mathbf{P}\left(\max_{1\leq i\leq n_1'} W_i > \sqrt{\frac{\alpha\log n}{4\theta}}\right). \end{aligned} \tag{28}$$

By (25), (26), (27) and (28), we obtain

$$\begin{aligned} P_2(n) &\leq 2n_2\mathbf{P}\left(\max_{1\leq i\leq n_1'} W_i > \sqrt{\frac{\alpha\log n}{4\theta}}\right) \\ &\quad + 2n_2\mathbf{P}\left(W > \sqrt{\frac{\alpha\log n}{4(1-\theta)}}\right). \end{aligned} \tag{29}$$

Since $W_1 \sim N(0,1)$, $\mathbf{P}(W_1 > w) \leq \frac{1}{\sqrt{2\pi}w} e^{-w^2/2}$ holds. Furthermore, we have $\mathbf{P}(\max_{1\leq i\leq n_1'} W_i > w) \leq n_1'\mathbf{P}(W_1 > w)$. Thus, we obtain

$$\begin{aligned} & \mathbf{P}\left(\max_{1\leq i\leq n_1'} W_i > \sqrt{\frac{\alpha\log n}{4\theta}}\right) \\ \leq\ & \frac{C_3'(\alpha,\theta,\gamma)}{\sqrt{\log n}} n^{4\gamma - \alpha/(8\theta)}, \end{aligned} \tag{30}$$

where $C_3'(\alpha,\theta,\gamma) = C_\gamma\sqrt{2\theta/(\alpha\pi)}$. On the other hand, by using the same bound for the upper probability for a Gaussian distribution in the above, we obtain

$$\mathbf{P}\left(W > \sqrt{\frac{\alpha\log n}{4(1-\theta)}}\right) \leq \frac{C_4'(\alpha,\theta,\gamma)}{\sqrt{\log n}} n^{-\alpha/(8(1-\theta))}, \tag{31}$$

where $C_4'(\alpha,\theta,\gamma) = C_\gamma\sqrt{2(1-\theta)/(\alpha\pi)}$. At last, by Lemma 1, $n_2 \leq C_\gamma(1/\epsilon_2)^{2\gamma}$ holds. Since we set $\epsilon_2 = \theta\delta/16$, $n_2 \leq C_\gamma(16/(\theta\delta))^{2\gamma} := C_5(\delta,\theta,\gamma)$ holds and, thus, this is constant. By setting the constants $C_m = 2C_m'(\cdot)C_5(\delta,\theta,\gamma)$ for $1 \leq m \leq 4$ and $\theta = 1/(32\gamma/\alpha + 8)$, the proof is completed. □