

# A Limit Theorem in Singular Regression Problem

Sumio Watanabe

Precision and Intelligence Laboratory

Tokyo Institute of Technology

4259 Nagatsuta, Midoriku, Yokohama, 226-8503A Japan

e-mail:swatanab@pi.titech.ac.jp

December 21, 2013

## Abstract

In statistical problems, a set of parameterized probability distributions is used to estimate the true probability distribution. If Fisher information matrix at the true distribution is singular, then it has been left unknown what we can estimate about the true distribution from random samples. In this paper, we study a singular regression problem and prove a limit theorem which shows the relation between the singular regression problem and two birational invariants, a real log canonical threshold and a singular fluctuation. The obtained theorem has an important application to statistics, because it enables us to estimate the generalization error from the training error without any knowledge of the true probability distribution.

## 1 Introduction

Let  $M$  and  $N$  be natural numbers, and  $\mathbb{R}^M$  and  $\mathbb{R}^N$  be  $M$  and  $N$  dimensional real Euclidean spaces respectively. Assume that  $(\Omega, \mathcal{B}, P)$  is a probability space and that  $(X, Y)$  is an  $\mathbb{R}^M \times \mathbb{R}^N$ -valued random variable which is subject to a simultaneous probability density function,

$$q(x, y) = \frac{q(x)}{(2\pi\sigma^2)^{N/2}} \exp\left(-\frac{|y - r_0(x)|^2}{2\sigma^2}\right),$$

where  $q(x)$  is a probability density function on  $\mathbb{R}^M$ ,  $\sigma > 0$  is a constant,  $r_0(x)$  is a measurable function from  $\mathbb{R}^M$  to  $\mathbb{R}^N$ , and  $|\cdot|$  is the Euclidean norm of  $\mathbb{R}^N$ . The function  $r_0(x)$  is called a regression function of  $q(x, y)$ . Assume that  $\{(X_i, Y_i); i = 1, 2, \dots, n\}$  is a set of random variables which are independently subject to the same probability distribution as  $(X, Y)$ . Let  $W$  be a subset of  $\mathbb{R}^d$ . Let  $r(x, w)$  be a function from  $\mathbb{R}^M \times W$  to  $\mathbb{R}^N$ . The square error  $H(w)$  is a real function on  $W$ ,

$$H(w) = \frac{1}{2} \sum_{i=1}^n |Y_i - r(X_i, w)|^2.$$

An expectation operator  $E_w[\cdot]$  on  $W$  is defined by

$$E_w[F(w)] = \frac{\int F(w) \exp(-\beta H(w)) \varphi(w) dw}{\int \exp(-\beta H(w)) \varphi(w) dw}, \quad (1)$$

where  $F(w)$  is a measurable function,  $\varphi(w)$  is a probability density function on  $W$ , and  $\beta > 0$  is a constant called an inverse temperature. Note that  $E_w[F(w)]$  is not a constant but a random variable because  $H(w)$  depends on random variables. Two random variables  $G$  and  $T$  are defined by

$$\begin{aligned} G &= \frac{1}{2} E_X E_Y [|Y - E_w[r(X, w)]|^2], \\ T &= \frac{1}{2n} \sum_{i=1}^n |Y_i - E_w[r(X_i, w)]|^2. \end{aligned}$$

These random variables  $G$  and  $T$  are called the generalization and training errors respectively. Since  $E_{X,Y} [|Y - r_0(X)|^2] = N\sigma^2$ , it is expected on some natural conditions that both  $E[G]$  and  $E[T]$  converge to  $S = N\sigma^2/2$  when  $n$  tends to infinity if there exists  $w_0 \in W$  such that  $r(x, w_0) = r_0(x)$ . In this paper, we ask how fast such convergences are, in other words, our study concerns with a limit theorem which shows the convergences  $n(E[G] - S)$  and  $n(E[T] - S)$ , when  $n \rightarrow \infty$ . If Fisher information matrix

$$I_{ij}(w) = \int \partial_i r(x, w) \cdot \partial_j r(x, w) q(x) dx,$$

where  $\partial_i = (\partial/\partial w_i)$ , is positive definite for arbitrary  $w \in W$ , then this problem is well known as a regular regression problem. In fact, in a regular regression problem, convergences  $n(E[G] - S) \rightarrow d\sigma^2/2$  and  $n(E[T] - S) \rightarrow -d\sigma^2/2$  hold. However, if  $I(w_0) = \{I_{ij}(w_0)\}$  is singular, that is to say, if  $\det I(w_0) = 0$ , then the problem is called a singular regression problem and convergences of  $n(E[G] - S)$  and  $n(E[T] - S)$  have been left unknown.

In general it has been difficult to study a limit theorem for the case when Fisher information matrix is singular. However, recently, we have shown that a limit theorem can be established based on resolution of singularities, and that there are mathematical relations between the limit theorem and two birational invariants in singular density estimation [16, 17, 18]. In this paper we prove a new limit theorem for the singular regression problem, which enables us to estimate birational invariants from random samples. The limit theorem proved in this paper has an important application to statistics, because the expectation value of the generalization error  $E[G]$  can be estimated from that of the training error  $E[T]$  without any knowledge of the true probability distribution.

**Example** Let  $M = N = 1$ ,  $d = 4$ ,  $w = (a, b, c, d)$ , and  $W = \{w \in \mathbb{R}^4; |w| \leq 1\}$ . If the function  $r(x, w)$  is defined by

$$r(x, w) = a \sin(bx) + c \sin(dx),$$

and  $r_0(x) = 0$ , then the set  $\{w \in W; r(x, w) = r_0(x)\}$  is not one point, and Fisher information matrix at  $(a, b, c, d) = (0, 0, 0, 0)$  is singular. A lot of functions used in statistics, information science, brain informatics, and bio-informatics are singular, for example, artificial neural networks, radial basis functions, and wavelet functions.

## 2 Main Results

We prove the main theorems based on the following assumptions.

### Basic Assumptions.

(A1) The set of parameters  $W$  is defined by

$$W = \{w \in \mathbb{R}^d; \pi_j(w) \geq 0 \ (j = 1, 2, \dots, k)\},$$

where  $\pi_j(w)$  is a real analytic function. It is assumed that  $W$  is a compact set in  $\mathbb{R}^d$  whose open kernel is not the empty set. The probability density function  $\varphi(w)$  on  $W$  is given by

$$\varphi(w) = \varphi_1(w)\varphi_2(w),$$

where  $\varphi_1(w) \geq 0$  is a real analytic function and  $\varphi_2(w) > 0$  is a function of class  $C^\infty$ .

(A2) Let  $s \geq 8$  be the number that is equal to 4 times of some integer. There exists an open set  $W^* \supset W$  such that  $r(x, w) - r_0(x)$  is an  $L^s(q)$ -valued analytic function on  $W^*$ , where  $L^s(q)$  is a Banach space defined by using its norm  $\|\cdot\|_s$ ,

$$L^s(q) = \{f; \|f\|_s = \left( \int |f(x)|^s q(x) dx \right)^{1/s} < \infty\}.$$

(A3) There exists a parameter  $w_0 \in W$  such that  $r(x, w_0) = r_0(x)$ .

If these basic assumptions are satisfied, then

$$K(w) = \frac{1}{2} \int |r(x, w) - r_0(x)|^2 q(x) dx \quad (2)$$

is a real analytic function on  $W^*$ . A subset  $W_a \subset W$  is defined by

$$W_a = \{w \in W; K(w) \leq a\}.$$

Note that  $W_0$  is the set of all points that satisfy  $K(w) = 0$ . In general,  $W_0$  is not one point and it contains singularities. This paper gives a limit theorem for such a case. Proofs of lemmas and theorems in this section are given in section 6.

**Lemma 1.** *Assume (A1), (A2), and (A3) with  $s \geq 4$ . Then*

$$\zeta(z) = \int_W K(w)^z \varphi(w) dw$$

is a holomorphic function on  $\text{Re}(z) > 0$  which can be analytically continued to the unique meromorphic function on the entire complex plane whose poles are all real, negative, and rational numbers.

**Lemma 2.** Assume (A1), (A2), and (A3) with  $s \geq 8$ . Then there exists a constant  $\nu = \nu(\beta) \geq 0$  such that

$$V = \sum_{i=1}^n \left( E_w[|r(X_i, w)|^2] - |E_w[r(X_i, w)]|^2 \right)$$

satisfies

$$\lim_{n \rightarrow \infty} E[V] = \frac{2\nu}{\beta}. \quad (3)$$

Based on Lemma 1 and 2, we define two important values  $\lambda, \nu > 0$ .

**Definition 2.1.** Let the largest pole of  $\zeta(z)$  be  $(-\lambda)$  and its order  $m$ . The constant  $\lambda > 0$  is called a real log canonical threshold. The constant  $\nu = \nu(\beta)$  is referred to as a singular fluctuation.

The real log canonical threshold is an important invariant of an analytic set  $K(w) = 0$ . For its relation to algebraic geometry and algebraic analysis, see [4, 5, 6, 9, 10, 11]. It is also important in statistical learning theory, and it can be calculated by resolution of singularities [16, 3]. The singular fluctuation is an invariant of  $K(w) = 0$  which is found in statistical learning theory [15, 18], whose relation to singularity theory is still unknown. The followings are main theorems of this paper.

**Theorem 1.** Assume the basic assumptions (A1), (A2), and (A3) with  $s \geq 8$ . Let  $S = N\sigma^2/2$ . Then

$$\lim_{n \rightarrow \infty} n(E[G] - S) = \frac{\lambda - \nu}{\beta} + \nu\sigma^2, \quad (4)$$

$$\lim_{n \rightarrow \infty} n(E[T] - S) = \frac{\lambda - \nu}{\beta} - \nu\sigma^2. \quad (5)$$

This theorem shows that both the real log canonical threshold  $\lambda$  and singular fluctuation  $\nu$  determine the singular regression problem.

**Theorem 2.** Assume the basic assumptions (A1), (A2), and (A3) with  $s \geq 12$ . Then

$$E[G] = E\left[\left(1 + \frac{2\beta V}{nN}\right)T\right] + o_n,$$

where  $o_n$  is a function of  $n$  which satisfies  $no_n \rightarrow 0$ .

By this theorem,  $V$  and  $T$  can be calculated from random samples without any direct knowledge of the true regression function  $r_0(x)$ . Therefore,  $E[G]$  can be estimated from random samples, resulting that we can find the optimal model or hyperparameter for the smallest generalization error. If the model is

regular, then  $\lambda = \nu = d/2$  for arbitrary  $0 < \beta \leq \infty$ , resulting that Theorem 2 coincides with AIC [1] of a regular statistical model. Therefore, Theorem 2 is a widely applicable information criterion, which we can apply to both regular and singular problems. We use Theorem 2 without checking that the true distribution is a singularity or not.

### 3 Preparation of Proof

We use notations,  $S = N\sigma_2/2$  and

$$\begin{aligned} S_i &= Y_i - r_0(X_i), \\ f(x, w) &= r(x, w) - r_0(x). \end{aligned}$$

Then  $\{S_i\}$  are independent random variables which are subject to the normal distribution with average zero and covariance matrix  $\sigma^2 I$  where  $I$  is the  $d \times d$  identity matrix. It is immediately derived that

$$\begin{aligned} E[T] &= S - E\left[\frac{1}{n} \sum_{i=1}^n S_i \cdot E_w[f(X_i, w)]\right] \\ &\quad + E\left[\frac{1}{2n} \sum_{i=1}^n |E_w[f(X_i, w)]|^2\right], \\ E[G] &= S + \frac{1}{2} E[E_X[|E_w[f(X, w)]|^2]], \\ E[V] &= E\left[\sum_{i=1}^n \{E_w[|f(X_i, w)|^2] - |E_w[f(X_i, w)]|^2\}\right]. \end{aligned}$$

The function  $f(x, w)$  is an  $L^s(q)$ -valued analytic function on  $W^*$ . In eq.(1), we can define  $E_w[\ ]$  by replacing  $H(w)$  by  $H_0(w)$ ,

$$H_0(w) = \frac{1}{2} \sum_{i=1}^n |f(X_i, w)|^2 - \sum_{i=1}^n S_i \cdot f(X_i, w),$$

which can be rewritten as

$$H_0(w) = nK(w) - \sqrt{n} \eta_n(w),$$

where  $K(w)$  is given in eq.(2), and

$$\begin{aligned} \eta_n(w) &= \eta_n^{(1)}(w) + \eta_n^{(2)}(w), \\ \eta_n^{(1)}(w) &= \frac{1}{\sqrt{n}} \sum_{i=1}^n S_i \cdot f(X_i, w), \\ \eta_n^{(2)}(w) &= \frac{1}{\sqrt{n}} \sum_{i=1}^n (K(w) - \frac{1}{2} |f(X_i, w)|^2). \end{aligned}$$

We define a norm  $\| \cdot \|$  of a function of  $f$  on  $W$  by

$$\|f\| = \sup_{w \in W} |f(w)|.$$

Since  $W$  is a compact set of  $\mathbb{R}^d$ , the set  $B(W)$  that is a set of all continuous and bounded function on  $W$  is a Polish space, and both  $\eta_n^{(1)}(w)$  and  $\eta_n^{(2)}(w)$  are  $B(W)$ -valued random variables. Because  $f(X, w)$  is an  $L^s(q)$ -valued analytic function,  $\{\eta_n^{(1)}\}$  and  $\{\eta_n^{(2)}\}$  are tight random processes, resulting that  $\eta_n^{(1)}$  and  $\eta_n^{(2)}$  weakly converge to unique tight gaussian processes  $\eta^{(1)}$  and  $\eta^{(2)}$  respectively which have the same covariance matrices as  $\eta_n^{(1)}$  and  $\eta_n^{(2)}$  respectively when  $n \rightarrow \infty$  [13, 17, 18].

**Lemma 3.** Assume (A1), (A2), and (A3) with  $s \geq 8$ . Then

$$\begin{aligned} E[\|\eta_n^{(1)}\|^s] &< \infty, \\ E[\|\eta_n^{(2)}\|^{s/2}] &< \infty, \end{aligned}$$

*Proof.* Since  $f(x, w)$  is an  $L^s(q)$ -valued analytic function, it is represented by the absolutely convergent power series  $f(x, w) = \sum_j a_j(x)w^j$  which satisfies  $|a_j(x)| \leq M(x)/r^j$  for some function  $M(x) \in L^s(q)$  where  $r = (r_1, \dots, r_d)$  is the associative convergence radii. By using this fact, the former inequality is proved [17, 18]. Also  $K(w) - (1/2)f(x, w)^2$  is an  $L^{s/2}(q)$ -valued analytic function, the latter inequality is proved.  $\square$

**Lemma 4.** For arbitrary natural number  $n$ ,

$$\begin{aligned} E[E_w[\sqrt{n} \eta_n^{(1)}(w)]] &= \sigma^2 \beta E[V], \\ E[E_w[\sqrt{n} \eta_n^{(2)}(w)]] &= E[E_w[nK(w) - \frac{1}{2} \sum_{i=1}^n |f(X_i, w)|^2]]. \end{aligned}$$

*Proof.* The second equation is trivial. Let us prove the first equation. Let the left hand side of the first equation be  $A$ . Since  $\{S_i\}$  are independently subject to the normal distribution with covariance matrix  $\sigma^2 I$ ,

$$\begin{aligned} A &= E\left[\sum_{i=1}^n S_i \cdot E_w[f(X_i, w)]\right] \\ &= \sigma^2 E\left[\sum_{i=1}^n \nabla_{S_i} \cdot E_w[f(X_i, w)]\right] \\ &= \sigma^2 E\left[\sum_{i=1}^n \nabla_{S_i} \cdot \left(\frac{\int f(X_i, w) \exp(-\beta H_0(w)) \varphi(w) dw}{\int \exp(-\beta H_0(w)) \varphi(w) dw}\right)\right] \\ &= \beta \sigma^2 E\left[\sum_{i=1}^n E_w[|f(X_i, w)|^2] - |E_w[f(X_i, w)]|^2\right], \end{aligned}$$

which is equal to the right hand side of the first equation.  $\square$

**Definition 3.1.** *Let us define five random variables.*

$$\begin{aligned}
D_1 &= nE_w[E_X[|f(X, w)|^2]], \\
D_2 &= nE_X[|E_w[f(X, w)]|^2], \\
D_3 &= \sum_{i=1}^n E_w[|f(X_i, w)|^2], \\
D_4 &= \sum_{i=1}^n |E_w[f(X_i, w)]|^2, \\
D_5 &= E_w[\sqrt{n} \eta_n(w)].
\end{aligned}$$

Then, by using Lemma 4, it follows that

$$E[G] = S + \frac{1}{2n}E[D_2], \quad (6)$$

$$E[T] = S - \frac{\beta\sigma^2}{n}E[D_3 - D_4] + \frac{1}{2n}E[D_4], \quad (7)$$

$$E[V] = E[D_3 - D_4], \quad (8)$$

$$E[D_5] = \beta\sigma^2 E[D_3 - D_4] + (1/2)E[D_1 - D_3]. \quad (9)$$

We show that five expectation values  $E[D_j]$  ( $j = 1, 2, 3, 4, 5$ ) converge to constants. To show such convergences, it is sufficient to prove that each  $D_j$  weakly converges to some random variable and that  $E[(D_j)^{1+\delta}] < C$  for some  $\delta > 0$  and constant  $C > 0$  [13].

**Definition 3.2.** *For a given constant  $\epsilon > 0$ , a localized expectation operator  $E_w^\epsilon[\ ]$  is defined by*

$$E_w^\epsilon[F(w)] = \frac{\int_{K(w) \leq \epsilon} F(w) \exp(-\beta H_0(w)) \varphi(w) dw}{\int_{K(w) \leq \epsilon} \exp(-\beta H_0(w)) \varphi(w) dw}. \quad (10)$$

Let  $D_i^\epsilon$  ( $i = 1, 2, 3, 4, 5$ ) be random variables that are defined by replacing  $E_w[\ ]$  by  $E_w^\epsilon[\ ]$ .

**Lemma 5.** *Let  $0 < \delta < s/4 - 1$ . For arbitrary  $\epsilon > 0$ ,  $j = 1, 2, 3, 4, 5$ ,*

$$\lim_{n \rightarrow \infty} E[|D_j - D_j^\epsilon|^{1+\delta}] = 0.$$

*Proof.* We can prove five equations by the same way. Let us prove the case  $j = 3$ . Let  $L(w) = \sum_{i=1}^n |f(X_i, w)|^2$ . Because  $f(x, w)$  is  $L^s(q)$ -valued analytic

function,  $E[(\|L\|/n)^{1+\delta}] < \infty$ .

$$\begin{aligned}
|D_3 - D_3^\epsilon| &\leq \frac{\int_{K(w) \geq \epsilon} L(w) \exp(-\beta H_0(w)) \varphi(w) dw}{\int_{K(w) \leq \epsilon} \exp(-\beta H_0(w)) \varphi(w) dw} \\
&\leq \frac{\|L\| e^{-n\beta\epsilon + 2\beta\sqrt{n}\|\eta_n\|}}{\int_{K(w) \leq \epsilon} \exp(-\beta n K(w)) \varphi(w) dw} \\
&\leq C_1 n^{d/2} \|L\| \exp(-n\beta\epsilon/2 + (2\beta/\epsilon)\|\eta_n\|^2)
\end{aligned}$$

where we used  $2\sqrt{n}\|\eta_n\| \leq (n\epsilon/2 + (2/\epsilon)\|\eta_n\|^2)$  and  $C_1 > 0$  is a constant. From Lemma 3,  $E[\|\eta_n\|^{s/2}] \equiv C_2 < \infty$ , hence by using  $C_3 = (8\epsilon^2)^{s/4} C_2$ ,

$$P(\|\eta_n\|^2 \geq n/(8\epsilon^2)) \leq C_3/n^{s/4}.$$

Let  $E[F]_A$  be the expectation value of  $F(x)I_A(x)$  where  $I_A(x)$  is the defining function of a set  $A$ , in other words,  $I_A(x) = 1$  if  $x \in A$  or 0 if otherwise.

$$\begin{aligned}
E[|D_3 - D_3^\epsilon|^{1+\delta}] &= E[|D_3 - D_3^\epsilon|^{1+\delta}]_{\{\|\eta_n\|^2 \geq n/(8\epsilon^2)\}} \\
&\quad + E[|D_3 - D_3^\epsilon|^{1+\delta}]_{\{\|\eta_n\|^2 < n/(8\epsilon^2)\}}.
\end{aligned}$$

The first term of the right hand side is not larger than  $C_3 E[\|L\|^{1+\delta}]/n^{s/4}$  and the second term is not larger than  $E[(C_1\|L\|)^{1+\delta}] n^{d/2} \exp(-n\beta\epsilon/4)$ . Both of them converge to zero.  $\square$

## 4 Resolution of Singularities

To study the expectation on the region  $W_\epsilon$  we need resolution of singularities because  $W_0$  contains singularities in general. Let  $\epsilon > 0$  be a sufficiently small constant. Then by applying Hironaka's theorem [7] to the real analytic function  $K(w) \prod_{j=1}^k \pi_j(w) \varphi_1(w)$ , all functions  $K(w)$ ,  $\pi_j(w)$ , and  $\varphi_1(w)$  are made normal crossing. In fact, there exist an open set  $W_\epsilon^* \subset W^*$  which contains  $W_\epsilon$ , a manifold  $U^*$ , and a proper analytic map  $g : U^* \rightarrow W_\epsilon^*$  such that in each local coordinate of  $U^*$ ,

$$\begin{aligned}
K(g(u)) &= u^{2k}, \\
\varphi(g(u)) |g(u)'| &= \phi(u) |u^h|,
\end{aligned}$$

where  $k = (k_1, \dots, k_d)$  and  $h = (h_1, \dots, h_d)$  are multi-indices ( $k_j$  and  $h_j$  are nonnegative integers),  $u^{2k} = \prod_j u_j^{2k_j}$ ,  $u^h = \prod_j u_j^{h_j}$ ,  $|g(u)'|$  is the absolute value of Jacobian determinant of  $w = g(u)$ , and  $\phi(u) > 0$  is a function of class  $C^\infty$ . Let  $U = g^{-1}(W_\epsilon)$ . Since  $g$  is a proper map and  $W_\epsilon$  is compact,  $U$  is also compact. Moreover, it is covered by a finite sum

$$U = \cup_\alpha U_\alpha,$$



where each  $U_\alpha$  can be taken to be  $[0, b]^d$  in each local coordinate using some  $b > 0$ , and

$$\int_{W_\epsilon} F(w) \varphi(w) dw = \sum_\alpha \int_{U_\alpha} F(g(u)) \phi_\alpha(u) |u^h| du,$$

where  $\phi_\alpha(u) \geq 0$  is a function of class  $C^\infty$ . In this paper, we apply these facts to analyzing the singular regression problem. For resolution of singularities and its applications, see [7] and [4],[16]. Lemma 1 is directly proved by these facts [4, 8, 16]. Moreover, the following lemma is simultaneously obtained.

**Lemma 6.** *The largest pole  $(-\lambda)$  and its order  $m$  of  $\zeta(z)$  are given by*

$$\lambda = \min_\alpha \min_j \left( \frac{h_j + 1}{2k_j} \right), \quad (11)$$

$$m = \max_\alpha \# \left\{ j; \lambda = \frac{h_j + 1}{2k_j} \right\}, \quad (12)$$

where, if  $k_j = 0$ ,  $(h_{j+1} + 1)/2k_j$  is defined to be  $+\infty$  and  $\#$  shows the number of elements of the set. Let  $\{U_{\alpha^*}\}$  be the set of all local coordinates that attain both  $\min_\alpha$  in eq.(11) and  $\max_\alpha$  in eq.(12). Such coordinates are referred to as the essential coordinates.

For a given real analytic function  $K(w)$ , there are infinitely many different resolutions of singularities. However,  $\lambda$  and  $m$  do not depend on the pair  $(U^*, g)$ . They are called birational invariants. By the definition of  $K(w)$  in eq.(2), there exists an  $L^s(q)$ -valued analytic function  $a(x, u)$  on each local coordinate in  $U^*$  such that

$$f(x, u) = a(x, u) u^k$$

and  $E_X[|a(X, u)|^2] = 2$ . Therefore,

$$H_0(g(u)) = n u^{2k} - \sqrt{n} u^k \xi_n(u),$$

where

$$\xi_n(u) = \frac{1}{\sqrt{n}} \sum_{i=1}^n S_i \cdot a(X_i, u) + \frac{1}{\sqrt{n}} \sum_{i=1}^n u^k \left( 1 - \frac{a(X_i, u)^2}{2} \right).$$

Then  $E[\|\xi_n\|^{s/2}] < \infty$  and  $E[\|\nabla \xi_n\|^{s/2}] < \infty$ , because both  $a(x, u)$  and  $\nabla a(x, u)$  are  $L^s(q)$ -valued analytic function, where  $\|\nabla \xi_n\| = \max_j \sup_w |\partial_j \xi_n(u)|$ . The expectation operator  $E_u[\ ]$  on  $U$  is defined so that it satisfies  $E_w^\epsilon[F(w)] = E_u[F(g(u))]$ . Then

$$\begin{aligned} D_1^\epsilon &= n E_u[2u^{2k}], \\ D_2^\epsilon &= n E_X[|E_u[a(X, u)u^k]|^2], \\ D_3^\epsilon &= \sum_{i=1}^n E_u[|a(X_i, u)|^2 u^{2k}], \\ D_4^\epsilon &= \sum_{i=1}^n |E_u[a(X_i, u)u^k]|^2, \\ D_5^\epsilon &= E_u[\sqrt{n} \xi_n(u) u^k]. \end{aligned}$$

**Lemma 7.** *Let  $s \geq 12$  and  $0 < \delta < s/6 - 1$ . For  $i = 1, 2, 3, 4, 5$ , there exists a constant  $C > 0$  such that  $E[(D_i^\epsilon)^{1+\delta}] < C$  holds.*

*Proof.* Since  $0 \leq D_4^\epsilon \leq D_3^\epsilon$ ,  $0 \leq D_2^\epsilon \leq D_1^\epsilon$ , and  $|D_5^\epsilon| \leq (\|\xi_n\|^2 + 2D_1^\epsilon)/2$ , it is sufficient to prove  $j = 1, 3$ . The proof for  $j = 1, 3$  can be done by the same way. Let us prove the case  $j = 3$ . In  $l = 1, 2, \dots, d$ , at least one of  $k_l \geq 1$ . By using partial integration for  $du_l$ , we can show that there exists  $c_1 > 0$  such that

$$E_u[u^{2k}] \leq \frac{c_1}{n} \{1 + \|\xi_n\|^2 + \|\nabla \xi_n\|^2\}. \quad (13)$$

Therefore by using  $L = (1/n) \sum_{i=1}^n \|a(X_i)\|^2$  and Hölder's inequality with  $1/3 + 1/(3/2) = 1$ ,

$$\begin{aligned} E[(D_3^\epsilon)^{1+\delta}] &\leq E[(c_1 L (1 + \|\xi_n\|^2 + \|\nabla \xi_n\|^2))^{1+\delta}] \\ &\leq E[(c_1 L)^{3+3\delta}]^{1/3} E[(1 + \|\xi_n\|^2 + \|\nabla \xi_n\|^2)^{(3+3\delta)/2}]^{3/2}. \end{aligned}$$

Since  $E[\|a(X)\|^s] < \infty$ ,  $E[\|\xi_n\|^{s/2}] < \infty$ , and  $E[\|\nabla \xi_n\|^{s/2}] < \infty$ , this expectation is finite.  $\square$

## 5 Renormalized distribution

**Definition 5.1.** *For a given function  $h(u)$  on  $U$ , the renormalized expectation operator  $E_{u,t}^*[\cdot|h]$  is defined by*

$$E_{u,t}^*[F(u, t)|h] = \frac{\sum_{\alpha^*} \int_0^\infty dt \int D(du) F(u, t) t^{\lambda-1} e^{-\beta t + \beta \sqrt{t} h(u)}}{\sum_{\alpha^*} \int_0^\infty dt \int D(du) t^{\lambda-1} e^{-\beta t + \beta \sqrt{t} h(u)}},$$

where  $D(du)$  is a measure which is defined in eq.(16) and  $\sum_{\alpha^*}$  shows the sum of all essential coordinates. Also we define

$$\begin{aligned} D_1^*(h) &= E_{u,t}^*[2t|h], \\ D_2^*(h) &= E_X[|E_{u,t}^*[a(X, u)\sqrt{t}]|^2|h], \\ D_5^*(h) &= E_{u,t}^*[h(u)\sqrt{t}|h]. \end{aligned}$$

**Lemma 8.** *The following convergences in probability hold.*

$$\begin{aligned} D_1^\epsilon - D_1^*(\xi_n) &\rightarrow 0, \\ D_2^\epsilon - D_2^*(\xi_n) &\rightarrow 0, \\ D_3^\epsilon - D_1^*(\xi_n) &\rightarrow 0, \\ D_4^\epsilon - D_2^*(\xi_n) &\rightarrow 0, \\ D_5^\epsilon - D_5^*(\xi_n) &\rightarrow 0. \end{aligned}$$

*Proof.* These five convergences can be proved by the same way. We show  $D_3^\epsilon - D_1^*(\xi_n) \rightarrow 0$ . Let  $L(u) = (1/n) \sum_{i=1}^n |a(X_i, u)|^2$ . Since  $E_X[|a(X, u)|^2] = 2$ ,

$$|D_3^\epsilon - D_1^*(\xi_n)| \leq |E_u[nL(u)u^{2k}] - E_u[E_X[a(X, u)^2]u^{2k}]| + |E_u[2u^{2k}] - E_{u,t}^*[2t|\xi_n]|.$$

Let the first and second terms of the left hand side of this inequality be  $D_6$  and  $D_7$  respectively. Then

$$D_6 \leq \|L - a(X)\|^2 E_u[nu^{2k}].$$

By the convergence in probability  $\|L - a(X)\| \rightarrow 0$  and eq.(13),  $D_6$  converges to zero in probability. From Lemma 10 and 11 in Appendix, it is derived that

$$|E_u[u^{2k}] - E_{u,t}^*[t|\xi_n]| \leq \frac{c_1}{\log n} \frac{e^{2\beta\|\xi_n\|^2}}{\min(\phi)^2} \{1 + \beta\|\nabla\xi_n\|\}, \quad (14)$$

which shows  $D_7 \rightarrow 0$  in probability.  $\square$

**Lemma 9.** *For arbitrary function  $h(u)$ , the following equality holds.*

$$D_1^*(h) = D_5^*(h) + \frac{2\lambda}{\beta}.$$

*Proof.* Let  $F_p(u)$  be a function defined by

$$F_p(u) = \int_0^\infty t^p t^{\lambda-1} e^{-\beta t + \beta\sqrt{t}h(u)} dt.$$

Then by using the partial integration of  $dt$ ,

$$F_1(u) = \frac{1}{2}h(u)F_{1/2}(u) + \frac{\lambda}{\beta}F_0(u).$$

By the definition of  $D_1^*(h) = E_{u,t}^*[2t|h]$  and  $D_5^*(h) = E_{u,t}^*[h(u)\sqrt{t}|h]$ , we obtain the lemma.  $\square$

## 6 Proof of Main Theorems

### 6.1 Proof of Lemma 1

*Proof.* Lemma 1 is already proved in section 4.  $\square$

### 6.2 Proof of Lemma 2

*Proof.* By the definition,  $V = D_3 - D_4$ . By Lemma 5 and 7,  $E[V^{1+\delta}] < \infty$ . Recall that the convergence in law  $\xi_n \rightarrow \xi$  holds. The random variable  $D_1^*(\xi_n) - D_2^*(\xi_n)$  is a continuous function of  $\xi_n$ , hence it converges to a random variable  $D_1^*(\xi) - D_2^*(\xi)$  in law. Therefore, by Lemma 5 and 8,  $D_3 - D_4$  converges to the same random variable in law. Hence  $E[V]$  converges to a constant when  $n$  tends to infinity.  $\square$

### 6.3 Proof of Theorem 1

*Proof.* By the same way as proof of Lemma 2, both  $E[D_1]$  and  $E[D_3]$  converge to  $E[D_1^*(\xi)]$  whereas both  $E[D_2]$  and  $E[D_4]$  converge to  $E[D_2^*(\xi)]$ . From eqs.(6), (7), and (8)

$$\begin{aligned} E[n(G - S)] &\rightarrow \frac{1}{2}E[D_2^*(\xi)], \\ E[n(T - S)] &\rightarrow -2\sigma^2\nu + \frac{1}{2}E[D_2^*(\xi)], \\ E[V] &\rightarrow E[D_1^*(\xi)] - E[D_2^*(\xi)], \end{aligned}$$

where we used the definition of  $\nu$ , that is to say,  $E[D_1^*(\xi) - D_2^*(\xi)] = 2\nu/\beta$ . From Lemma 9,

$$E[D_1^*(\xi)] = 2\sigma^2\nu + \frac{2\lambda}{\beta},$$

resulting that

$$E[D_2^*(\xi)] = 2\sigma^2\nu + \frac{2\lambda - 2\nu}{\beta},$$

which completes the theorem.  $\square$

### 6.4 Proof of Theorem 2

*Proof.* From Theorem 1,

$$\begin{aligned} E[G] &= \frac{N\sigma^2}{2} + \left(\frac{\lambda - \nu}{\beta} + \nu\sigma^2\right)\frac{1}{n} + o_n, \\ E[T] &= \frac{N\sigma^2}{2} + \left(\frac{\lambda - \nu}{\beta} - \nu\sigma^2\right)\frac{1}{n} + o_n, \end{aligned}$$

where  $no_n \rightarrow 0$ . Therefore

$$\begin{aligned} E[G] &= E[T] + \frac{2\nu\sigma^2}{n} + o_n \\ &= E[T] \left(1 + \frac{2\beta E[V]}{Nn}\right) + o_n. \end{aligned}$$

To prove Theorem 2, it is sufficient to show  $E[VT] - E[V]E[T] \rightarrow 0$ .

$$E[|V(T - E[T])|] \leq E[V^2]^{1/2} E[(T - E[T])^2]^{1/2}.$$

Since  $s/4 - 1 \geq 2$ ,

$$0 \leq E[V^2] \leq E[(D_3)^2] < \infty.$$

Let  $S^{(n)} = \frac{1}{n} \sum_{i=1}^n |S_i|^2/2$ ,  $S = \sigma^2 N/2$ . Then

$$E[(T - E[T])^2] \leq 3E[(T - S^{(n)})^2 + (S^{(n)} - S)^2 + (S - E[T])^2].$$

Firstly, from

$$T - S^{(n)} = \frac{E_w[\eta_n(w)]}{\sqrt{n}} + \frac{D_3}{2n^2},$$

we obtain

$$E[(T - S^{(n)})^2] \leq \frac{2E[\|\eta\|^2]}{n} + \frac{E[D_3^2]}{n},$$

which converges to zero. Secondly,  $\{S_i\}$  are independently subject to the normal distribution, hence  $E[(S^{(n)} - S)^2] \rightarrow 0$ . And lastly,

$$T - S = \frac{D_1}{n},$$

hence  $E[(T - S)^2]$  also converges to zero.  $\square$

## 7 Conclusion

In this paper, we proved that singular regression problem is mathematically determined by two birational invariants, the real log canonical threshold and singular fluctuation. Moreover, there is a universal relation between the generalization error and the training error, by which we can estimate two birational invariants from random samples.

## Appendix

To prove eq.(14), we use the following lemmas. Let  $\xi$  and  $\varphi$  are functions of  $C^1$  class from  $[0, b]^d$  to  $\mathbb{R}$ . Assume that  $\varphi(u) > 0$ ,  $u = (x, y) \in [0, b]^d$ . The partition function of  $\xi$ ,  $\varphi$ ,  $n > 1$ , and  $p \geq 0$  is defined by

$$\begin{aligned} Z^p(n, \xi, \varphi) &= \int_{[0, b]^m} dx \int_{[0, b]^{d-m}} dy K(x, y)^p x^h y^{h'} \varphi(x, y) \\ &\quad \times \exp(-n\beta K(x, y)^2 + \sqrt{n}\beta K(x, y) \xi(x, y)). \end{aligned} \quad (15)$$

where  $K(x, y) = x^k y^{k'}$ . Let us use

$$\begin{aligned} \|\xi\| &= \max_{(x, y) \in [0, b]^d} |\xi(x, y)|, \\ \|\nabla \xi\| &= \max_{1 \leq j \leq m} \max_{(x, y) \in [0, b]^d} \left| \frac{\partial \xi}{\partial x_j} \right|. \end{aligned}$$

Without loss of generality, we can assume that four multi-indices  $k, k', h, h'$  satisfy

$$\frac{h_1 + 1}{2k_1} = \dots = \frac{h_r + 1}{2k_m} = \lambda < \frac{h'_j + 1}{2k'_j} \quad (j = m + 1, 2, \dots, d).$$

In this appendix, we define  $a(n, p) \equiv (\log n)^{m-1}/n^{\lambda+p}$ .

**Lemma 10.** *There exist constants  $c_1, c_2 > 0$  such that for arbitrary  $\xi$  and  $\varphi$  ( $\varphi(x) > 0 \in [0, b]^d$ ) and an arbitrary natural number  $n > 1$ ,*

$$c_1 a(n, p) e^{-\beta \|\xi\|^2/2} \min(\varphi) \leq Z^p(n, \xi, \varphi) \leq c_2 a(n, p) e^{\beta \|\xi\|^2/2} \|\varphi\|$$

*holds, where  $\min(\varphi) = \min_{u \in [0, b]^d} \varphi(u)$ .*

Let  $\xi$  and  $\varphi$  be functions of class  $C^1$ . We define

$$Y^p(n, \xi, \varphi) \equiv \gamma a(n, p) \int_0^\infty dt \int_{[0, b]^s} dy t^{\lambda+p-1} y^\mu e^{-\beta t + \beta \sqrt{t} \xi_0(y)} \varphi_0(y),$$

where we use notations,  $\gamma = b^{|h|+m-2|k|\lambda} / (2^m(m-1)! \prod_{j=m+1}^d k_j)$ ,  $\xi_0(y) = \xi(0, y)$ ,  $\varphi_0(y) = \varphi(0, y)$ ,  $\mu = h' - 2\lambda k'$ . A measure  $D(du)$  on  $\mathbb{R}^d$  is defined by

$$D(du) = \gamma \delta(x) y^\mu. \quad (16)$$

**Lemma 11.** *There exists a constant  $c_3 > 0$  such that, for arbitrary  $n > 1$ ,  $\xi$ ,  $\varphi$ , and  $p \geq 0$ ,*

$$\begin{aligned} & |Z^p(n, \xi, \varphi) - Y^p(n, \xi, \varphi)| \\ & \leq \frac{c_1 a(n, p)}{\log n} e^{\beta \|\xi\|^2/2} \{\beta \|\nabla \xi\| \|\varphi\| + \|\nabla \varphi\| + \|\varphi\|\}. \end{aligned}$$

*Moreover, there exist constant  $c_4, c_5 > 0$  such that, for arbitrary  $\xi$ ,  $\varphi$ ,  $n > 1$ ,*

$$c_4 a(n, p) e^{-\beta \|\xi\|^2/2} \min(\varphi) \leq Y^p(n, \xi, \varphi) \leq c_5 a(n, p) e^{\beta \|\xi\|^2/2} \|\varphi\|.$$

*Proof.* Lemmas 10 and 11 are proved by direct but rather complicated calculations [17, 18]. Let us introduce the outline of the proof. Let  $F_p(x, y)$  be the integrated function in eq.(15) and  $Z^p = Z^p(n, \xi, \phi)$ .

$$Z^p = \int dx \int dy F_p(x, y),$$

which is equal to

$$Z^p = \int_0^\infty dt \int_{[0, b]^d} dx dy \delta(t - K(x, y)^2) F_p(x, y). \quad (17)$$

Therefore, the problem results in  $\delta(t - K(x, y)^2)$ . For arbitrary function  $\Psi(x, y)$  of class  $C^\infty$ , the function

$$\zeta(z) = \int_{[0, b]^d} K(x, y)^{2z} \Psi(x, y) dx dy$$

is the meromorphic function whose poles are  $(-\lambda_j)$  and its order  $m_j$ , hence it has Laurent expansion,

$$\zeta(z) = \zeta_0(z) + \sum_{j=1}^{\infty} \frac{c_j(\Psi)}{(z + \lambda_j)^{m_j}},$$

where  $\zeta_0(z)$  is a holomorphic function and  $c_j(\Psi)$  is a Schwartz distribution. Since  $\int \delta(t - K(x, y)^2) \Psi(x, y) dx dy$  is the Mellin transform of  $\zeta(z)$ , we have an asymptotic expansion of  $\delta(t - K(x, y)^2)$  for  $t \rightarrow +0$ ,

$$\delta(t - K(x, y)^2) = \sum_{j=1}^{\infty} \sum_{m=1}^{m_j} t^{\lambda_j-1} (-\log t)^{m-1} c_{jm}(x, y),$$

where  $c_{jm}(x, y)$  is a Schwartz distribution. By applying this expansion to eq.(17), we obtain two lemmas.  $\square$

## References

- [1] H. Akaike. A new look at the statistical model identification. *IEEE Trans. on Automatic Control*, Vol.19, pp.716-723, 1974.
- [2] S. Amari, N. Murata. Statistical theory of learning curves under entropic loss. *Neural Computation*, Vol. 5, pp.140-153, 1993.
- [3] M.Aoyagi, S.Watanabe. Stochastic complexities of reduced rank regression in Bayesian estimation. *Neural Networks*, Vol.18, No.7, pp.924-933, 2005.
- [4] M.F. Atiyah. Resolution of singularities and division of distributions. *Communications of Pure and Applied Mathematics*, Vol.13, pp.145-150. 1970.
- [5] I.N. Bernstein. The analytic continuation of generalized functions with respect to a parameter. *Functional Analysis and Applications*, Vol.6, pp.26-40, 1972.
- [6] I.M. Gelfand and G.E. Shilov. *Generalized Functions*. Academic Press, San Diego, 1964.
- [7] H. Hironaka. Resolution of singularities of an algebraic variety over a field of characteristic zero. *Annals of Mathematics*, Vol.79, pp.109-326, 1964.
- [8] M. Kashiwara. B-functions and holonomic systems. *Inventiones Mathematicae*, Vol. 38, pp.33-53, 1976.
- [9] M. Mustata. Singularities of pairs via jet schemes. *Journal of the American Mathematical Society*, Vol.15, pp.599-615. 2002.
- [10] T. Oaku. Algorithms for b-functions, restrictions, and algebraic local cohomology groups of D-modules. *Advances in Applied Mathematics*, Vol.19, pp.61-105, 1997.
- [11] M. Saito. On real log canonical thresholds, *arXiv:0707.2308v1*, 2007.
- [12] G. Schwarz. Estimating the dimension of a model. *Annals of Statistics*, Vol.6, No.2, pp.461-464. 1978.

- [13] A. W. van der Vaart, J. A. Wellner. *Weak Convergence and Empirical Processes*. Springer, 1996.
- [14] S. Watanabe. Equations of States in Singular Statistical Estimation. *arXiv:0712.0653*, 2007.
- [15] S. Watanabe. A formula of equations of states in singular learning machines. *Proc. of IEEE World Congress in Computational Intelligence*, 2008.
- [16] S. Watanabe, Algebraic analysis for nonidentifiable learning machines, *Neural Computation*, **13(4)** (2001) 899–933.
- [17] S. Watanabe, *Algebraic geometry and learning theory*, Morikita Publishing, Tokyo, 2006.
- [18] S. Watanabe, *Algebraic geometry and statistical learning theory*, Cambridge University Press, Cambridge, 2009.