Advanced Studies in Pure Mathematics 57, 2010
Probabilistic Approach to Geometry
pp. 473–492

A limit theorem in singular regression problem

Sumio Watanabe

Abstract.

In statistical problems, a set of parameterized probability distributions is often used to estimate the true probability distribution. If the 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 accuracy of singular regression 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,

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

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 $\{(X_i, Y_i); i = 1, 2, ..., n\}$ is a set of random variables which are independently subject to the same probability distribution as (X, Y).

Received January 15, 2009.
Revised July 7, 2009.
2000 *Mathematics Subject Classification.* 60D05.
Key words and phrases. Singular regression, real log canonical threshold, singular fluctuation, resolution of singularities, generalization error.

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 $E_w[\ ]$ on W is defined by

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

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

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

where $E_X E_Y[\ ]$ shows the expectation value over the random variable (X,Y). 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 $n \to \infty$. If the Fisher information matrix

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

where $\partial_i = (\partial/\partial w_i)$ and $\cdot$ shows the inner product of $\mathbb{R}^N$, 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 $n(E[G] - S) \to d\sigma^2/2$ and $n(E[T] - S) \to -d\sigma^2/2$ 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 the 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 $\{w \in W; r(x, w) = r_0(x)\}$ consists of not one point but of an analytic set, and the 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. Let $s \geq 4$ be a natural number which is equal to 4 times of some integer. Each theorem or lemma in this paper depends on the natural number s.

Basic assumptions.
(A1) The set of parameters W is defined by

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

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^∞.

(A2) 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 $|\ |_s$,

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

(A3) There exists a parameter w_0 in the open kernel of W such that $r(x,w_0) = r_0(x)$.

If these basic assumptions are satisfied, then

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

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.

Remarks. (1) If $s' > s$, then by the Hölder inequality, $|f|_{s'} \geq |f|_s$. Hence if the assumption (A2) with s' holds, then (A2) with s also holds. (2) The set of the assumptions (A1), (A2), and (A3) is a sufficient condition for Theorems. It is an important future study to generalize assumptions.

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 $Re(z) > 0$ 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*

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

satisfies

$$\lim_{n\to\infty} E[V] = \frac{2\nu}{\beta}. \tag{3}$$

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]. The value λ is also important in statistical learning theory, which 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*

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

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

This theorem shows that both the real log canonical threshold λ and singular fluctuation ν determine the singular regression problem.

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

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

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

This theorem shows the following fact. The values V and T can be calculated from random samples $(X_1, Y_1), \cdots, (X_n, Y_n)$ and the statistical model $r(x, w)$ without any direct knowledge of the true regression function $r_0(x)$. The generalization error $E[G]$ can be estimated from T and V, 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. In other words, we can use Theorem 2 without checking that the true distribution is a singularity or not. Note that Theorem 2 holds even if the true distribution is not contained in the statistical models [19].

§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

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

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 $H_0(w)$,

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

Moreover, $H_0(w)$ 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 $\| \ \|$ 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 uniformly tight random processes. Let $\eta^{(1)}$ and $\eta^{(2)}$ be respectively the tight gaussian processes which have the same expectations and the same covariance matrices as $\eta_n^{(1)}$ and $\eta_n^{(2)}$. It is well known in empirical process theory that $\eta_n^{(1)}$ and $\eta_n^{(2)}$ weakly converge to $\eta^{(1)}$ and $\eta^{(2)}$ respectively, as $n \to \infty$ [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 $r = (r_1, .., r_d)$ 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. Q.E.D.

Lemma 4. *Assume (A1), (A2), and (A3) with $s \geq 4$. For an 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. Since $\{S_i; i = 1, 2, \cdots, n\}$ are independently subject to the normal distribution with the average 0 and the covariance matrix $\sigma^2 I$, it follows that

$$E[S_i \cdot F(S_i)] = E[\nabla_{S_i} \cdot F(S_i)],$$

where $F(x)$ is a function of $x \in \mathbb{R}^N$ which satisfies

$$\lim_{|x|\to\infty} e^{-|x|^2/2}|F(x)| = 0,$$
$$\int e^{-|x|^2/2}|\nabla F(x)|dx < \infty,$$

and ∇_{S_i} is defined by

$$\nabla_{S_i} \cdot F(S_i) = \sum_{j=1}^{N} \frac{\partial F}{\partial x_j}(S_i).$$

Let the left hand side of the first equation be A. Because $H_0(w)$ is a function of S_i, it follows that,

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

which is equal to the right hand side of the first equation. Q.E.D.

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

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

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

$$E[V] = E[D_3 - D_4], \tag{8}$$

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

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*

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

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

Lemma 5. *Assume (A1), (A2), and (A3) with $s \ge 8$. Let $0 < \delta < s/4 - 1$. For arbitrary $\epsilon > 0$, $j = 1, 2, 3, 4, 5$,*

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

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

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

where we used $2\sqrt{n}\|\eta_n\| \leq (n\epsilon/2 + (2/\epsilon)\|\eta_n\|^2)$ and a lower bound,

$$\int_{K(w)\leq\epsilon} \exp(-\beta n K(w))\varphi(w)dw \geq \frac{1}{C_1 n^\lambda}$$

with a constant $C_1 > 0$ [16]. 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. Q.E.D.

§4. Resolution of singularities

To study the expectation on the region W_ϵ 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 $K(w)\prod_{j=1}^J \pi_j(w)\varphi_1(w)$, 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_ϵ, a manifold U^*, and a proper analytic map $g : U^* \to W_\epsilon^*$ 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 $k = (k_1, ..., k_d)$ and $h = (h_1, ..., h_d)$ are multi-indices (k_k and h_l are nonnegative integers) and

$$\begin{aligned} u^{2k} &= \prod_{j=1}^d u_j^{2k_j}, \\ u^h &= \prod_{j=1}^d u_j^{h_j}. \end{aligned}$$

Here $|g(u)'|$ is the absolute value of Jacobian determinant of $w = g(u)$, and $\phi(u) > 0$ is a function of class C^∞. Let $U = g^{-1}(W_\epsilon)$. Since g is a proper map and W_ϵ is compact, U is also compact. Moreover, it is covered by a finite sum

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

where each U_α 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^∞. 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. *Assume (A1), (A2), and (A3) with $s \geq 4$. The largest pole $(-\lambda)$ and its order m of $\zeta(z)$ are given by*

$$\lambda = \min_\alpha \min_j \Big(\frac{h_j + 1}{2k_j}\Big), \tag{11}$$

$$m = \max_\alpha \#\Big\{j; \lambda = \frac{h_j + 1}{2k_j}\Big\}, \tag{12}$$

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. Although the multi-indices k and h depend on the choise of resolution, the constants λ and m do not depend on the pair (U^*, g). Such values are called birational invariants. By the definition of $K(w)$ in eq.(2),

$$K(g(u)) = \frac{1}{2}\int |f(x, g(u))|^2 q(x)dx. \tag{13}$$

Because $K(g(u)) = u^{2k}$ and $f(x, g(u))$ is an $L^s(q)$-valued analytic function with $s \geq 4$, 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

$$\xi_n(u) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} S_i \cdot a(X_i,u) + \frac{1}{\sqrt{n}}\sum_{i=1}^{n} u^k\Big(1 - \frac{a(X_i,u)^2}{2}\Big). \tag{14}$$

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 $\|\nabla\xi_n\| = \max_j \sup_u |\partial_j \xi_n(u)|$. The function $\xi_n(u)$ is defined on each local coordinate in U, which is an empirical process on each local coordinate. By the same way as Lemma 3, it converges in law to the tight gaussian process which is denoted by $\xi(u)$. The expectation operator $E_u[\ ]$ on U is defined so that it satisfies $E_w^\epsilon[F(w)] = E_u[F(g(u))]$. Then

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

Lemma 7. *Assume (A1), (A2), and (A3) with $s \geq 12$ 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 $l = 1,2,..,d$, 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

$$E_u[u^{2k}] \leq \frac{c_1}{n}\{1 + \|\xi_n\|^2 + \|\nabla\xi_n\|^2\}. \tag{15}$$

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

$$\begin{aligned}
&E[(D_3^\epsilon)^{1+\delta}] \leq E[(c_1L(1 + \|\xi_n\|^2 + \|\nabla\xi_n\|^2))^{1+\delta}]\\
&\leq E[(c_1L)^{3+3\delta}]^{1/3}E[(1 + \|\xi_n\|^2 + \|\nabla\xi_n\|^2)^{(3+3\delta)/2}]^{3/2}.
\end{aligned}$$

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

§5. Renormalized distribution

Definition 5.1. *For a given function $h(u)$ on U, the renormalized expectation operator $E^*_{u,t}[\ |h]$ 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.(20) 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. *Assume (A1), (A2), and (A3) with $s \geq 12$. The following convergences in probability hold.*

$$\begin{aligned} D^\epsilon_1 - D^*_1(\xi_n) &\to 0, \\ D^\epsilon_2 - D^*_2(\xi_n) &\to 0, \\ D^\epsilon_3 - D^*_1(\xi_n) &\to 0, \\ D^\epsilon_4 - D^*_2(\xi_n) &\to 0, \\ D^\epsilon_5 - D^*_5(\xi_n) &\to 0. \end{aligned}$$

Proof. These five convergences can be proved by the same way. We show $D^\epsilon_3 - D^*_1(\xi_n) \to 0$. Let $L(u) = (1/n)\sum_{i=1}^n |a(X_i,u)|^2$. Since $E_X[|a(X,u)|^2] = 2$,

$$\begin{aligned} |D^\epsilon_3 - D^*_1(\xi_n)| \leq\ & |E_u[nL(u)u^{2k}] - E_u[E_X[a(X,u)^2]u^{2k}]| \\ & + |E_u[2u^{2k}] - E^*_{u,t}[2t|\xi_n]|. \end{aligned}$$

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 $\|L - a(X)\| \to 0$ and eq.(15), D_6 converges to zero in probability. From Lemma 10 and 11 in Appendix, it is derived that

$$|E_u[u^{2k}] - E^*_{u,t}[t|\xi_n]| \leq \frac{c_1}{\log n} \frac{e^{2\beta\|\xi_n\|^2}}{\min(\phi)^2} \{1 + \beta\|\nabla\xi_n\|\}, \tag{16}$$

which shows $D_7 \to 0$ in probability. Q.E.D.

Lemma 9. *Assume (A1), (A2), and (A3) with $s \geq 12$. 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. Q.E.D.

§6. Proof of Main Theorems

6.1. Proof of Lemma 1

Proof. Lemma 1 is already proved in section 4. Q.E.D.

6.2. Proof of Lemma 2

Proof. By the definition, $V = D_3 - D_4$. By Lemma 5 and 7, $E[V^{1+\delta}] < \infty$. Reall that the convergence in law $\xi_n \to \xi$ holds. The random variable $D_1^*(\xi_n) - D_2^*(\xi_n)$ is a continuous function of ξ_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. Q.E.D.

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)

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

where we used the definition of ν, 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. Q.E.D.

6.4. Proof of Theorem 2

Proof. From Theorem 1,

$$E[G] = \frac{N\sigma^2}{2} + \left(\frac{\lambda-\nu}{\beta} + \nu\sigma^2\right)\frac{1}{n} + o_n, \tag{17}$$

$$E[T] = \frac{N\sigma^2}{2} + \left(\frac{\lambda-\nu}{\beta} - \nu\sigma^2\right)\frac{1}{n} + o_n, \tag{18}$$

where $no_n \to 0$. 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 $E[VT] - E[V]E[T] \to 0$.

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

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

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

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

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

which converges to zero. Secondly, $\{S_i\}$ are independently subject to the normal distribution, hence $E[(S^{(n)} - S)^2] \to 0$. And lastly, by using eq.(18), $E[(E[T]-S)^2] = (E[T]-S)^2$ converges to zero, which completes the proof. Q.E.D.

§7. Discussion

In this paper, we proved a limit theorem in a singular regression problem. In general statistical estimation problems, a true probability density function $q(x)$ on $\mathbb{R}^N$ is estimated by a parametric density function $p(x|w)$, where $w \in \mathbb{R}^d$. Also in such problems, there are mathematical issues caused by the singular Fisher information matrix. There are some mathematical results for general statistical estimation problems [14, 18].

From the statistical point of view, the regression problem might be understood as a special one contained in the general statistical estimation problems. In fact, in a regression problem, it seems that the true probability density function $q(x, y)$ is estimated by using the parametric probability density function,

$$p(x, y|w) = \frac{q(x)}{(2\pi\sigma^2)^{N/2}} \exp\Big(-\frac{|y - r(x, w)|^2}{2\sigma^2}\Big).$$

However, there are some mathematical differences between the general statistical estimation problems and a regression problem.

Firstly, in the regression problem of this paper, the standard deviation σ of the true distribution is not directly estimated. Strictly speaking, the Kullback–Leibler distance between $q(x, y)$ and the Bayes predictive distribution $E_w[p(x, y|w)]$ is not equal to the generalization error defined by the square error G in this paper. The formula in Theorem 2 are given only in the regression problem.

Secondly, in general statistical estimation problems, $K(w)$ is not equal to the square error of the log density ratio function, hence eq.(13)

does not hold, resulting that we need the different assumptions and proofs to obtain theorems. Therefore the results of the general statstical estimation does not contain the results of this paper mathematically.

And lastly, in general statistical estimation problems, Lemma 4 does not hold, because, in the proof of this lemma, we used the fact that $\{S_i\}$ are subject to the normal distribution. In general statistical estimation problems, we need the different assumptions and proofs in stead of this lemma.

By these reasons, although a regression problem has a strong relation to the general statistical estimation problems, the former is not contained in the latter. In this paper, we mainly studied a regression problem, and proved a limit theorem in a regression problem.

§8. Conclusion

In this paper, we proved that a singular regression problem is mathematically determined by two birational invariants, the real log canonical threshold and the 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

In the proof of eq.(16), we used the following lemmas.

Let ξ and φ 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 ξ, φ, $n > 1$, and $p \geq 0$ is defined by

$$\begin{aligned} Z^p(n,\xi,\varphi) &= \int_{[0,b]^m} dx \int_{[0,b]^{d-m}} dy\ K(x,y)^p\ x^h y^{h'}\ \varphi(x,y) \\ &\quad \times \exp(-n\beta\ K(x,y)^2 + \sqrt{n}\beta\ K(x,y)\ \xi(x,y)). \end{aligned} \tag{19}$$

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

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

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 ξ and φ ($\varphi(x) > 0 \in [0,b]^d$) 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 ξ and φ be functions of class C^1. We define

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

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

$$D(du) = \gamma\delta(x) y^\mu dy. \tag{20}$$

Lemma 11. *There exists a constant $c_3 > 0$ such that, for arbitrary $n > 1$, ξ, φ, 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 ξ, φ, $n > 1$,

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

Proof. Lemmas 10 and 11 are proved by direct but rather complicated calculations. They are shown by applying Theorems 28 and 29 in [17] or directly in Theorems 4.8 and 4.9 in [18].

Let us introduce the outline of the proof. Let $F_p(x,y)$ be the integrated function in eq.(19) and $Z^p = Z^p(n,\xi,\phi)$.

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

which is equal to

$$Z^p = \int_0^\infty dt \int_{[0,b]^d} dx\, dy\, \delta(t - K(x,y)^2)\, F_p(x,y). \tag{21}$$

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

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

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 $\int \delta(t-K(x,y)^2)\Psi(x,y)dxdy$ is the Mellin transform of $\zeta(z)$, we have an asymptotic expansion of $\delta(t-K(x,y)^2)$ for $t\to+0$,

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

where $c_{jm}(x,y)$ is a Schwartz distribution. By applying this expansion to eq.(21), we obtain two lemmas. Q.E.D.

References

[1] H. Akaike, A new look at the statistical model identification, IEEE Trans. Automatic Control, **19** (1974), 716–723.

[2] S. Amari and N. Murata, Statistical theory of learning curves under entropic loss, Neural Computation, **5** (1993), 140–153.

[3] M. Aoyagi and S.Watanabe, Stochastic complexities of reduced rank regression in Bayesian estimation, Neural Networks, **18** (2005), 924–933.

[4] M. F. Atiyah, Resolution of singularities and division of distributions, Comm. Pure Appl. Math., **13** (1970), 145–150.

[5] I. N. Bernstein, The analytic continuation of generalized functions with respect to a parameter, Funct. Anal. Appl., **6** (1972), 26–40.

[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, Ann. of Math. (2), **79** (1964), 109–326.

[8] M. Kashiwara, B-functions and holonomic systems, Invent. Math., **38** (1976), 33–53.

[9] M. Mustata, Singularities of pairs via jet schemes, J. Amer. Math. Soc., **15** (2002), 599–615.

[10] T. Oaku, Algorithms for b-functions, restrictions, and algebraic local cohomology groups of D-modules, Adv. Appl. Math., **19** (1997), 61–105.

[11] M. Saito, On real log canonical thresholds, arXiv:0707.2308v1.
[12] G. Schwarz, Estimating the dimension of a model, Ann. Statist., **6** (1978), 461–464.
[13] A. W. van der Vaart and J. A. Wellner, Weak Convergence and Empirical Processes, Springer-Verlag, 1996.
[14] S. Watanabe, Equations of states in singular statistical estimation, Neural Networks, **23** (2010), 35–43.
[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** (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 Univ. Press, Cambridge, 2009.
[19] S. Watanabe, Equations of states in statistical learning for an unrealizable and regular case, IEICE Transactions, to appear.

Precision and Intelligence Laboratory
Tokyo Institute of Technology
4259 Nagatsuta, Midoriku
Yokohama, 226-8503
Japan

E-mail address: swatanab@pi.titech.ac.jp