Algebraic Geometrical Method in Singular Statistical Estimation

Sumio Watanabe
Tokyo Institute of Technology

Contents

1

Singular Models

Statistical Estimation

Unknown Information Source (x in R^N)

Random Samples

$$q(x) \longrightarrow X^n = (X_1, X_2, \ldots, X_n)$$

Statistical Model $p(x|w)$ (w in W)

A priori distribution $\varphi(w)$

Regular and Singular

Definition. If the map $w \longmapsto p(x|w)$ is one-to-one, and Fisher information matrix is positive definite, p(x|w) is called regular, if otherwise singular.

Regular Models

Normal distribution, Binomial distribution,

Polynomial regression, …

Singular Models

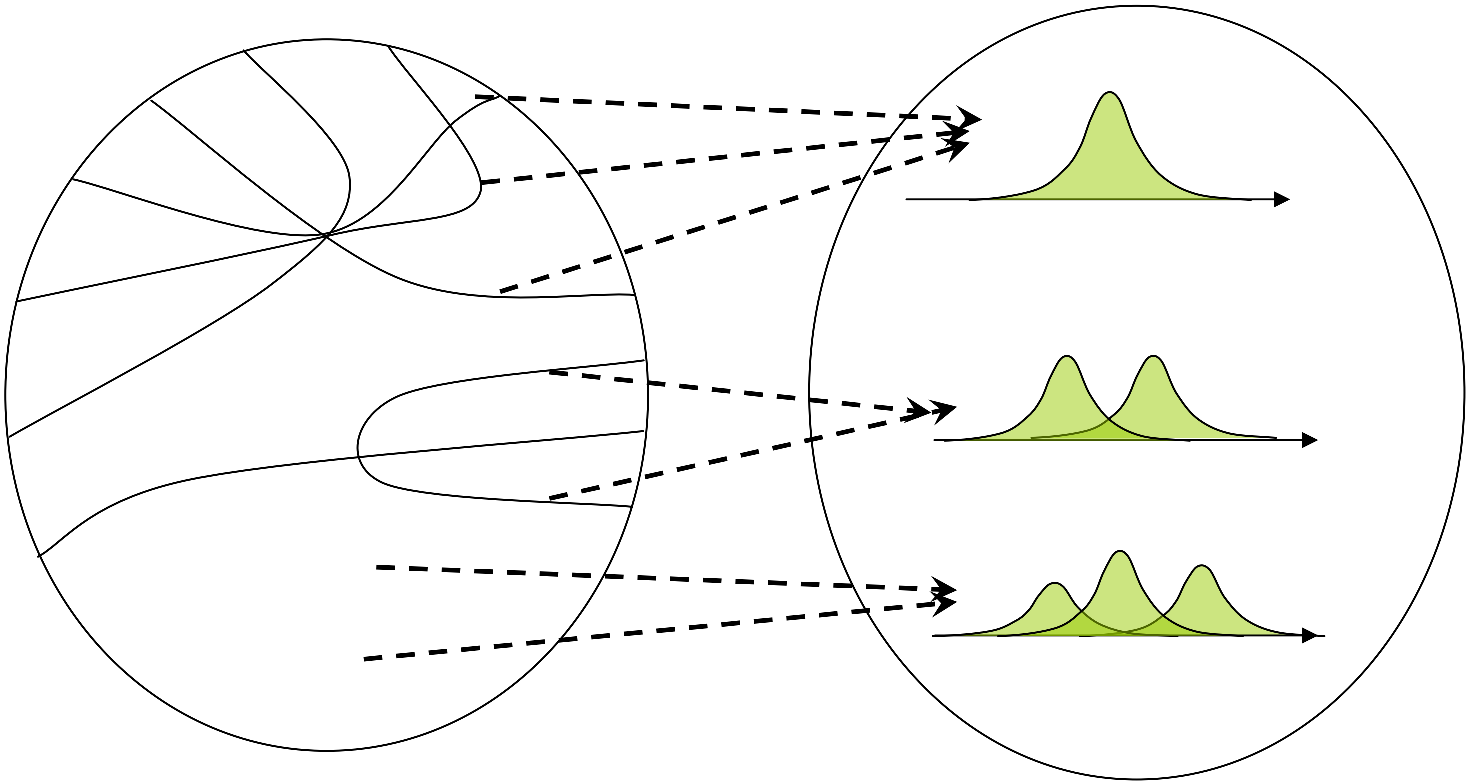
Normal mixture, Binomial mixture, Neural network,

Reduced rank regression, Hidden Markov model,

Kalman filter, Stochastic context-free grammar, …

Singular Statistical Models

$$w \longmapsto p(x|w)$$



Set of parameters

Set of probability distributions

Definition. A posteriori distribution ($0<\beta<\infty$)

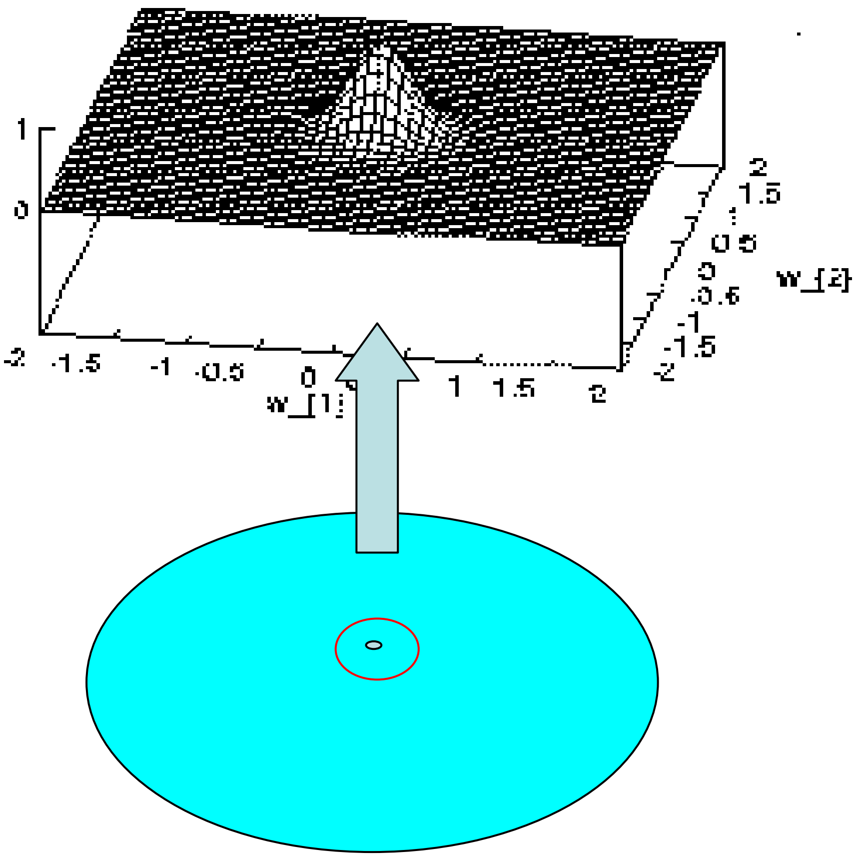
$$p(w|X^n) = \frac{\prod_{i=1}^{n} p(X_i|w)^{\beta}\, \varphi(w)}{Z}$$

Expectation value by $p(w|X^n)$ is denoted by

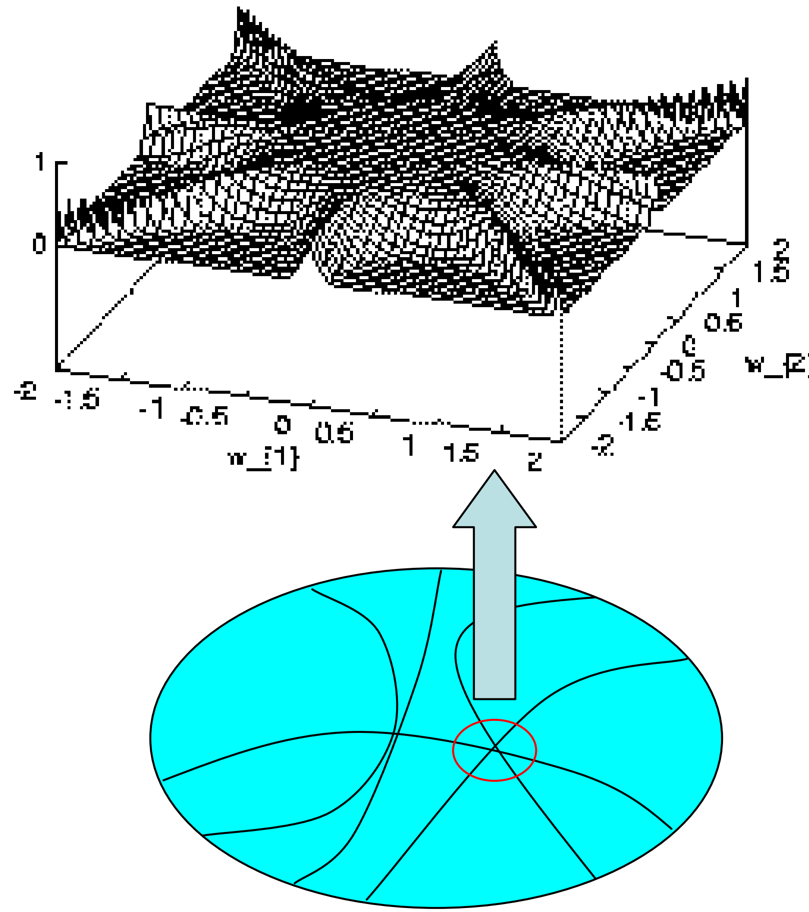
$$E_w[\quad] = \int [\qquad]\, p(w|X^n)\, dw$$

A Posteriori Distribution

regular model



singular model



Important random variables I

Stochastic complexity or marginal likelihood

$$F = -\log \int \prod_{i=1}^{n} p(X_i|w)^{\beta} \varphi(w) dw$$

(1) F is the minus likelihood of $p(x|w)$ and $\varphi(w)$.

(2) If F is smaller, (p,φ) is more appropriate for data.

(3) F is often used for hyperparameter optimization.

Important random variables II

Bayes generalization error

$$B_g = E_X \log \left(\frac{q(X)}{E_w\, p(X|w)} \right)$$

(1) B_g is a function of $p(x|w)$ and $\varphi(w)$ for given data.

(2) If B_g is smaller, prediction by (p,φ) is more precise.

(3) However, B_g can not be directly calculated from data.

Important random variables III

Bayes Training error

$$B_t = \left(\frac{1}{n} \sum_{i=1}^{n} \log \frac{q(X_i)}{E_w p(X_i|w)} \right)$$

(1) B_t is a function of $p(x|w)$ and $\varphi(w)$ for given data.

(2) B_t can be calculated by data.

BIC and MDL

If a model is regular and $0 < \beta < \infty$,

$$E[\,F\,] = \beta n S + (d/2) \log n + O(1),$$

where S : entropy of true distribution
d : dimension of parameter space

Ref: Schwarz(1978), Rissanen(1984).

AIC

If a model is regular and $0 < \beta \leq \infty$,

$$E[\,B_g\,] = E[\,B_t\,] + d/n + o(1/n),$$

where d : dimension of parameter space

Ref: Akaike(1974)

In singular models

If a model is singular,

$$E[\,F\,] \neq \beta n S + (d/2)\log n + O(1),$$

$$E[\,B_g\,] \neq E[\,B_t\,] + d/n + o(1/n),$$

Singular theory is necessary.

Mathematics : Limit Theorem

1. Limit theorem of random variable,

$$Z = \int p(X_1|w)p(X_2|w) \dots p(X_n|w)\, \varphi(w)dw$$

2. Limit theorem of probability distribution

$$E_w[\quad] = \frac{\int [\quad]\, p(X_1|w)p(X_2|w) \dots p(X_n|w)\, \varphi(w)dw}{\int p(X_1|w)p(X_2|w) \dots p(X_n|w)\, \varphi(w)dw}$$

Central limit theorem --- expectation and variance

Singular learning theory --- RLCT and SF.

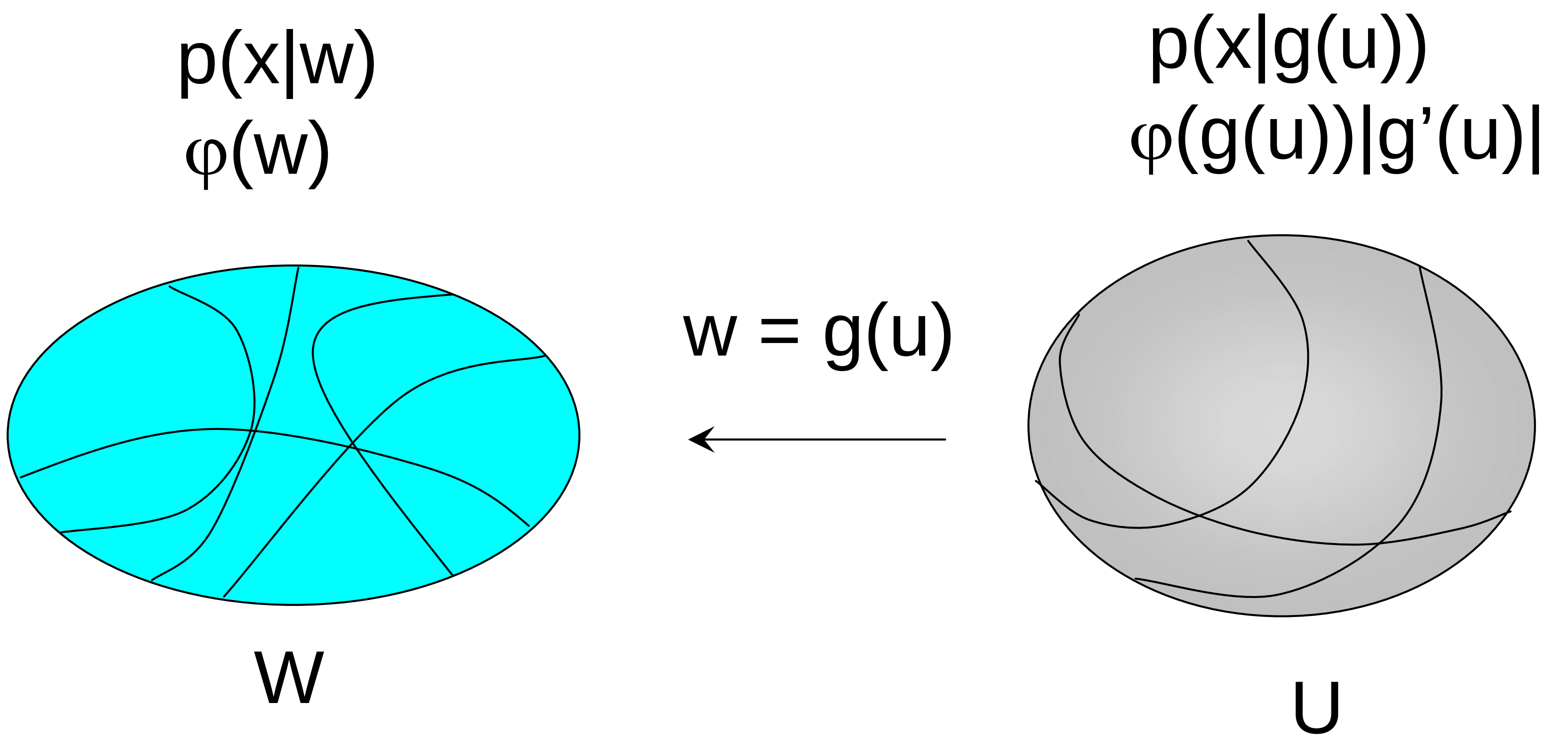
2 Birational Invariants I

Real log canonical threshold

Birational Map

A birational map gives a new parameter space.

Example: Blow-up, Toric Modification, …



Important functions

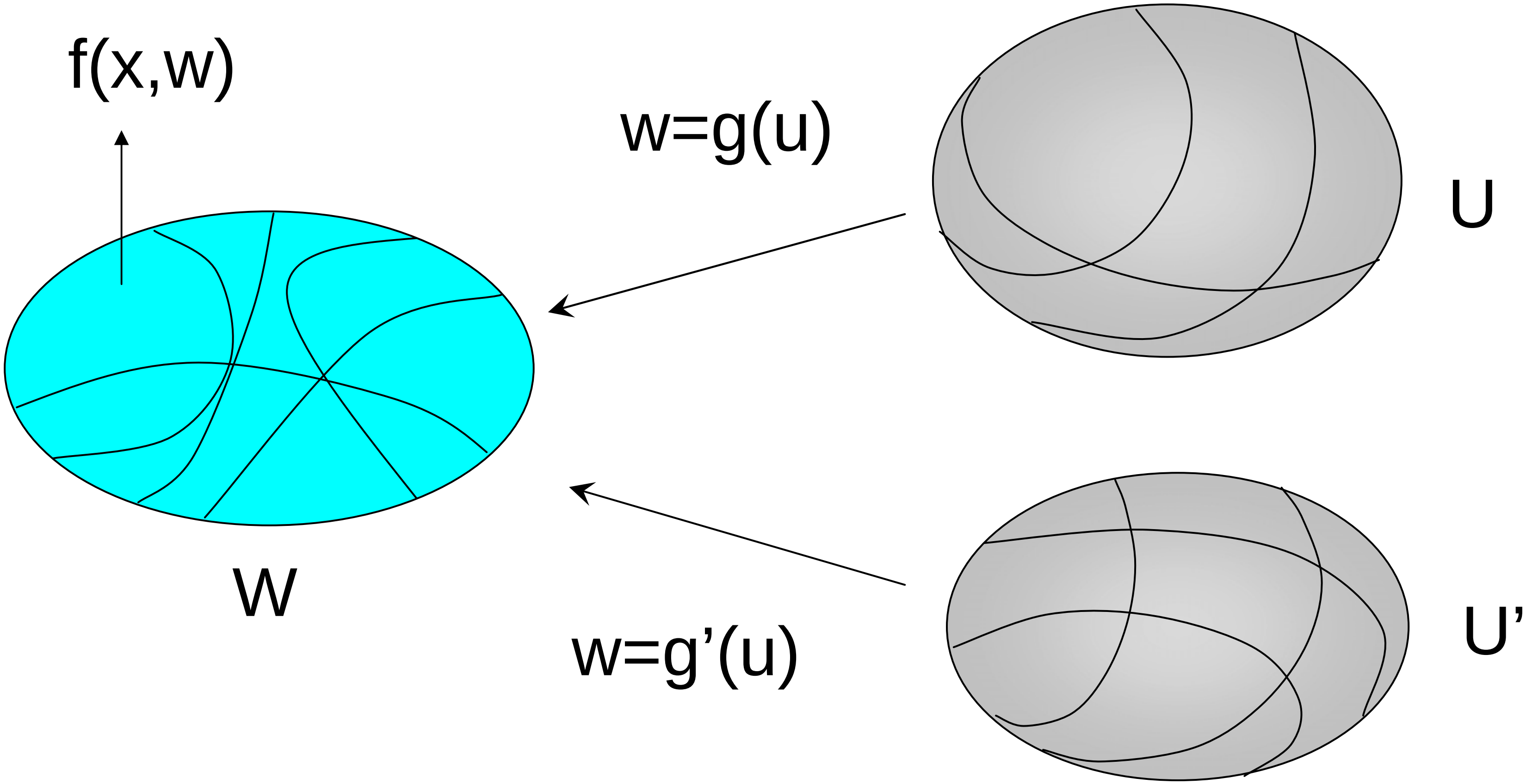
(1) Log density ratio $f(x,w) = \log (q(x)/p(x|w))$

(2) Kullback-Leibler $K(w) = \int q(x) f(x,w)\, dx.$

(3) Log likelihood ratio $K_n(w) = \dfrac{1}{n} \sum_{i=1}^{n} f(X_i,w)$

These functions are defined on W, which are also defined on U, by w=g(u).

Birational Invariant



Statistical theorem should be birational invariant.

Resolution theorem (Hironaka, 1964)

R

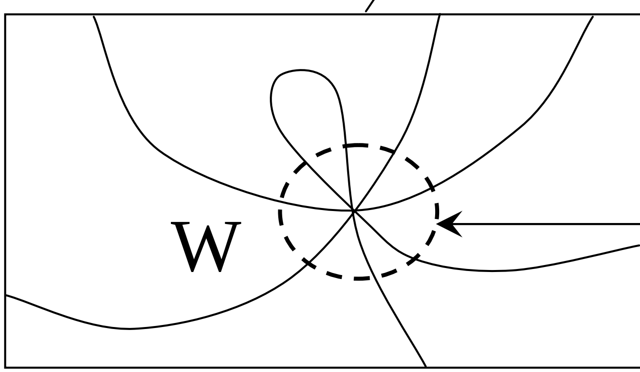
Kullback
-Leibler

$K(w) \geq 0$

In every local coordinate α,

$$K(g(u)) = u_1^{2k_1} u_2^{2k_2} \cdots$$

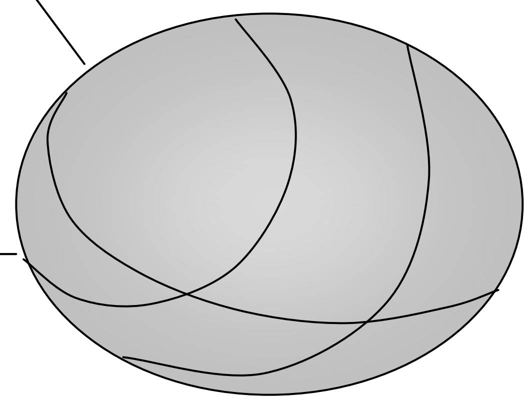
$$\varphi(u)\,|g'(u)| = b(u)\, u_1^{h_1} u_2^{h_2} \cdots u_d^{h_d}$$



R^d

$w=g(u)$

Proper
Analytic
map



Manifold U

Real Log Canonical Threshold

In a local coordinate α,

$$K(g(u)) = u_1^{2k_1} u_2^{2k_2} \cdot \ \cdot \ u_d^{2k_d}$$

$$\varphi(u)\ |g'(u)| = b(u)\ u_1^{h_1} u_2^{h_2} \cdot \ \cdot \ u_d^{h_d}$$

Definition. RLCT $\lambda = \min_{\alpha} \ \min_{j=1,2,\ldots,d} (h_j+1)/2k_j$

Order $m = \max_{\alpha} \sharp\{\, j ; \lambda = (h_j+1)/2k_j \,\}$

α^* : local coordinate s.t. λ and m are attained.

Essential coordinate

(Cf. LCT : *Mori, Mustata, Saitoh, …   Algebraic Geometers*)

Zeta function in Statistics

RLCT is a birational invariant because ($-\lambda$) is equal to the largest pole of the zeta function.

Zeta function

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

$$= \sum_\alpha \int K(g(u))^z \varphi(g(u)) |g'(u)| du$$

$$= \sum_\alpha \int \Pi u_j^{2k_j z+h_j} b(u) du$$

$\rightarrow$ Laurent expansion of $\zeta(z) = \Sigma \frac{C_{km}}{(z+\lambda_k)^{m_k}}$.

BIC is generalized

Theorem.1. (1999, Watanabe)

Assume a model is regular or singular, $0 < \beta < \infty$.

Let λ and m be RLCT and its order of Kullback-Leibler information K(w). Then

$$E[\,F\,] = \beta nS + \lambda \log n - (m-1)\log\log n + O(1).$$

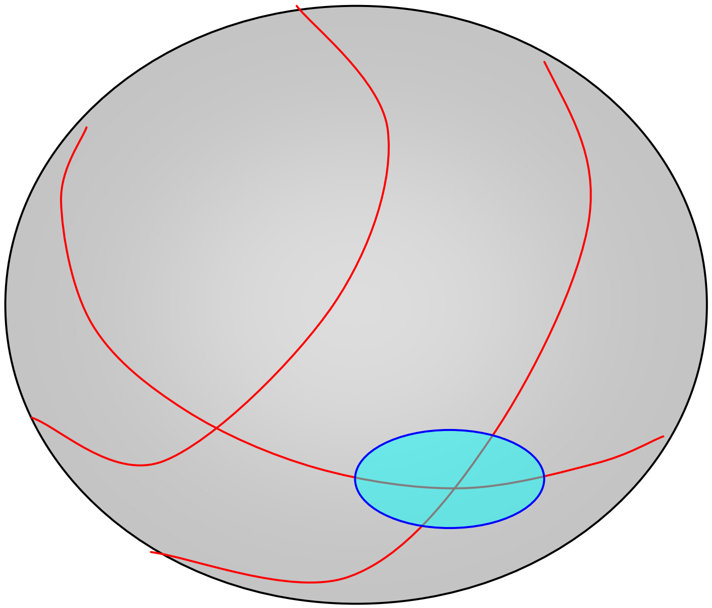
Remark. For a given K(w), several methods to calculate λ are studied in algebraic geometry, commutative ring theory, and algebraic analysis.

3

Birational Invariant II

Singular Fluctuation

Decomposition of log likelihood ratio



U Manifold

By using Hironaka theorem,

in every local coordinate,

$$K(g(u)) = \prod_{j=1}^{d} u_j^{2k_j}$$

$$f(x,g(u)) = a(x,u) \prod_{j=1}^{d} u_j^{k_j}$$

$$\Longrightarrow \quad E_X[a(X,u)] = \prod u_j^{k_j}$$

Empirical Process

We define a random process on U,

$$\xi_n(u) = \frac{1}{n^{1/2}} \sum_{i=1}^{n} \left(a(X_i,u) - \prod_{j=1}^{d} u_j^{\,k_j} \right)$$

$C^0(U)$: Set of continuous functions on compact U

$C^0(U)$ is separable and complete metric space with

$$||f|| = \max_{u \text{ in } U} |f(u)|$$

Lemma.1. $\xi_n \to \xi$: convergence in law in $C^0(U)$

ξ is a unique gaussian random process whose average is zero and covariance is $E_x[a(X,u)a(X,v)]-u^k v^k$.

Remark. Empirical process theory is the central limit theorem in Banach space.

Def. $E_\xi[\ \ ]$: expectation over gaussian process ξ

Decomposition of State Density

Lemma 2. There exists a measure $D(u)du$ such that

$$\delta(t-K(g(u)))\ \varphi(g(u))\ |g'(u)|\ du$$
$$= \sum_{\alpha^*} D(u)du\ t^{\lambda-1}(-\log t)^{m-1} + \cdot\ \cdot\ \cdot \qquad (t \to 0)$$

(Proof) State density function

$$v(t) = \int \delta(t-K(g(u)))\ \varphi(g(u))\ |g'(u)|\ dt$$

is the inverse Mellin transform of zeta function.
Laurent expansion of zeta gives Lemma2.

Renormalized A posteriori distribution

Definition

Renormalized A Posteriori distribution is defined by

$$E_{u,t}[\quad] = \frac{\sum_{\alpha^*} \iint [\quad] S_\lambda(\xi(u))\, D(u)\, du\, dt}{\sum_{\alpha^*} \iint S_\lambda(\xi(u))\, D(u)\, du\, dt}$$

$$S_\lambda(a) = t^{\lambda-1} \exp(-\beta t + a\,\beta\, t^{1/2})$$

Renormalization

Lemma.3 For s=1,2,3, convergence in law holds,

$$E_w[\ (\ n^{1/2}\ f(X,w)\ )^{s}\ ] \rightarrow E_{u,t}[\ (\ a(X,u)t^{1/2}\ )^{s}\ ]$$

(Proof) $f(x,u) = a(x,u)\ u^k$ and

$$p(g(u)|X^n) = (1/C)\ \exp(\ -\ nu^{2k} - n^{1/2}u^k\xi(u))\ u^h\ b(u)$$

$$=(1/C)\int_0^\infty \exp(\ -\ t - t^{1/2}\xi(u))\ \delta(t-nu^{2k})\ u^h\ dt$$

$$\rightarrow (1/C)\ \frac{(\log n)^{m-1}}{n^\lambda}\int_0^\infty \exp(\ -\ t - t^{1/2}\xi(u))\ t^{\lambda-1}\ D(u)du\ dt$$

Singular Fluctuation

Definition. Singular Fluctuation is defined by

$$\nu(\beta) = \frac{\beta}{2} \; E_{\xi}\, E_{x} \left[E_{u,t}[a(X,u)^2\, t\,] - E_{u,t}[a(X,u)t^{1/2}\,]^2 \right]$$

Remarks.

(1) ν is the variance of renormalized log density ratio $a(x,u)t^{1/2}$.

(2) If a model is regular, ν=d/2.

Renormalization II

Lemma.4 Singular Fluctuation is birational invariant because

$$\nu/\beta = \lim_{n \to \infty} E\left[\sum_{i=1}^{n} \{ \log E_w[\, p(X_i|w)] - E_w[\log p(X_i|w)] \} \right]$$

(Proof) From Lemma.1,2, and 3, Lemma 4 is derived.

Remark: This lemma shows that SF can be estimated by random samples.

4

Main Theorem

Generalization and Training

Lemma. 5

There exist random variables Bg*, Bt*
such that both convergences in law and
convergences of expectation values hold.

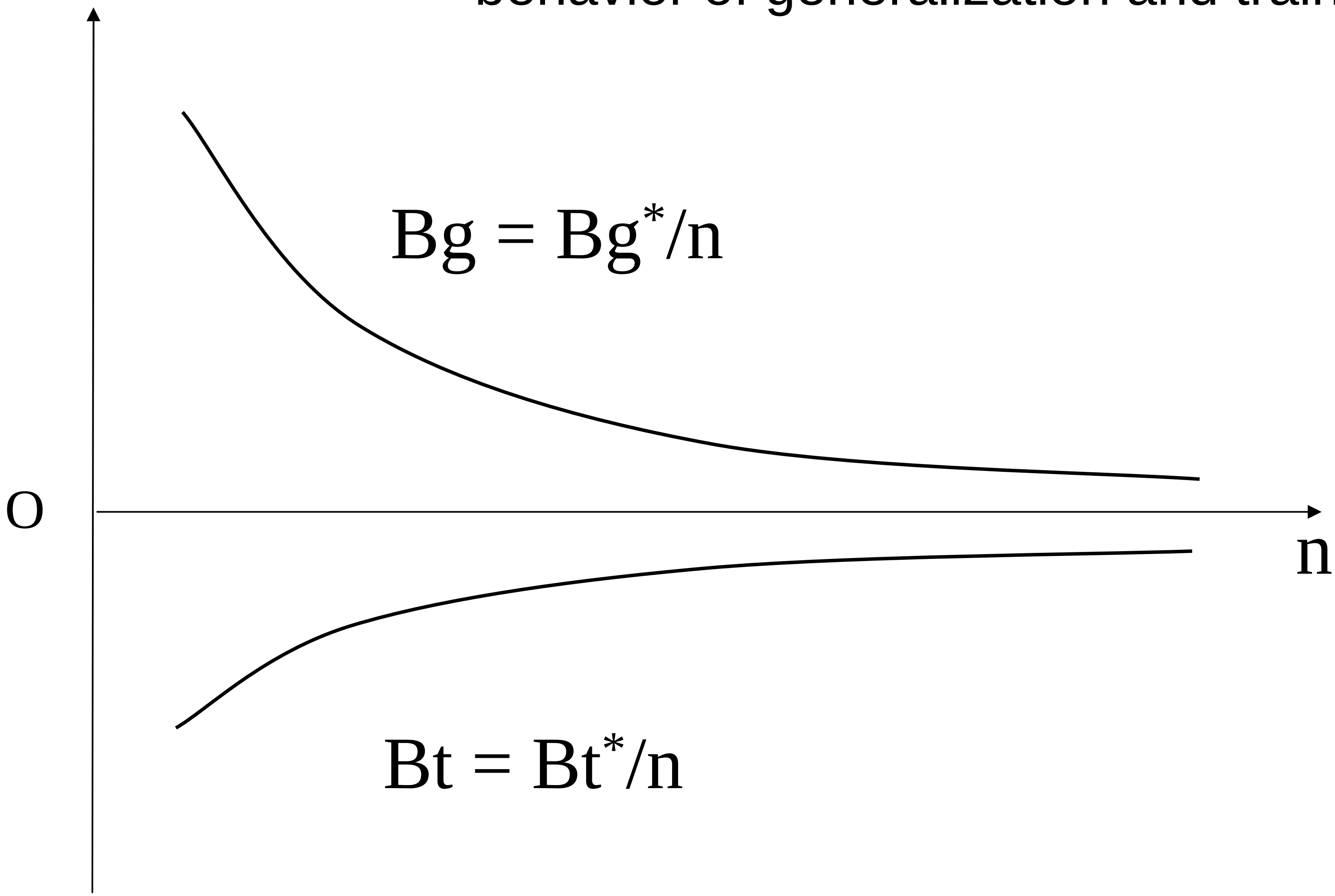
$nBg \to Bg^*$

$E[nBg] \to E[Bg^*]$

$nB\ t \to B$
$t\ ^*$

$E[nB\ t\ ] \to E[B\ t\ ^*]$

Theorem 4 shows the asymptotic behavior of generalization and training



Theorem. 2
Expectations of generalization and training errors are given by real log canonical threshold λ and singular fluctuation $\nu(\beta)$.

$$E[Bg^*] = \frac{\lambda}{\beta} \qquad (1/\beta-1)\,\nu(\beta)$$

$$E[B\ t\ ^*] = \frac{\lambda}{\beta} \qquad -(1/\beta+1)\,\nu(\beta)$$

Remark: If a model is regular, $\lambda=\nu$=d/2,
hence E[Bg*]=d/2, E[Bt*]=-d/2.

(Proof of theorem 2)

By using the fact that $\xi(u)$ is a gaussian random process,

$$\left\{\begin{array}{l} \frac{1}{2}\ E_\xi E_x E_{u,t}[\ a(X,u)t^{1/2}]^2 = \lambda/\beta+\nu(1-1/\beta) \\ \\ \frac{1}{2}\ E_\xi[E_{u,t}[\ \xi(u)t^{1/2}\ ]\ ] = \nu \\ \\ E_\xi E_{u,t}[\ t\ ] = \lambda/\beta+\nu \end{array}\right.$$

From definition, Bg* and Bt* are linear sums of three terms.

AIC is generalized

Theorem.3. (2007, Watanabe).

If a model is regular or singular model and $0 < \beta < \infty$,

$$E[\,B_g\,] = E[\,B_t\,] + 2\nu/n + o(1/n),$$

where $\nu = \nu(\beta)$ is a singular fluctuation which satisfies

$$\nu/\beta = \lim_{n \to \infty} E\left[\sum_{i=1}^{n} \{\, \log E_w[\,p(X_i|w)\,] - E_w[\,\log p(X_i|w)\,]\,\}\right]$$

Application to statistics

By defining widely applicable information criterion

$$\mathrm{WAIC}(p,\varphi) = -\sum_{i=1}^{n} \log E_w[\, p(X_i|w)\,] + 2\nu_n$$

$$\nu_n/\beta = \sum_{i=1}^{n} \{\ \log E_w[\, p(X_i|w)] - E_w[\log p(X_i|w)]\ \},$$

$\Rightarrow$ $E[\,\mathrm{WAIC}(p,\varphi)\,] = n\ (E[\,B_g(p,\varphi)\,]+S)$

Hyperparameters in (p,φ) are optimized.

Summary

Regular $\quad E[F] = \beta nS + (d/2) \log n + O(1),$

Singular $\quad E[F] = \beta nS + \lambda \log n - (m-1) \log\log n + O(1),$

λ : RLCT

Regular $\quad E[B_g] = E[B_t] + d/n + o(1/n),$

Singular $\quad E[B_g] = E[B_t] + 2\nu/n + o(1/n),$

ν : SF

Conclusion

1. Singular Models
2. & 3. Two invariants of singularities
   Real log canonical threshold and Singular Fluctuation
4. Expectations of Bayes generalization and training
   are determined by two invariants of singularities.

Future Study

In some models, RLCTs are obtained by resolution theorem.
Singular fluctuations are still unknown.

Notes from Media provided

RRR

λ

$f = (ab + cd)^2 + (ab^3 + cd^3)^2$

$\lambda = 2/3 \quad m = 1$

$\beta = 1 \quad E[B_g] = \frac{\lambda}{n} \quad \lambda = d/2 \quad \lambda \leq d/2$