What is Singular Learning Theory?

Shaowei Lin (UC Berkeley)

shaowei@math.berkeley.edu

23 Sep 2011

McGill University

Singular Learning Theory

A statistical model is *regular* if it is identifiable and its Fisher information matrix is postive definite. Behavior of regular models for large samples is well-understood, e.g. *central limit theorems*.

A model is *singular* if it is not regular. Many hidden variable models are singular. Singular learning theory teaches us how to study the *asymptotic behavior* of singular models: *by monomializing the Kullback-Leibler distance*.

Statistical Model

Let X be a random variable with state space $\mathcal{X}$ (e.g. $\{1, 2, \ldots, k\}$, $\mathbb{R}^k$).
Let $X_1, \ldots, X_N$ be N independent random samples of X.
Let the *true distribution* on X be $q(x)dx$.

Let $\mathcal{M}$ be a statistical model on $\mathcal{X}$ with parameter space Ω,
where the distribution at $\omega \in \Omega$ is denoted by $p(x|\omega)dx$
and the prior distribution on Ω is given by $\varphi(\omega)d\omega$.

likelihood function: $$L_N(\omega) = \prod_{i=1}^{N} p(X_i|\omega)$$

log likelihood ratio function: $$K_N(\omega) = \frac{1}{N} \sum_{i=1}^{N} \log \frac{q(X_i)}{p(X_i|\omega)}$$

Kullback-Leibler distance: $$K(\omega) = \int_{\mathcal{X}} q(x) \log \frac{q(x)}{p(x|\omega)} dx$$

Asymptotic Behavior

In *statistical learning theory*, we are interested in using the data to select a model $\mathcal{M}$ that best represents the true distribution. For this purpose, many *model selection criteria* (e.g. maximum likelihood, marginal likelihood, AIC, BIC) have been designed.

It is important to analyze how these criteria behave as the number of samples grow large. This analysis often depends on understanding the *log likelihood ratio* $K_N(\omega)$.

e.g. marginal likelihood

$$Z_N = \int_\Omega \prod_{i=1}^{N} p(X_i|\omega)\varphi(\omega)d\omega = \prod_{i=1}^{N} q(X_i) \cdot \int_\Omega e^{-NK_N(\omega)}\varphi(\omega)d\omega$$

The expectation of $K_N(\omega)$ is the *Kullback-Leibler distance* $K(\omega)$.

Regular and Singular Models

A model is *identifiable* if

$$p(x|\omega_1) = p(x|\omega_2) \text{ for all } x \in \mathcal{X} \quad \Rightarrow \quad \omega_1 = \omega_2.$$

The *Fisher information matrix* $I(\omega_0)$ is the Hessian matrix of the Kullback-Leibler distance $K(\omega)$ between $p(x|\omega_0)dx$ and $p(x|\omega)dx$. This matrix is always *positive semidefinite*.

A model is *regular* if it is identifiable and the Fisher information matrix $I(\omega)$ is *positive definite* for all $\omega \in \Omega$.

A model is *singular* if it is not regular. In particular, singular models are either nonidentifiable, or $\det I(\omega) = 0$ for some $\omega \in \Omega$.

The asymptotic behavior of regular models is well-understood.
[See Schwarz(1978), Haughton(1988), Lauritzen(1996).]
Unfortunately, many important models in learning theory are singular.

Resolution of Singularities

Watanabe's insight: find a change of variables $\rho : \mathscr{M} \to \Omega$ such that $K(\omega)$ becomes locally *monomial* on the *manifold* $\mathscr{M}$.

Such a change of variables always exists, due to a deep theorem in algebraic geometry known as *resolution of singularities*.
[Proved in 1964, this theorem won Hironaka the Fields Medal.]

Fundamental Theorem of Singular Learning (Watanabe)

> Given mild conditions on the model $\mathcal{M}$, there exists a change of variable $\rho : \mathscr{M} \to \Omega$ such that (μ^κ denotes $\mu_1^{\kappa_1} \cdots \mu_d^{\kappa_d}$)
>
> $$K_N(\rho(\mu)) = \mu^{2\kappa} - \frac{1}{\sqrt{N}} \mu^\kappa \xi_N(\mu)$$
>
> where $\xi_N(\mu)$ converges in law to a Gaussian process on $\mathscr{M}$.

The above formula is the *standard form* of the log likelihood ratio.

Learning Coefficient

Define stochastic complexity $F_N = -\log Z_N$.

Define empirical entropy $S_N = -\frac{1}{N}\sum_{i=1}^{N} \log q(X_i)$.

Convergence of stochastic complexity (Watanabe)

Given mild conditions on the model $\mathcal{M}$, the stochastic complexity has the asymptotic expansion

$$F_N = NS_N + \lambda \log N - (\theta - 1) \log \log N + F_N^R$$

where F_N^R converges in law to a random variable. Moreover, λ is the smallest pole, and θ its order, of the zeta function

$$\zeta(z) = \int_\Omega K(\omega)^{-z} \varphi(\omega) d\omega, \quad z \in \mathbb{C}.$$

We call λ the *learning coefficient* of the model $\mathcal{M}$ at the true distribution, and θ its *order*. We compute them by *monomializing* $K(\omega)$ and $\varphi(\omega)$.

Computation

Suppose $K(\omega) = \omega_1^{\kappa_1} \cdots \omega_d^{\kappa_d}$, $\varphi(\omega) = \omega_1^{\tau_1} \cdots \omega_d^{\tau_d}$ and $\Omega = [0, \varepsilon]^d$.

Then, the zeta function is

$$\begin{aligned}\zeta(z) &= \int_{[0,\varepsilon]^d} \omega_1^{-\kappa_1 z + \tau_1} \cdots \omega_d^{-\kappa_d z + \tau_d} \, d\omega \\ &= \frac{\varepsilon^{-\kappa_1 z + \tau_1 + 1}}{-\kappa_1 z + \tau_1 + 1} \cdots \frac{\varepsilon^{-\kappa_d z + \tau_d + 1}}{-\kappa_d z + \tau_d + 1}\end{aligned}$$

The poles of this function are $(\tau_i + 1)/\kappa_i$ for each i.

Thus, the learning coefficient is given by

$$\lambda = \min_i \frac{\tau_i + 1}{\kappa_i}$$

and its order θ is the number of times this minimum is attained.

The most *difficult* computation
in singular learning
is *finding* a change of variables
which monomializes $K(\omega)$.

Algebraic Geometry

Linear Algebra is the study of
systems of *linear* equations.

Commutative Algebra is the study of
systems of *polynomial* equations.

Algebraic Geometry is the study of
solutions of systems of polynomial equations.

Simple Example

Let $V \subset \mathbb{C}^2$ be the set of all solutions to $\{y - x^2 = 0, y = 0\}$.
The solution set to a system of polynomial equations is called a *variety*.

Given a subset $V \subset \mathbb{C}^2$ and polynomials $f(x, y), g(x, y)$,
if $f(x, y) = 0$ and $g(x, y) = 0$ for all $(x, y) \in V$, then

1. $f(x, y) + g(x, y) = 0$ for all $(x, y) \in V$,
2. $p(x, y)f(x, y) = 0$ for all $(x, y) \in V$ and all polynomials $p(x, y)$.

Let I be the set of polynomials generated by $y - x^2$ and y
via addition and polynomial multiplication. Notation: $I = \langle y - x^2, y \rangle$
A set of polynomials closed under these operations is called an *ideal*.

Is $x^2 \in I$? Is $x \in I$?

Ideals and Varieties

Let $\mathcal{R} = \mathbb{C}[x_1, x_2, \ldots, x_d]$ be a polynomial ring.

Given a subset $I \subset \mathcal{R}$, we define the *variety*

$$\mathcal{V}(I) = \{x \in \mathbb{C}^d \mid f(x) = 0 \text{ for all } f \in I\}.$$

Given a subset $V \subset \mathbb{C}^d$, we define the *ideal*

$$\mathcal{I}(V) = \{f \in \mathcal{R} \mid f(x) = 0 \text{ for all } x \in V\}.$$

The *algebraic closure* of V is the set $\overline{V} = \mathcal{V}(\mathcal{I}(V))$.

The *radical* of I is the set

$$\sqrt{I} = \{f \mid f^n \in I \text{ for some positive integer } n\}.$$

Fundamental Theorems

Hilbert Basis Theorem

Every ideal in $\mathbb{C}[x_1, \ldots, x_d]$ is finitely generated.

Hilbert's Nullstellensatz

$$\mathcal{I}(\mathcal{V}(I)) = \sqrt{I}$$

There is a bijective correspondence between radical ideals in $\mathbb{C}[x_1, \ldots, x_d]$ and varieties in $\mathbb{C}^d$.

BIG IDEA: Study varieties by studying their ideals.

Gröbner Bases

Every system of linear equations has a *row echelon form*, which is computed using *Gaussian elimination*.

Every system of polynomial equations has a *Gröbner basis*, which is computed using *Buchberger's algorithm*.

Determine ideal membership (e.g. Is $x^2 \in I$? Is $x \in I$?), dimension, degree, number of solutions, radicals, irreducible components, elimination of variables, etc.

Textbook:
"Ideals, Varieties, and Algorithms," Cox-Little-O'Shea(1997)

Software: `Macaulay2`, `Singular`, `Maple`, etc.

Disclaimer

The modern view of algebraic geometry, however, is quite different. Following *Grothendieck's* approach, it explores mathematical objects by studying *maps* between them.

Instead of studying a *solution set* V or its *ideal* I, we study the *coordinate ring* $\mathbb{C}[x_1, \ldots, x_d]/I$ of regular functions from the variety V to $\mathbb{C}$.

Such rings are *glued* together to form *sheafs* and *schemes*.

Real Log Canonical Thresholds

The Kullback-Leibler distance $K(\omega)$
is a *nonpolynomial* function that is
computationally difficult to monomialize.

Many singular models, however,
are regular models whose parameters are
polynomial functions of new parameters.

We want to *exploit* this polynomiality
in computing their learning coefficients.

Regularly Parametrized Models

A model $\mathcal{M}$ is *regularly parametrized* if it can be expressed as a regular model whose parameters $u = (u_i)$ are analytic functions $u_i(\omega)$ of new parameters $\omega = (\omega_i)$.

e.g. Discrete models $(p_1(\omega), p_2(\omega), \ldots, p_k(\omega))$
Gaussian models $X \sim \mathcal{N}(\mu, \Sigma)$, $\mu = (\mu_i(\omega))$, $\Sigma = (\sigma_{ij}(\omega))$

Suppose the true distribution lies in the model $\mathcal{M}$,
i.e. $q(x) = p(x|\omega^*)$ for some $\omega^* \in \Omega$.

Define the *fiber ideal* $I = \langle u_i(\omega) - u_i(\omega_i^*)$ for all $i \rangle$.
It is the ideal of the variety $V = \{\omega \in \Omega \mid q(x) = p(x|\omega)$ for all $x\}$.

[Note that if the model is identifiable, then V is a single point.]

Real Log Canonical Thresholds

In algebraic geometry, the *real log canonical threshold* of an ideal $\langle f_1(\omega), \ldots, f_k(\omega) \rangle$ is the pair (λ, θ) where λ is the smallest pole of the zeta function

$$\zeta(z) = \int_\Omega (f_1^2(\omega) + \cdots + f_k^2(\omega))^{-z/2} \, |\varphi(\omega)| d\omega$$

and θ its order. We denote $(\lambda, \theta) = \mathrm{RLCT}_\Omega(I; \varphi)$.

- This definition is independent of the choice of generators for I.
- Fix I, Ω and φ. For each point $x \in \Omega$, there exists a sufficiently small open neighborhood Ω_x of x in Ω such that $\mathrm{RLCT}_U(I; \varphi)$ is the same for all open neighborhoods U of x contained in Ω_x.
- We order the pairs (λ, θ) by the value of $\lambda \log N - (\theta - 1) \log \log N$ for sufficiently large N.

Exploiting Polynomiality

Theorem (L.)

Let $\mathcal{M}$ be a regularly parametrized model, and let the true distribution $q(x)dx$ be in $\mathcal{M}$. Given mild conditions on $\mathcal{M}$, the learning coefficient λ and its order θ of the model is given by

$$(2\lambda, \theta) = \min_{x \in \mathcal{V}(I)} \mathrm{RLCT}_{\Omega_x}(I; \varphi)$$

where I is the fiber ideal at the true distribution and $\mathcal{V}(I) \subset \Omega$ is the variety of the ideal.

[We compute $\mathrm{RLCT}(I; \varphi)$ by *monomializing* the ideal I and function φ, and using a geometric method involving *Newton polyhedra*.]

Examples

Example 1: **Bayesian Information Criterion**

When the model is regular, the fiber ideal is $I = \langle \omega_1, \ldots, \omega_d \rangle$.
Using Newton polyhedra, the RLCT of this ideal is $(d, 1)$.

By our theorem, the learning coefficient is $(\lambda, \theta) = (d/2, 1)$.
By Watanbe's theorem, the stochastic complexity is asymptotically

$$NS_N + \frac{d}{2} \log N.$$

This formula is the *Bayesian Information Criterion* (BIC).

Examples

Example 2: **132 Schizophrenic Patients**

Evans-Gilula-Guttman(1989) studied schizophrenic patients for connections between recovery time (in years Y) and frequency of visits by relatives.

	$2 \leq Y < 10$	$10 \leq Y < 20$	$20 \leq Y$	$Totals$
Regularly	43	16	3	62
Rarely	6	11	10	27
Never	9	18	16	43
$Totals$	58	45	29	$\mathbf{132}$

They wanted to find out if the data can be explained by a *naïve Bayesian network* with two hidden states (e.g. male and female).

Examples

Example 2: **132 Schizophrenic Patients**

The model is parametrized by $(t, a, b, c, d) \in \Delta_1 \times \Delta_2 \times \Delta_2 \times \Delta_2 \times \Delta_2$.

	$2 \leq Y < 10$	$10 \leq Y < 20$	$20 \leq Y$
Regularly	$ta_1b_1 + (1-t)c_1d_1$	$ta_1b_2 + (1-t)c_1d_2$	$ta_1b_3 + (1-t)c_1d_3$
Rarely	$ta_2b_1 + (1-t)c_2d_1$	$ta_2b_2 + (1-t)c_2d_2$	$ta_2b_3 + (1-t)c_2d_3$
Never	$ta_3b_1 + (1-t)c_3d_1$	$ta_3b_2 + (1-t)c_3d_2$	$ta_3b_3 + (1-t)c_3d_3$

As a model selection criteria, we compute the *marginal likelihood* of this model, given the above data and a uniform prior on the parameter space.

Examples

Example 2: **132 Schizophrenic Patients**

Lin-Sturmfels-Xu(2009) computed this integral *exactly*.
It is the rational number with numerator

$$27801948853106338912064360032498932910387614080528524283958209256935726588667532284587409752803399493069713103633199906939405711180837568853737$$

and denominator

$$12288402873591935400678094796599848745442833177572204504488199792864569951855421959468150731124291699780133503900169921912167352239204153786645029153951176422432983280461634722619620284616504320243563397065411323437531847188027481866765742374912000000000000000000.$$

Examples

Example 2: **132 Schizophrenic Patients**

We want to approximate the integral using asymptotic methods.
The EM algorithm gives us the *maximum likelihood distribution*

$$q = \frac{1}{132}\begin{pmatrix} 43.002 & 15.998 & 3.000 \\ 5.980 & 11.123 & 9.897 \\ 9.019 & 17.879 & 16.102 \end{pmatrix}.$$

Compare this distribution with the data

$$\begin{pmatrix} 43 & 16 & 3 \\ 6 & 11 & 10 \\ 9 & 18 & 16 \end{pmatrix}.$$

We use the ML distribution as the *true distribution* for our approximations.

Examples

Example 2: **132 Schizophrenic Patients**

Recall that stochastic complexity $= -\log$ (marginal likelihood).

- The BIC approximates the stochastic complexity as

$$NS_N + \frac{9}{2}\log N.$$

- By computing the RLCT of the fiber ideal, our approximation is

$$NS_N + \frac{7}{2}\log N.$$

- Summary:

	Stochastic Complexity
Exact	273.1911759
BIC	278.3558034
RLCT	275.9144024

“Algebraic Methods for Evaluating Integrals in Bayesian Statistics”

`http://math.berkeley.edu/~shaowei/swthesis.pdf`

(PhD dissertation, May 2011)

References

1. D. A. Cox, J. B. Little, and D. O'Shea: *Ideals, Varieties, and Algorithms: An Introduction to Computational Algebraic Geometry and Commutative Algebra*. Springer-Verlag, New York, 1997.
2. M. Evans, Z. Gilula and I. Guttman: Latent class analysis of two-way contingency tables by Bayesian methods, *Biometrika* **76** (1989) 557–563.
3. D. M. A. Haughton: On the choice of a model to fit data from an exponential family. *Ann. Statist.*, 16(1):342–355, 1988.
4. H. Hironaka: Resolution of singularities of an algebraic variety over a field of characteristic zero I, II, *Ann. of Math.* (2) **79** (1964) 109–203.
5. S. L. Lauritzen: *Graphical models*, volume 17 of *Oxford Statistical Science Se ries*. The Clarendon Press Oxford University Press, New York, 1996.
6. S. Lin, B. Sturmfels and Z. Xu: Marginal likelihood integrals for mixtures of independence models, *J. Mach. Learn. Res.* **10** (2009) 1611–1631.
7. S. Lin: Algebraic methods for evaluating integrals in Bayesian statistics, PhD dissertation, Dept. Mathematics, UC Berkeley (2011).
8. G. Schwarz: Estimating the dimension of a model. *Ann. Statist.*, 6(2):461–464, 1978.
9. S. Watanabe: *Algebraic Geometry and Statistical Learning Theory*, Cambridge Monographs on Applied and Computational Mathematics **25**, Cambridge University Press, Cambridge, 2009.