Singular Learning Theory:
A view from Algebraic Geometry

Shaowei Lin (UC Berkeley)

12 Dec 2011

American Institute of Mathematics

Integral Asymptotics

An Appetizer

For large N, approximate

$$Z(N) \;=\; \int_{[0,1]^2} (1 - x^2y^2)^{N/2}\, dxdy.$$

- Write $Z(N)$ as $\int e^{-Nf(x,y)} dxdy$ where

$$f(x,y) = -\frac{1}{2}\log(1 - x^2y^2).$$

- Can we use the Gaussian integral

$$\int_{\mathbb{R}^d} e^{-\frac{N}{2}(\omega_1^2+\cdots+\omega_d^2)} d\omega = \left(\frac{2\pi}{N}\right)^{d/2}$$

by finding a suitable change of coordinates for x, y?

Laplace Approximation

Ω small nbhd of origin, $f : \Omega \to \mathbb{R}$ analytic function with unique minimum $f(0)$ at origin, $\partial^2 f$ Hessian of f. If $\det \partial^2 f(0) \neq 0$,

$$Z(N) = \int_\Omega e^{-Nf(\omega)} d\omega \approx e^{-Nf(0)} \cdot \sqrt{\frac{(2\pi)^d}{\det \partial^2 f(0)}} \cdot N^{-d/2}.$$

- e.g. Bayesian Information Criterion (BIC)

$$-\log Z(N) \approx \left(-\sum_{i=1}^{N} \log q^*(X_i) \right) + \frac{d}{2} \log N$$

- e.g. Stirling's approximation

$$N! = N^{N+1} \int_0^\infty e^{-N(x - \log x)}\, dx \approx N^{N+1} e^{-N} \sqrt{\frac{2\pi}{N}}$$

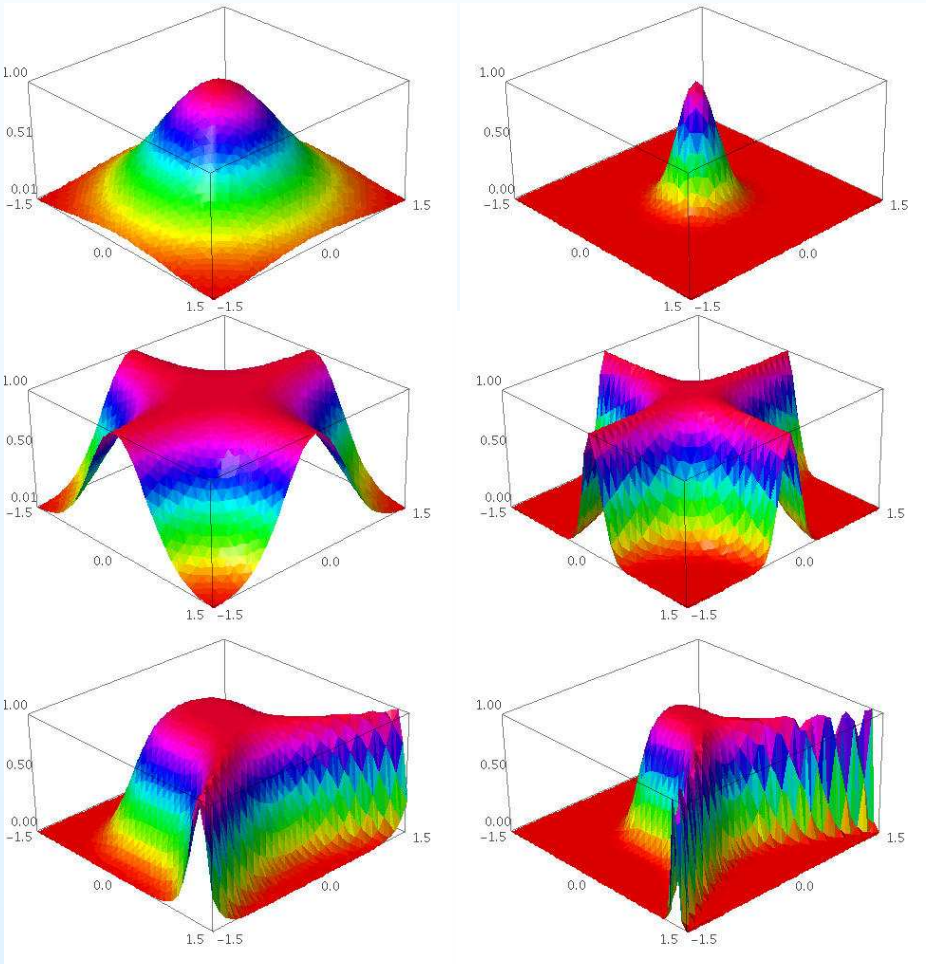
Geometry of the Integral

Because $\det \partial^2 f(0) = 0$ in our example, we cannot apply Laplace approximation. More important to study *minimas* of f.

$f(x, y) = x^2 + y^2$

$f(x, y) = (xy)^2$

$f(x, y) = (y^2 - x^3)^2$



Plots of $z = e^{-N f(x,y)}$ for $N = 1$ and $N = 10$

Monomial Functions

Notation: $\omega^\kappa = \omega_1^{\kappa_1} \cdots \omega_d^{\kappa_d}$.

Asymptotic theory of Arnol'd, Guseĭn-Zade and Varchenko (1974).

Theorem (AGV). Given $\kappa, \tau \in \mathbb{Z}_{\geq 0}^d$,

$$Z(N) \;=\; \int_{\mathbb{R}_{\geq 0}^d} e^{-N\omega^\kappa} \omega^\tau \, d\omega \;\approx\; C N^{-\lambda} (\log N)^{\theta - 1}$$

where C is a constant,

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

$\theta =$ number of times minimum is attained.

Resolution of Singularities

Let $\Omega \subset \mathbb{R}^d$ and $f : \Omega \to \mathbb{R}$ analytic function.

- We say $\rho : \mathscr{M} \to \Omega$ *desingularizes* f if

  1. $\mathscr{M}$ is a d-dimensional real analytic manifold covered by patches $U_1, \ldots, U_s$ ($\simeq$ subsets of $\mathbb{R}^d$).

  2. For each restriction $\rho : U_i \to \Omega$, $\mu \mapsto \omega$,

  $$f \circ \rho(\mu) = a(\mu)\mu^\kappa, \quad \det \rho'(\mu) = b(\mu)\mu^\tau$$

  where $a(\mu)$ and $b(\mu)$ are nonzero on U_i.

- Deep result in algebraic geometry (Hironaka's Theorem, 1964) that desingularizations always exist.

- The preimage of the *variety* $\{\omega : f(\omega) = 0\}$ is a *transform* $\{\mu : f \circ \rho(\mu) = 0\}$ that has *simple normal crossings*.

Algorithm for Computing Integral Asymptotics

$$Z(N) = \int_\Omega e^{-Nf(\omega)}\varphi(\omega)d\omega \approx e^{-Nf^*} \cdot CN^{-\lambda}(\log N)^{\theta-1}$$

Input:

Semialgebraic set $\Omega = \{\omega : g_1(\omega) \geq 0, \ldots, g_l(\omega) \geq 0\} \subset \mathbb{R}^d$

Analytic functions $f, \varphi : \Omega \to \mathbb{R}$

Output:

Asymptotic coefficients f^*, λ, θ

1. Find minimum f^* of f over Ω.
2. Find a desingularization ρ for product $(f - f^*)g_1 \cdots g_l\varphi$.
3. Use AGV Theorem to find coefficients λ_i, θ_i on each patch U_i.
4. $\lambda = \min\{\lambda_i\}$, $\theta = \max\{\theta_i : \lambda_i = \lambda\}$.

How do we desingularize $f(x, y) = -\frac{1}{2}\log(1 - x^2y^2)$?

Singular Learning

Statistical Model

X random variable with state space $\mathcal{X}$ (e.g. $\{1, 2, \ldots, k\}$, $\mathbb{R}^k$)

$\Delta_{\mathcal{X}}$ space of probability distributions on $\mathcal{X}$

$\mathcal{M} \subset \Delta_{\mathcal{X}}$ statistical model, image of $p : \Omega \to \Delta_{\mathcal{X}}$

Ω parameter space

$p(x|\omega)dx$ distribution at $\omega \in \Omega$

$\varphi(\omega)d\omega$ prior distribution on Ω

Given samples $X_1, \ldots, X_N$ of X, define *marginal likelihood*

$$Z_N = \int_\Omega \prod_{i=1}^N p(X_i|\omega)\, \varphi(\omega) d\omega.$$

Given $q \in \Delta_{\mathcal{X}}$, define *Kullback-Leibler function*

$$K(\omega) = \int_{\mathcal{X}} q(x) \log \frac{q(x)}{p(x|\omega)} dx.$$

Learning Coefficient

Suppose samples $X_1, \ldots, X_N$ are drawn from distribution $q \in \mathcal{M}$. Define *empirical entropy* $S_N = -\frac{1}{N} \sum_{i=1}^{N} \log q(X_i)$.

Convergence of stochastic complexity (Watanabe)

The *stochastic complexity* has the asymptotic expansion

$$-\log Z_N = N S_N + \lambda_q \log N - (\theta_q - 1) \log \log N + F_N^R$$

where F_N^R converges in law to a random variable. Moreover, λ_q, θ_q are asymptotic coefficients of the deterministic integral

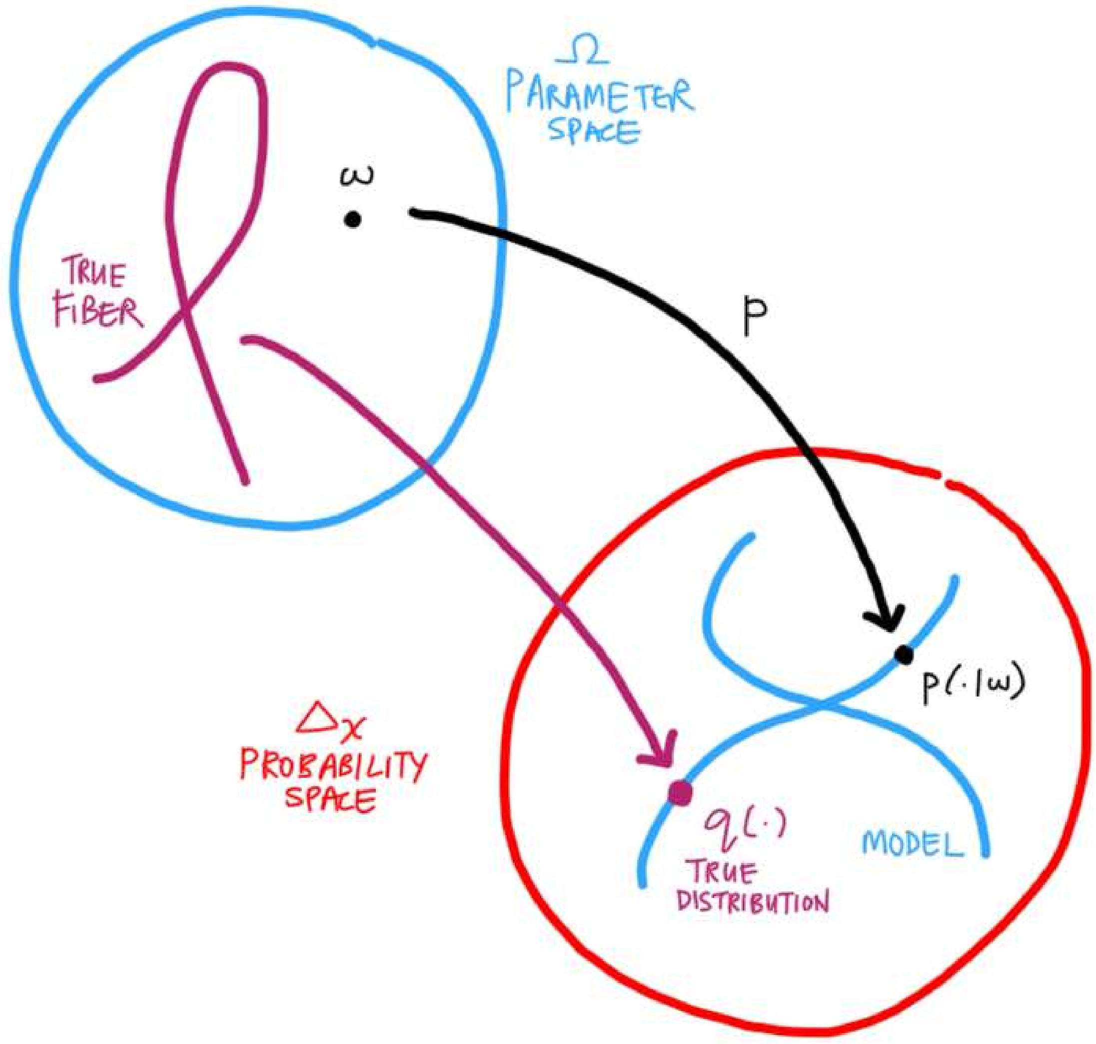
$$Z(N) = \int_\Omega e^{-NK(\omega)} \varphi(\omega) d\omega \;\approx\; C N^{-\lambda_q} (\log N)^{\theta_q - 1}.$$

Think of this as *generalized BIC* for singular models.

λ_q, θ_q *learning coefficient* of the model $\mathcal{M}$ at q, and its *order*.

Loosely speaking, a model is *regular* if the Laplace approximation applies to $Z(N)$. Otherwise, it is *singular*.

Geometry of Singular Models



Standard Form of Log Likelihood Ratio

Define *log likelihood ratio*. Note that its expectation is $K(\omega)$.

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

Standard Form of Log Likelihood Ratio (Watanabe)

Suppose $\rho : \mathscr{M} \to \Omega$ desingularizes $K(\omega)$. Then,

$$K_N \circ \rho(\mu) = \mu^{2\kappa} - \frac{1}{\sqrt{N}} \mu^\kappa \xi_N(\mu)$$

where $\xi_N(\mu)$ converges in law to a Gaussian process on $\mathscr{M}$.

Think of this as *generalized CLT* for singular models.
Classical central limit theorem (CLT):

$$\text{sample mean} = \frac{1}{N} \sum_{i=1}^N X_i = \mu + \frac{1}{\sqrt{N}} \sigma \xi_N$$

where ξ_N converges in law to standard normal distribution.

Mathematical Questions in Singular Learning

For each distribution q in the model $\mathcal{M}$,

1. Study the geometrical structure of the fiber $p^{-1}(q)$.
2. Study the asymptotics of the integral

$$Z(N) = \int_\Omega e^{-NK(\omega)}\varphi(\omega)d\omega$$

   and compute the learning coefficient λ_q and its order θ_q.
3. Desingularize the Kullback-Leibler function $K(\omega)$.

Algebraic Geometry

Exploiting Polynomiality

How do we desingularize $f(x,y) = -\frac{1}{2}\log(1-x^2y^2)$?

- Algorithms for desingularization (e.g. Bravo-Encinas-Villamayor) intractable when applied to nonpolynomial functions like $K(\omega)$.
- Many singular models parametrized by *polynomials*. Exploit this?

Regularly parametrized models

- A model is *regularly parametrized* if its map $p : \Omega \to \Delta_{\mathcal{X}}$ factors

$$\Omega \xrightarrow{u} U \xrightarrow{f} \Delta_{\mathcal{X}}$$

  where $f : U \to \Delta_{\mathcal{X}}$ defines a regular model and $u : \Omega \to U \subset \mathbb{R}^k$ is a polynomial map.
- e.g. Gaussian/discrete models parametrized by polynomials.
- Kullback-Leibler function $K(\omega)$ also factors

$$\Omega \xrightarrow{u} U \xrightarrow{f} \Delta_{\mathcal{X}} \xrightarrow{g} \mathbb{R}$$

  where Hessian of $g \circ f$ has nonzero determinant at q.

Ideals and Varieties

Polynomial system		Solution set (*variety*)
$\{y - x^2, y\} \subset \mathbb{R}[x, y]$	$\longleftrightarrow$	$V = \{(0,0)\} \subset \mathbb{R}^2$

- If $y - x^2$ and y vanish on V, so do all polynomials of the form
$$p(x,y) = (y - x^2)\, p_1(x,y) + (y)\, p_2(x,y).$$
This infinite set of polynomials is the *ideal* $I = \langle y - x^2, y\rangle$.

- Vector spaces: generated by addition, scalar multiplication.
Ideals: generated by addition, polynomial multiplication.

- Ideal membership. Is $x^2 \in I$? Is $x \in I$?

- Given subset $I \subset \mathcal{R} := \mathbb{R}[x_1, \ldots, x_d]$, define the *variety*
$$\mathcal{V}(I) = \{x \in \mathbb{R}^d : f(x) = 0 \text{ for all } f \in I\}.$$
Given subset $V \subset \mathbb{R}^d$, define the *ideal*
$$\mathcal{I}(V) = \{f \in \mathcal{R} : f(x) = 0 \text{ for all } x \in V\}.$$

Gröbner Bases

- Every system of linear equations has a *row echelon form*, which depends on the ordering of the coordinates and is computed using *Gaussian elimination*.
- Every system of polynomial equations has a *Gröbner basis*, which depends on the ordering of the monomials and is computed using *Buchberger's algorithm*.
- Determine ideal membership, dimension, degree, number of solutions, irreducible components, elimination of variables, etc. Also essential in resolution of singularities.
- **Textbook**:
  "Ideals, Varieties, and Algorithms," Cox-Little-O'Shea (1997)
  Software:
  `Macaulay2`, `Singular`, `Maple`, etc.

Fiber Ideals

Regularly Parametrized Functions

- A function $f : \Omega \to \mathbb{R}$ is *regularly parametrized* if it factors

$$\Omega \xrightarrow{u} U \xrightarrow{g} \mathbb{R}$$

where $U \subset \mathbb{R}^k$ nbhd of origin, u is polynomial, g has unique minimum $g(0) = 0$ at the origin and $\det \partial^2 g(0) \neq 0$.

- For such functions, define *fiber ideal*

$$I = \langle u_1(\omega), \ldots, u_k(\omega) \rangle \subset \mathbb{R}[\omega_1, \ldots, \omega_d].$$

It is the ideal of the fiber $f^{-1}(0)$.

- e.g. $f(x, y) = -\frac{1}{2} \log(1 - x^2 y^2)$

$$u(x, y) = xy, \;\; g(u) = -\tfrac{1}{2} \log(1 - u^2)$$

fiber ideal $I = \langle xy \rangle$

Real Log Canonical Thresholds

Definition

Given ideal $I = \langle f_1(\omega), \ldots, f_k(\omega) \rangle \subset \mathbb{R}[\omega_1, \ldots, \omega_d]$,
semialgebraic set $\Omega \subset \mathbb{R}^d$,
polynomial $\varphi(\omega) \in \mathbb{R}[\omega_1, \ldots, \omega_d]$,

the *real log canonical threshold* of I is the pair (λ, θ) where λ is the smallest pole of the zeta function

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

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

- Definition is independent of choice of generators $f_1, \ldots, f_k$.
- The poles of the zeta function are positive *rational* numbers.
- Order the pairs (λ, θ) by the value of $\lambda \log N - \theta \log\log N$ for sufficiently large N.

Learning Coefficients

Theorem (L.)

Let $f = g \circ u : \Omega \to \mathbb{R}$ be a regularly parametrized function. The asymptotic coefficient (λ, θ) of the integral

$$Z(N) \; = \; \int_\Omega e^{-Nf(\omega)} \varphi(\omega) d\omega \; \approx \; CN^{-\lambda} (\log N)^{\theta - 1}$$

is given by

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

where $I = \langle u_1, \ldots, u_k \rangle$ is the fiber ideal of f, $\mathcal{V}(I) \subset \Omega$ is the fiber $f^{-1}(0)$, and each Ω_x is a sufficiently small nbhd of x in Ω.

Corollary (L.)

Given regularly parametrized model $\mathcal{M}$ and $q \in \mathcal{M}$, the learning coefficient (λ_q, θ_q) is given by the above formula.

Newton Polyhedra

Given an ideal $I \subset \mathbb{R}[\omega_1, \ldots, \omega_d]$,

1. Plot $\alpha \in \mathbb{R}^d$ for each monomial ω^α appearing in some $f \in I$.
2. Take the convex hull $\mathcal{P}(I)$ of all plotted points.

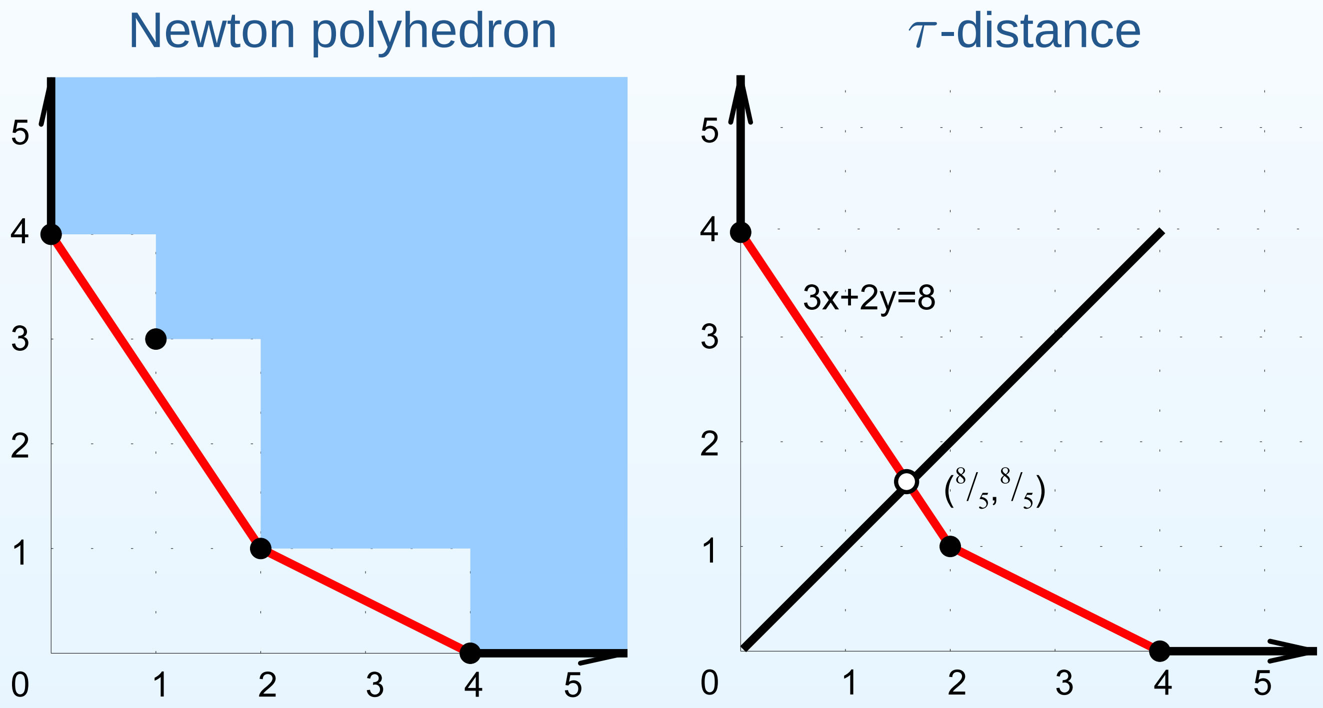
This convex hull $\mathcal{P}(I)$ is the *Newton polyhedron* of I.

Given a vector $\tau \in \mathbb{Z}^d_{\geq 0}$, define

1. *τ-distance* $l_\tau = \min\{t : t(\tau_1 + 1, \ldots, \tau_d + 1) \in \mathcal{P}(I)\}$.
2. *multiplicity* $\theta_\tau =$ codim of face of $\mathcal{P}(I)$ at this intersection.

Newton Polyhedra

e.g. Let $I = \langle x^4, x^2y, xy^3, y^4 \rangle$ and $\tau = (0,0)$.



The τ-distance is $l_\tau = 8/5$ and the multiplicity is $\theta_\tau = 1$.

Upper Bounds

Let $I \subset \mathbb{R}[\omega_1, \ldots, \omega_d]$ be an ideal.

Proposition (Trivial) $\quad \mathrm{RLCT}_\Omega(I;\varphi) \leq d$

Theorem (Watanabe) $\quad \mathrm{RLCT}_\Omega(I;\varphi) \leq \operatorname{codim} \mathcal{V}(I)$

Theorem (L.)

For a sufficiently small nbhd $U \subset \mathbb{R}^d$ of the origin,

$$\mathrm{RLCT}_U(I;\omega^\tau) \leq (1/l_\tau, \theta_\tau)$$

where l_τ is the τ-distance of Newton polyhedron $\mathcal{P}(I)$ and θ_τ its multiplicity. Equality occurs when I is a monomial ideal.

e.g. $f(x,y) = -\frac{1}{2}\log(1-x^2y^2),\ I = \langle xy \rangle,\ \tau = (0,0).$

The τ-distance l_τ is 1, and its multiplicity θ_τ is 2.
Therefore, $Z(N) \approx CN^{-1/2}(\log N)$.

Applications to Statistics

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 Watanabe's theorem, the stochastic complexity is asymptotically

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

This formula is the Bayesian Information Criterion (BIC).

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.

	$2 \le Y < 10$	$10 \le Y < 20$	$20 \le Y$	*Totals*
Regularly	43	16	3	*62*
Rarely	6	11	10	*27*
Never	9	18	16	*43*
Totals	*58*	*45*	*29*	**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).

Example 2: 132 Schizophrenic Patients

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

	$2 \le Y < 10$	$10 \le Y < 20$	$20 \le Y$
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	$ta_3b_1 + (1-t)c_3d_1$	$ta_3b_2 + (1-t)c_3d_2$	$ta_3b_3 + (1-t)c_3d_3$

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

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

$$
\begin{array}{l}
278019488531063389120643600324989329103876140805 \\
285242839582092569357265886675322845874097528033 \\
99493069713103633199906939405711180837568853737
\end{array}
$$

and denominator

$$
\begin{array}{l}
12288402873591935400678094796599848745442833177572204 \\
50448819979286456995185542195946815073112429169997801 \\
33503900169921912167352239204153786645029153951176422 \\
43298328046163472261962028461650432024356339706541132 \\
34375318471880274818667657423749120000000000000000.
\end{array}
$$

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

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

Use ML distribution as *true distribution* for our approximations.

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

Desingularization

Strategy for Regularly Parametrized Functions

Given a regularly parametrized function $f = g \circ u : \Omega \to \mathbb{R}$, we want to *exploit the polynomiality* in u in desingularizing f.
Let $I = \langle u_1, \ldots, u_k \rangle$ be the polynomial fiber ideal.
Given $\rho : M \to \Omega$, define *pullback* $\rho^* I = \langle u_1 \circ \rho, \ldots, u_k \circ \rho \rangle$.

1. **Monomialization** (polynomial):
   Find a map $\rho : M \to \Omega$ which *monomializes* I,
   i.e. $\rho^* I$ is a monomial ideal in each patch of M.
   Use algorithm of Bravo-Encinas-Villamayor.

2. **Principalization** (combinatorial):
   Find a map $\eta : \mathscr{M} \to M$ which *principalizes* $J = \rho^* I$,
   i.e. $\eta^* J$ is generated by one monomial in each patch of $\mathscr{M}$.
   Use toric blowups or Goward's principalization map.

Theorem (L.) The composition $\rho \circ \eta$ desingularizes f.

Higher Order Asymptotics

Using this strategy to desingularize $f(x,y) = -\frac{1}{2}\log(1-x^2y^2)$, we were able to compute higher order asymptotics of $Z(N)$.

$$\sqrt{\frac{\pi}{8}} N^{-\frac{1}{2}} \log N \quad - \sqrt{\frac{\pi}{8}} \left(\frac{1}{\log 2} - 2\log 2 - \gamma \right) N^{-\frac{1}{2}}$$
$$- \frac{1}{4} N^{-1} \log N \quad + \frac{1}{4} \left(\frac{1}{\log 2} + 1 - \gamma \right) N^{-1}$$
$$- \frac{\sqrt{2\pi}}{128} N^{-\frac{3}{2}} \log N \quad + \frac{\sqrt{2\pi}}{128} \left(\frac{1}{\log 2} - 2\log 2 - \frac{10}{3} - \gamma \right) N^{-\frac{3}{2}}$$
$$- \frac{1}{24} N^{-2} \quad + \cdots$$

Euler-Mascheroni constant $$\gamma = \lim_{n\to\infty} \left(\sum_{k=1}^{n} \frac{1}{k} - \log n \right) \approx 0.5772156649.$$

"Algebraic Methods for Evaluating Integrals in Bayesian Statistics"

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

(PhD dissertation, May 2011)

References

1. V. I. Arnol'd, S. M. Guseĭn-Zade and A. N. Varchenko: *Singularities of Differentiable Maps*, Vol. II, Birkhäuser, Boston, 1985.
2. A. Bravo, S. Encinas and O. Villamayor: A simplified proof of desingularisation and applications, *Rev. Math. Iberoamericana* **21** (2005) 349–458.
3. 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.
4. M. Evans, Z. Gilula and I. Guttman: Latent class analysis of two-way contingency tables by Bayesian methods, *Biometrika* **76** (1989) 557–563.
5. 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.
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. S. Watanabe: *Algebraic Geometry and Statistical Learning Theory*, Cambridge Monographs on Applied and Computational Mathematics **25**, Cambridge University Press, Cambridge, 2009.

Supplementary

The integral $Z(N)$ with $f(x,y) = -\frac{1}{2}\log(1-x^2y^2)$ comes from the coin toss model parametrized by

$$p_1(\omega,t) = \frac{1}{2}t + (1-t)\omega$$
$$p_2(\omega,t) = \frac{1}{2}t + (1-t)(1-\omega)$$

where the Kullback-Leibler function at the distribution (q_1, q_2)

$$K(\omega,t) = q_1 \log \frac{q_1}{p_1(\omega,t)} + q_2 \log \frac{q_2}{p_2(\omega,t)}.$$

The function $f(x,y)$ comes from $K(x,y)$ at $q_1 = q_2 = 1/2$ and substituting $\omega = (1+x)/2, t = 1-y$.