The Algebraic Geometry of Singular Learning Theory

Shaowei Lin (UC Berkeley)

13 June 2012

Pennsylvania State University

Singular Learning Theory

Bayesian Statistics

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

$\mathcal{M} \subset \Delta$	statistical model, image of $p : \Omega \to \Delta$
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 $q \in \Delta$, 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

$$-\log Z_N = NS_N + \lambda_q \log N - (\theta_q - 1)\log\log N + R_N$$

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

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

For regular models, this is the *Bayesian Information Criterion*.

Various names for (λ_q, θ_q):

statistics - *learning coefficient* of the model $\mathcal{M}$ at q

algebraic geometry - *real log canonical threshold* of $K(\omega)$

Integral Asymptotics

Geometry of the Integral

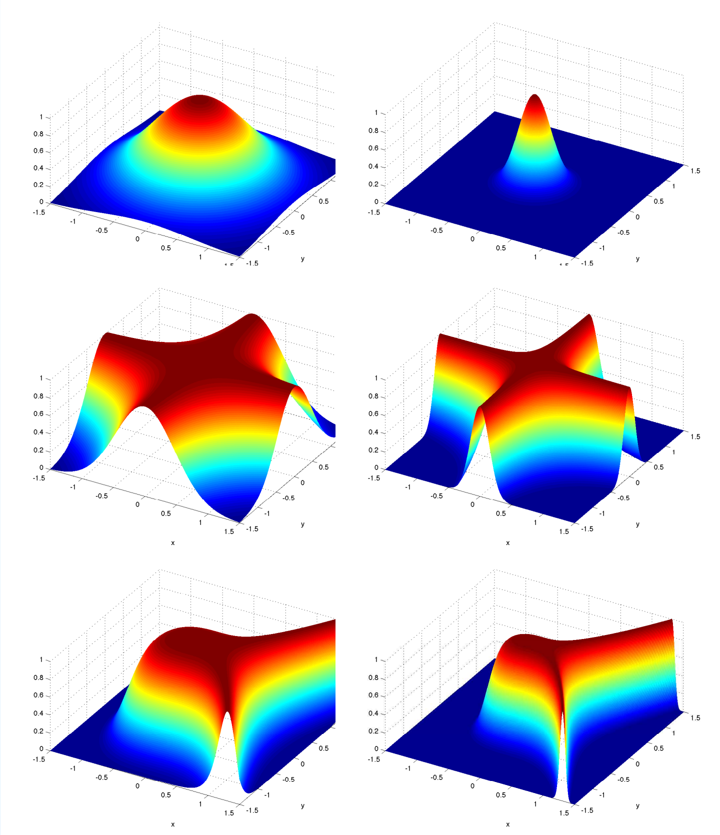
$$Z(N) = \int_\Omega e^{-Nf(\omega)}\varphi(\omega)d\omega \approx e^{-Nf^*} \cdot CN^{-\lambda}(\log N)^{\theta-1}$$

Integral asymptotics depend on *minimum locus* of exponent $f(\omega)$.

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

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

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



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

Desingularization and Monomial Functions

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

- We say $\rho : U \to \Omega$ *desingularizes* f if

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

  2. For each restriction $\rho : U_i \to \Omega$,
  $$f \circ \rho(\mu) = a(\mu)\mu^\kappa, \quad \det \partial\rho(\mu) = b(\mu)\mu^\tau$$
  where $a(\mu)$ and $b(\mu)$ are nonzero on U_i.

- Hironaka (1964) proved that desingularizations always exist.

RLCT of monomial functions (Arnol'd·Guseĭn-Zade·Varchenko)

$$Z(N) = \int_\Omega e^{-N\omega_1^{\kappa_1}\cdots\omega_d^{\kappa_d}} \omega_1^{\tau_1}\cdots\omega_d^{\tau_d} d\omega \approx CN^{-\lambda}(\log N)^{\theta-1}$$

where $\lambda = \min_i \frac{\tau_i+1}{\kappa_i}, \theta =$ number of times minimum is attained.

Algorithm for Computing RLCTs

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

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. $\lambda = \min\{\lambda_i\}$, $\theta = \max\{\theta_i : \lambda_i = \lambda\}$.

Upper bound (trivial) $\lambda \leq \frac{d}{2}$

Upper bound (Watanabe) $\lambda \leq \frac{1}{2}($ codim of minimum locus of f $)$

Real Log Canonical Thresholds

Exploiting Polynomiality

How do we desingularize $K(\omega) = \int_{\mathcal{X}} q(x) \log \frac{q(x)}{p(x|\omega)} dx$?

- Algorithms (e.g. Bravo-Encinas-Villamayor) intractable
- Many models parametrized by *polynomials*. Exploit this?

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*

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

  The variety $\mathcal{V}(I)$ is the fiber $f^{-1}(0)$.

Equivalence (Watanabe) RLCT of f = RLCT of $u_1^2 + \cdots + u_k^2$.

Real Log Canonical Thresholds of Ideals

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

The *real log canonical threshold* (λ, θ) of I at $x \in \Omega$ satisfies

$$\int_{\Omega_x} e^{-N(f_1^2+\cdots+f_k^2)}\,\varphi(\omega)d\omega \;\approx\; CN^{-\lambda}(\log N)^{\theta-1}$$

for suff small nbhd Ω_x of x in Ω. Denote $(\lambda, \theta) = \mathrm{RLCT}_{\Omega_x}(I;\varphi)$.

Properties

- Definition is independent of choice of generators $f_1, \ldots, f_k$.
- λ positive *rational* number, θ positive *integer*.
- Depends on structure of boundary $\partial\Omega$ if $x \in \partial\Omega$.
- Order the (λ, θ) by the value of $N^{\lambda}(\log N)^{-\theta}$ for large N.

Discrete and Gaussian Models

- *Discrete models* with state probabilities $p(\omega)$.
  Fiber ideal at a true distribution $\hat{p}$

$$I_{\hat{p}} = \langle p_i(\omega) - \hat{p}_i \rangle_i$$

- *Gaussian models* with mean $\mu(\omega)$ and covariance $\Sigma(\omega)$.
  Fiber ideal at a true distribution $\mathcal{N}(\hat{\mu}, \hat{\Sigma})$

$$I_{\hat{\mu},\hat{\Sigma}} = \langle \mu_i(\omega) - \hat{\mu}_i, \Sigma_{ij}(\omega) - \hat{\Sigma}_{ij} \rangle_{ij}$$

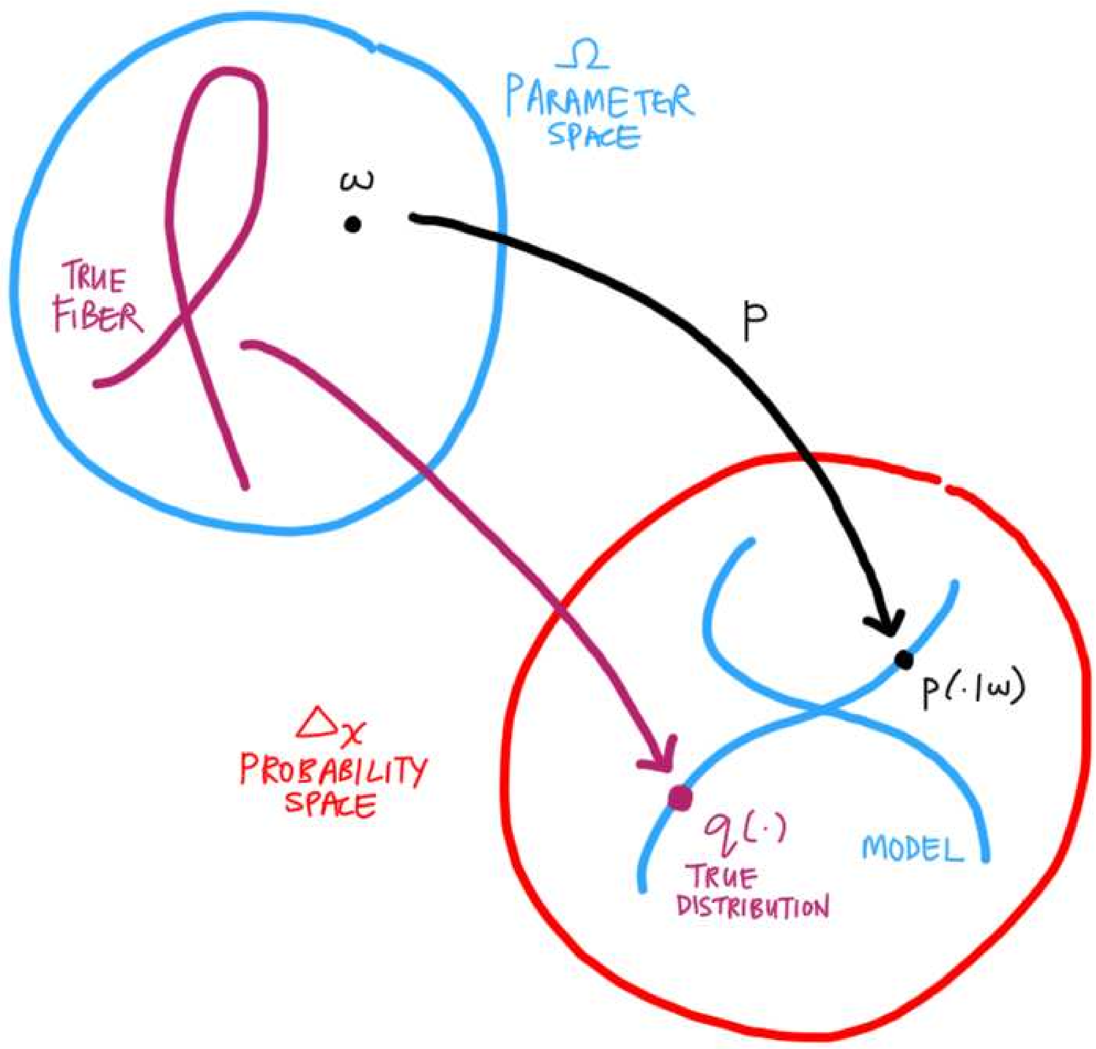
Learning coefficients and RLCTs of fiber ideals (L.)

If the true distribution q is in the model,
then the learning coefficient (λ_q, θ_q) is given by

$$(2\lambda_q, \theta_q) = \min_{x \in \mathcal{V}(I_q)} \mathrm{RLCT}_{\Omega_x}(I_q; \varphi)$$

where I_q is the fiber ideal at q and $\mathcal{V}(I_q) \subset \Omega$ is the fiber over q.

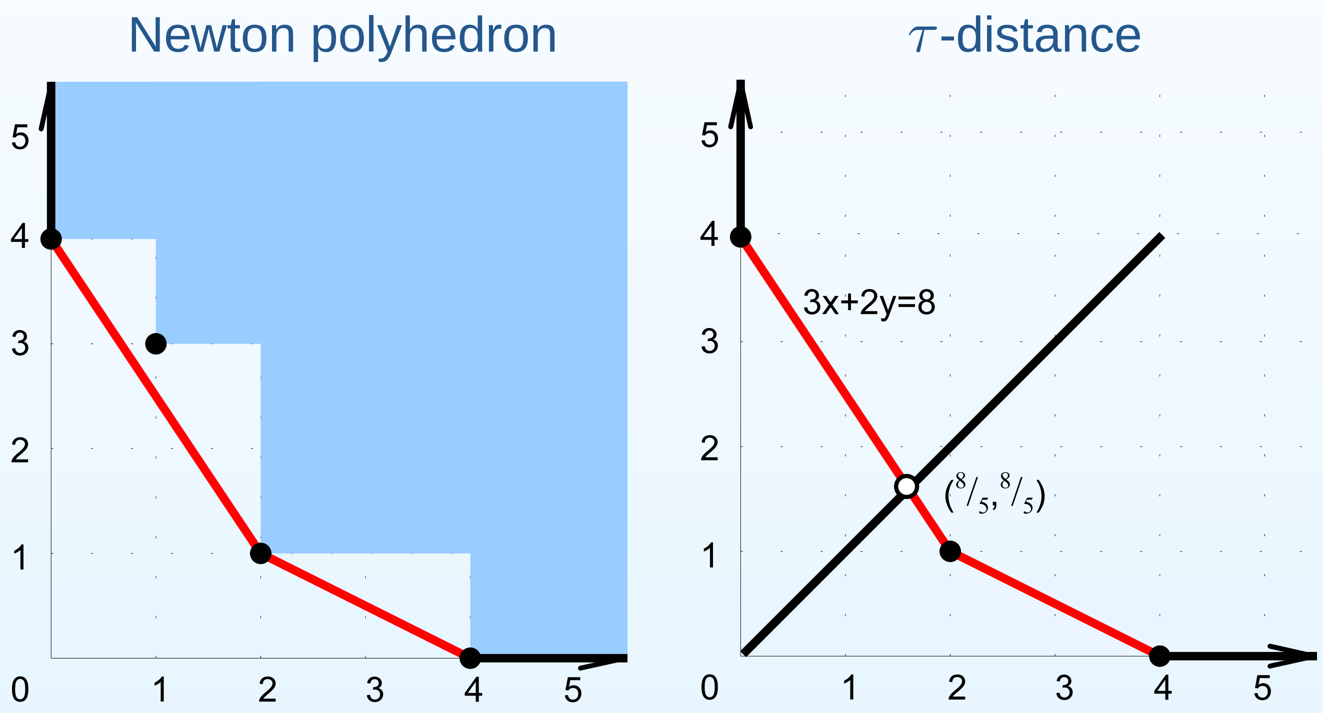
Geometry of Singular Models



Newton Polyhedra

Distance and Multiplicity

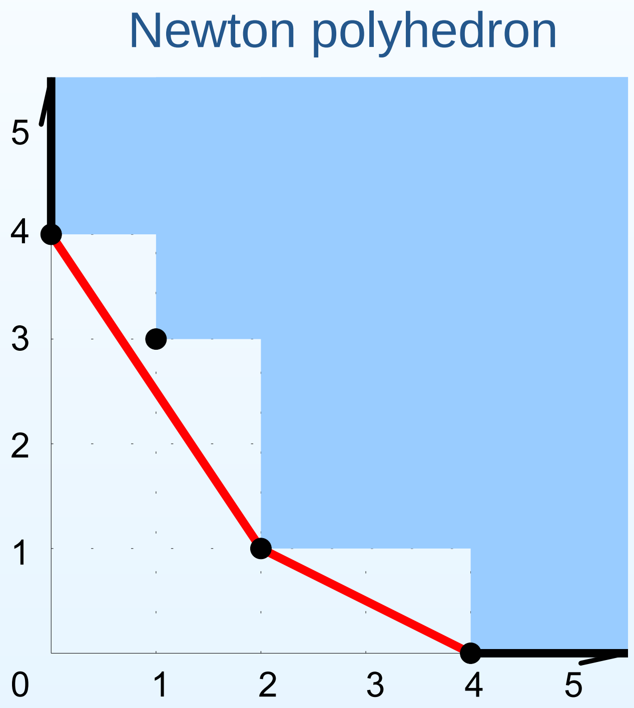
e.g. Let $I = \langle x^4, x^2y, xy^3, y^4 \rangle$ and $\tau = (1, 1)$.

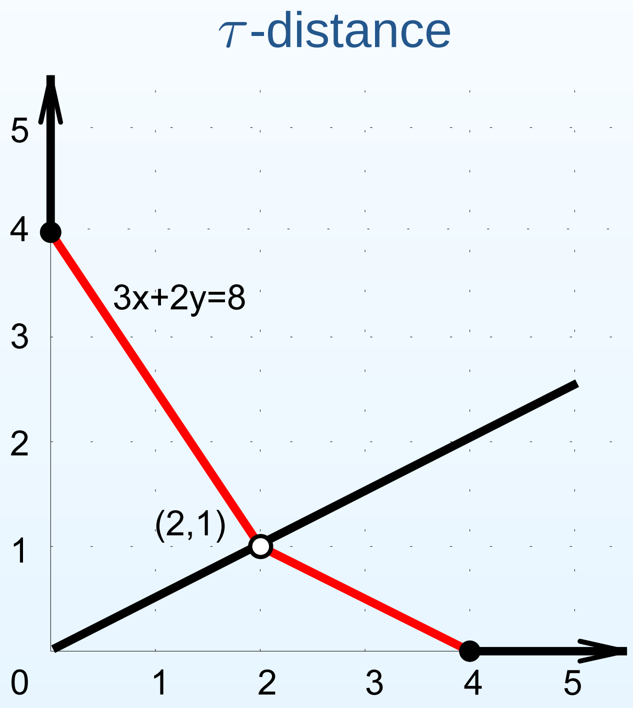


The τ-distance is $l_\tau = 8/5$ and the multiplicity is $\theta_\tau = 1$.

Distance and Multiplicity

e.g. Let $I = \langle x^4, x^2y, xy^3, y^4 \rangle$ and $\tau = (2,1)$.





The τ-distance is $l_\tau = 1$ and the multiplicity is $\theta_\tau = 2$.

Relation to RLCTs

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* $l_\tau = \min\{t : t\tau \in \mathcal{P}(I)\}$.
2. *multiplicity* $\theta_\tau =$ codim of face of $\mathcal{P}(I)$ at this intersection.

Upper bound and equality for RLCT (L.)

If l_τ is the τ-distance of $\mathcal{P}(I)$ and θ_τ is its multiplicity, then

$$\mathrm{RLCT}_\Omega(I; \omega^{\tau-1}) \leq (1/l_\tau, \theta_\tau).$$

Equality occurs when I is a monomial ideal.

`Macaulay2`
Computations

132 Schizophrenic Patients (Evans·Gilula·Guttman)

Naïve Bayes network with 2 ternary variables, 2 hidden states. Model parametrized in $\omega = (t, a_1, a_2, \ldots, d_3)$ by

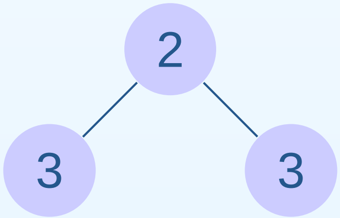
$$p = \begin{pmatrix} ta_1b_1 + (1-t)c_1d_1 & ta_1b_2 + (1-t)c_1d_2 & ta_1b_3 + (1-t)c_1d_3 \\ ta_2b_1 + (1-t)c_2d_1 & ta_2b_2 + (1-t)c_2d_2 & ta_2b_3 + (1-t)c_2d_3 \\ ta_3b_1 + (1-t)c_3d_1 & ta_3b_2 + (1-t)c_3d_2 & ta_3b_3 + (1-t)c_3d_3 \end{pmatrix}.$$

Assume true distribution $\hat{p}_{ij} = \frac{1}{9}$ for all i, j.

Compute RLCT of fiber ideal

$$I = \langle p_{11}(\omega) - \hat{p}, \ldots, p_{33}(\omega) - \hat{p} \rangle$$

at the point $\hat{w} = (\frac{1}{2}, \frac{1}{3}, \frac{1}{3}, \ldots, \frac{1}{3}) \in \mathcal{V}(I)$.



Computations using our library `asymptotics.m2` show that

$$\mathrm{RLCT}_{\hat{\omega}}(I; 1) = (6, 2).$$

All other learning coefficients can be computed in this fashion.

Model Definition

```
Macaulay2, version 1.4
with packages: ConwayPolynomials, Elimination,
               IntegralClosure, LLLBases,
               PrimaryDecomposition, ReesAlgebra,
               TangentCone

i1 : load "asymptotics.m2";
i2 : R = QQ[t,a1,a2,b1,b2,c1,c2,d1,d2];
i3 : A = matrix {{a1,a2,1-a1-a2}};
i4 : B = matrix {{b1,b2,1-b1-b2}};
i5 : C = matrix {{c1,c2,1-c1-c2}};
i6 : D = matrix {{d1,d2,1-d1-d2}};
i7 : P = t*(transpose A)*B + (1-t)*(transpose C)*D;
                   3       3
o7 : Matrix R  <--- R
```

Fiber Ideal

Maps for shifting the origin to $\hat{\omega}$ and evaluating a polynomial at $\hat{\omega}$.

```
i8 : shift = map(R,R,{t+1/2,a1+1/3,a2+1/3,b1+1/3,b2+1/3,
                            c1+1/3,c2+1/3,d1+1/3,d2+1/3});
i9 : eval = map(R,R,{1/2,1/3,1/3,1/3,1/3,
                         1/3,1/3,1/3,1/3});
```

The true distribution.

```
i10 : eval P
o10 = {-1} | 1/9 1/9 1/9 |
      {-1} | 1/9 1/9 1/9 |
      {-1} | 1/9 1/9 1/9 |
```

The fiber ideal.

```
i11 : I = ideal (shift P - eval P);
o11 : Ideal of R
```

Gröbner Basis

Gröbner basis of the fiber ideal.

```
i12 : I = ideal gens gb I
o12 = ideal (a2*d2, a1*d2, b2*d1 - b1*d2, a2*d1, a1*d1,
      b2*c2, b1*c2, b2*c1, b1*c1, a2*c1 - a1*c2,
      2t*b2 - 2t*d2 + b2 + d2, 2t*b1 - 2t*d1 + b1 + d1,
      2t*a2 - 2t*c2 + a2 + c2, 2t*a1 - 2t*c1 + a1 + c1,
      2t*c2*d2 - c2*d2, 2t*c1*d2 - c1*d2,
      2t*c2*d1 - c2*d1, 2t*c1*d1 - c1*d1)
```

Preliminary upper bound of the RLCT.

```
i13 : RLCT(I,1)
[RLCT] Warning: Output RLCT is an upper bound.

o13 = (8, 1)
```

To compute the RLCT, we transform I into a monomial ideal.

Gröbner Basis

Gröbner basis of the fiber ideal.

```
i12 : I = ideal gens gb I
o12 = ideal (a2*d2, a1*d2, b2*d1 - b1*d2, a2*d1, a1*d1,
      b2*c2, b1*c2, b2*c1, b1*c1, a2*c1 - a1*c2,
      2t*b2 - 2t*d2 + b2 + d2, 2t*b1 - 2t*d1 + b1 + d1,
      2t*a2 - 2t*c2 + a2 + c2, 2t*a1 - 2t*c1 + a1 + c1,
      2t*c2*d2 - c2*d2, 2t*c1*d2 - c1*d2,
      2t*c2*d1 - c2*d1, 2t*c1*d1 - c1*d1)
```

The red generator prevents I from being a monomial ideal. Replace it with new indeterminate β_2 via the change of variable

$$b_2 = \frac{\beta_2 - (1-2t)d_2}{1+2t}$$

which is a real-analytic isomorphism near the origin.

We can also accomplish this by introducing a new polynomial $-\beta_2 + 2tb_2 - 2td_2 + b_2 + d_2$ to the ideal and eliminating b_2.

Monomialization

Perform similar transformations to a_1, a_2, b_1, b_2.

```
i14 : R1 = QQ[t,a1,a2,b1,b2,c1,c2,d1,d2,bb1,bb2,cc1,cc2];
i15 : liftR1 = map(R1,R,{t,a1,a2,b1,b2,c1,c2,d1,d2});
i16 : I1 = (liftR1 I) + ideal(
           -bb2 + 2*t*b2 - 2*t*d2 + b2 + d2,
           -bb1 + 2*t*b1 - 2*t*d1 + b1 + d1,
           -cc2 + 2*t*a2 - 2*t*c2 + a2 + c2,
           -cc1 + 2*t*a1 - 2*t*c1 + a1 + c1);
i17 : I1 = eliminate({c1,c2,b1,b2},I1)
o17 = ideal (cc2, cc1, bb2, bb1,
             a2*d2, a1*d2, a2*d1, a1*d1)
```

Finally, we have a monomial ideal so we can compute its RLCT.

```
i18 : RLCT(I1,1)
o18 = (6, 2)
```

"Algebraic Methods for Evaluating Integrals in Bayesian Statistics"

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

(PhD dissertation, May 2011)

Singular Learning
Integral Asymptotics
RLCTs
Newton Polyhedra
Computations
- Schizo Patients
- Model Definition
- Fiber Ideal
- Gröbner Basis
- Monomialization

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 Material

Nondegenerate Ideals

Let $[\omega^\alpha]f$ denote coefficient of monomial ω^α in polynomial f.

Given $\gamma \subset \mathbb{R}^d$ and poly f, define *face poly* $f_\gamma = \sum_{\alpha\in\gamma}([\omega^\alpha]f)\omega^\alpha$.
Given $\gamma \subset \mathbb{R}^d$ and ideal I, define *face ideal* $I_\gamma = \langle f_\gamma : f \in I\rangle$.

We say I is *sos-nondegenerate* if for all compact faces $\gamma \subset \mathcal{P}(I)$, the real variety $\mathcal{V}(I_\gamma)$ does not intersect the torus $(\mathbb{R}^*)^d$.

Remark sos = sum-of-squares. Saia has similar notion of nondegeneracy for ideals of *complex* formal power series.

Proposition (L.) If $I = \langle f_1, \ldots, f_r\rangle$ and γ is a compact face of the Newton polyhedron $\mathcal{P}(I)$, then $I_\gamma = \langle f_{1\gamma}, \ldots, f_{r\gamma}\rangle$.

Proposition (L.) $\mathrm{RLCT}(I; \omega^{\tau-1}) = (1/l_\tau, \theta_\tau)$ if I is sos-ndg.

Proposition (Zwiernik) Monomial ideals are sos-ndg.

Higher Order Asymptotics

Using fiber ideals and toric blowups, we were able to compute higher order asymptotics of the statistical integral

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

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

Euler-Mascheroni constant

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

Learning Coefficients for Schizo Patients

$$Z_N = \int_\Omega \prod_{i,j} p_{ij}(\omega)^{U_{ij}} \varphi(\omega) d\omega$$

Using Watanabe's *Singular Learning Theory*,

$$-\log Z_N \approx -\sum_{i,j} U_{ij} \log q_{ij} + \lambda_q \log N - (\theta_q - 1) \log\log N$$

where the *learning coefficient* (λ_q, θ_q) is given by

$$(\lambda_q, \theta_q) = \begin{cases} (5/2, 1) & \text{if } \operatorname{rank} q = 1, \\ (7/2, 1) & \text{if } \operatorname{rank} q = 2,\ q \notin \left[\begin{smallmatrix} 0 & \times \\ \times & \times \end{smallmatrix}\right] \cup \left[\begin{smallmatrix} 0 & \times \\ \times & 0 \end{smallmatrix}\right], \\ (4, 1) & \text{if } \operatorname{rank} q = 2,\ q \in \left[\begin{smallmatrix} 0 & \times \\ \times & \times \end{smallmatrix}\right] \setminus \left[\begin{smallmatrix} 0 & \times \\ \times & 0 \end{smallmatrix}\right], \\ (9/2, 1) & \text{if } \operatorname{rank} q = 2,\ q \in \left[\begin{smallmatrix} 0 & \times \\ \times & 0 \end{smallmatrix}\right]. \end{cases}$$

Here, $q \in \left[\begin{smallmatrix} 0 & \times \\ \times & \times \end{smallmatrix}\right]$ if for some $i, j,\ q_{ii} = 0$ and $q_{ij}\, q_{ji}\, q_{jj} \neq 0$,
$q \in \left[\begin{smallmatrix} 0 & \times \\ \times & 0 \end{smallmatrix}\right]$ if for some $i, j,\ q_{ii} = q_{jj} = 0$ and $q_{ij}\, q_{ji} \neq 0$.

Model Selection (Joint work with Russell Steele)

Question: The learning coefficients (λ_q, θ_q) of a statistical model $\mathcal{M}$ depend on the true distribution q of the data which is unknown. How do we use these coefficients for model selection?

Proposal: The ML criterion and BIC may be expressed as:

$$\text{ML} = \max_{q \in \mathcal{M}} \{-\textstyle\sum_{i=1}^{N} \log q(X_i)\},$$

$$\text{BIC} = \max_{q \in \mathcal{M}} \{-\textstyle\sum_{i=1}^{N} \log q(X_i) + \frac{d}{2} \log N\}.$$

For singular models, the BIC naturally generalizes to

$$\max_{q \in \mathcal{M}} \{-\textstyle\sum_{i=1}^{N} \log q(X_i) + \lambda_q \log N - (\theta_q - 1) \log \log N\}.$$

(Maximize marginal likelihood approx over all true distributions.)

Conjecture: The generalized BIC for singular models is consistent.