Studying Model Asymptotics with Singular Learning Theory

Shaowei Lin (UC Berkeley)
shaowei@math.berkeley.edu
Joint work with Russell Steele (McGill)

13 July 2012
MMDS 2012, Stanford University
Workshop on Algorithms for Modern Massive Data Sets

Sparsity Penalties

Linear Regression

Model $Y = \omega \cdot X + \varepsilon, \quad Y \in \mathbb{R}, \quad \omega, X \in \mathbb{R}^d, \quad \varepsilon \in \mathcal{N}(0,1)$

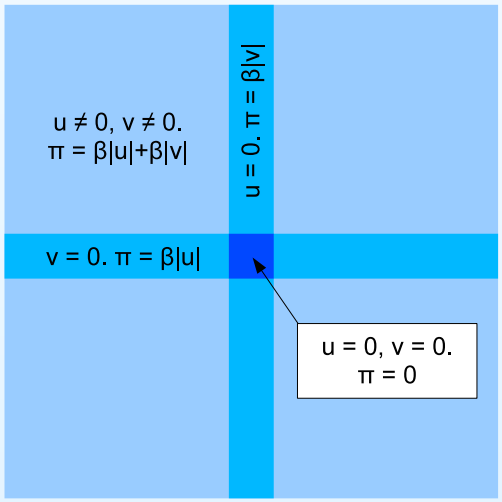
Data $(Y_1, X_1), \ldots, (Y_N, X_N)$

Least squares $\min_\omega \sum_{i=1}^N |Y_i - \omega \cdot X_i|^2$

Penalized regression $\min_\omega \sum_{i=1}^N |Y_i - \omega \cdot X_i|^2 + \pi(\omega)$

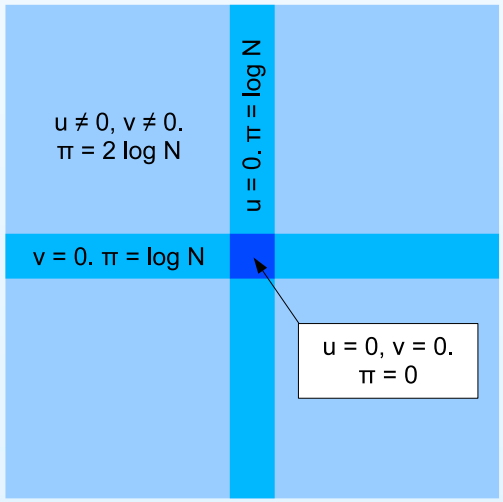
LASSO

$\pi(\omega) = |\omega|_1 \cdot \beta$



Bayesian Info Criterion (BIC)

$\pi(\omega) = |\omega|_0 \cdot \log N$



Parameter space is partitioned into regions (submodels).

Bayesian Information Criterion

- Given region Ω of parameters and a prior $\varphi(\omega)d\omega$ on Ω, the *marginal likelihood* of the data is proportional to

$$Z_N = \int_\Omega e^{-Nf(\omega)}\,\varphi(\omega)d\omega$$

where $f(\omega) = \frac{1}{2N}\sum_{i=1}^N |Y_i - \omega \cdot X_i|^2$.

- *Laplace approximation*: Asymptotically as sample size $N \to \infty$,

$$-\log Z_N \approx Nf(\omega^*) + \frac{d}{2}\log N + O(1)$$

where $\omega^* = \operatorname{argmin}_{\omega\in\Omega} f(\omega)$ and $d = \dim \Omega$.

- Studying model asymptotics allows us to derive the BIC.
But Laplace approx only works when the model is regular.
Many models in machine learning are *singular*,
e.g. mixtures, neural networks, hidden variables.

Integral Asymptotics

Estimating Integrals

Generally, there are three ways to estimate statistical integrals.

1. *Exact methods*
   Compute a closed form formula for the integral,
   e.g. (Lin·Sturmfels·Xu, 2009).

2. *Numerical methods*
   Approximate using Markov Chain Monte Carlo (MCMC)
   and other sampling techniques.

3. *Asymptotic methods*
   Analyze how the integral behaves for large samples.

Real Log Canonical Threshold

Asymptotic theory (Arnol'd·Guseĭn-Zade·Varchenko, 1985) states that for a Laplace integral,

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

asymptotically as $N \to \infty$ for some positive constants C, λ, θ and where $f^* = \min_{\omega \in \Omega} f(\omega)$.

The pair (λ, θ) is the *real log canonical threshold* of $f(\omega)$ with respect to the measure $\varphi(\omega) d\omega$.

Geometry of the Integral

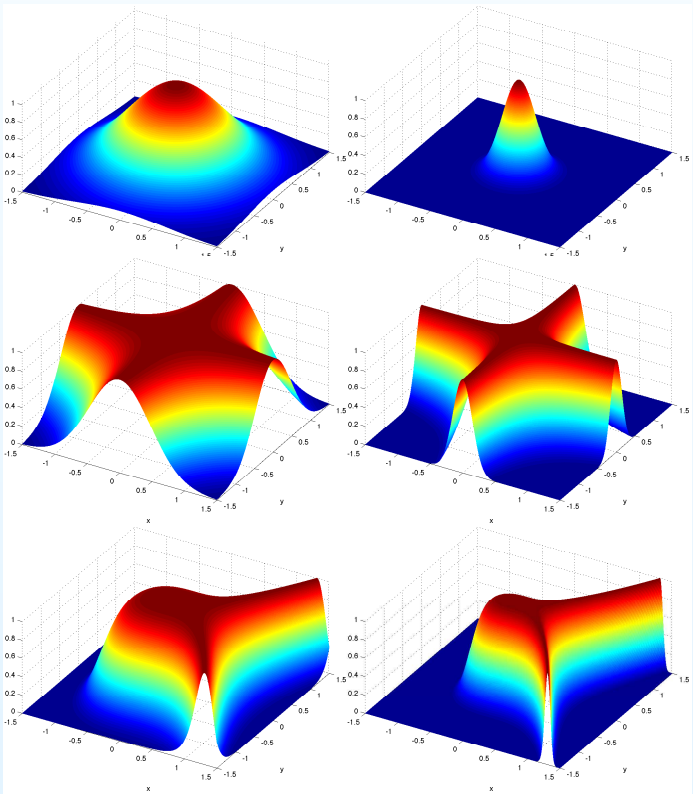
$$Z(N) = \int_\Omega e^{-Nf(\omega)}\varphi(\omega)d\omega \approx e^{-Nf^*} \cdot CN^{-\lambda}(\log N)^{\theta-1}$$

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 $N = 10$

Desingularizations

Let $\Omega \subset \mathbb{R}^d$ and $f : \Omega \to \mathbb{R}$ real 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. ρ is a proper real analytic map that is an isomorphism onto the subset $\{\omega \in \Omega : f(\omega) \neq 0\}$.

  3. 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.

Algorithm for Computing RLCTs

- We know how to find RLCTs of *monomial functions* (AGV, 1985).

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

  where $\lambda = \min_i \frac{\tau_i+1}{\kappa_i}$, $\theta = |\{i : \frac{\tau_i+1}{\kappa_i} = \lambda\}|$.

- To compute the RLCT of any function $f(\omega)$:

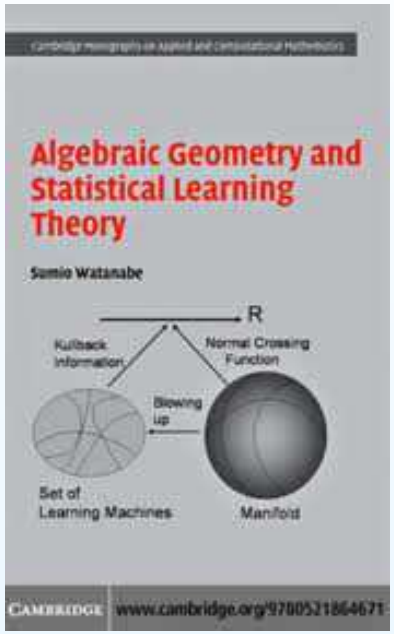
  1. Find minimum f^* of f over Ω.
  2. Find a desingularization ρ for $f - f^*$.
  3. Use AGV Theorem to find (λ_i, θ_i) on each patch U_i.
  4. $\lambda = \min\{\lambda_i\}$, $\theta = \max\{\theta_i : \lambda_i = \lambda\}$.

- The difficult part is finding a desingularization, e.g (Bravo·Encinas·Villamayor, 2005).

Singular Learning Theory

Sumio Watanabe

Sumio Watanabe



Heisuke Hironaka

In 1998, Sumio Watanabe discovered how to study the asymptotic behavior of singular models. His insight was to use a deep result in algebraic geometry known as *Hironaka's Resolution of Singularities*.

Heisuke Hironaka proved this celebrated result in 1964.
His accomplishment won him the Field's Medal in 1970.

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

Suppose samples $X_1, \ldots, X_N$ drawn from *true distribution* $q \in \mathcal{M}$.

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

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

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)

If $\rho : U \to \Omega$ desingularizes $K(\omega)$, then on each patch U_i,

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

where $\xi_N(\mu)$ converges in law to a Gaussian process on U.

For regular models, this is a *Central Limit Theorem*.

Learning Coefficient

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

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

where λ_q, θ_q describe the asymptotics of the deterministic integral

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

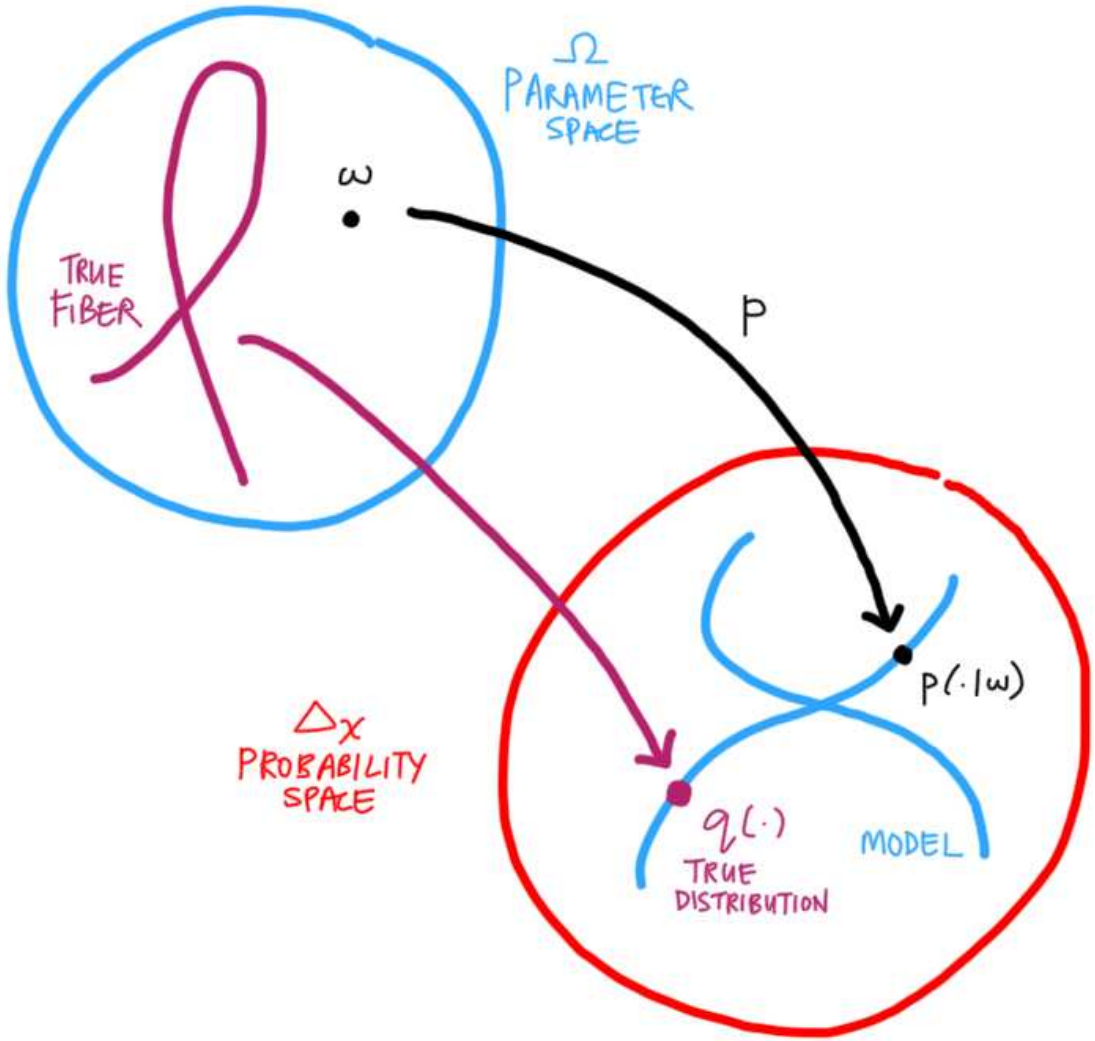
For regular models, this is the Bayesian Information Criterion.

Various names for (λ_q, θ_q):

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

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

Geometry of Singular Models



AIC and DIC

Bayes generalization error B_N. The Kullback-Leibler distance from the true distribution $q(x)$ to the predictive distribution $p(x|D)$.

Asymptotically, B_N is equivalent to

- Akaike Information Criterion for regular models

$$\mathrm{AIC} = -\sum_{i=1}^{N} \log p(X_i|\omega^*) + d$$

- Akaike Information Criterion for singular models

$$\mathrm{AIC} = -\sum_{i=1}^{N} \log p(X_i|\omega^*) + 2(\textit{singular fluctuation})$$

Numerically, B_N can be estimated using MCMC methods.

- Deviance Information Criterion for regular models

$$\mathrm{DIC} = \mathbb{E}_X[\log p(X|\mathbb{E}_\omega[\omega])] - 2\,\mathbb{E}_\omega[\mathbb{E}_X[\log p(X|\omega)]]$$

- Widely Applicable Information Criterion for singular models

$$\mathrm{WAIC} = \mathbb{E}_X[\log \mathbb{E}_\omega[p(X|\omega)]] - 2\,\mathbb{E}_\omega[\mathbb{E}_X[\log p(X|\omega)]]$$

Real Log Canonical Thresholds

Sparsity Penalty

Local RLCTs. Given $x \in \Omega$, there exist a small nbhd $\Omega_x \subset \Omega$ of x and exponents (λ_x, θ_x) such that for all nbhds $U \subset \Omega_x$ of x,

$$\int_U e^{-Nf(\omega)}\,\varphi(\omega)d\omega \;\approx\; CN^{-\lambda_x}(\log N)^{\theta_x - 1}.$$

Maximum likelihood estimation. Find $\min_{\omega\in\Omega} \ell_N(\omega)$ where

$$\ell_N(\omega) = -\sum_{i=1}^{N} \log p(X_i|\omega).$$

Sparsity penalty for MLE. Find $\min_{\omega\in\Omega} \ell_N(\omega) + \pi(\omega)$ where

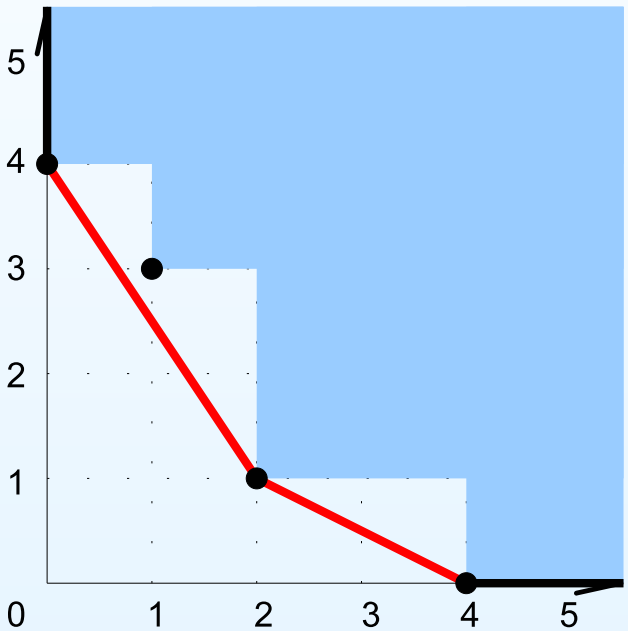
$$\pi(\omega) = \lambda_\omega \log N - (\theta_\omega - 1)\log\log N.$$

This is a generalization of the BIC to singular models. It can also teach us how to penalize parameters appropriately in LASSO.

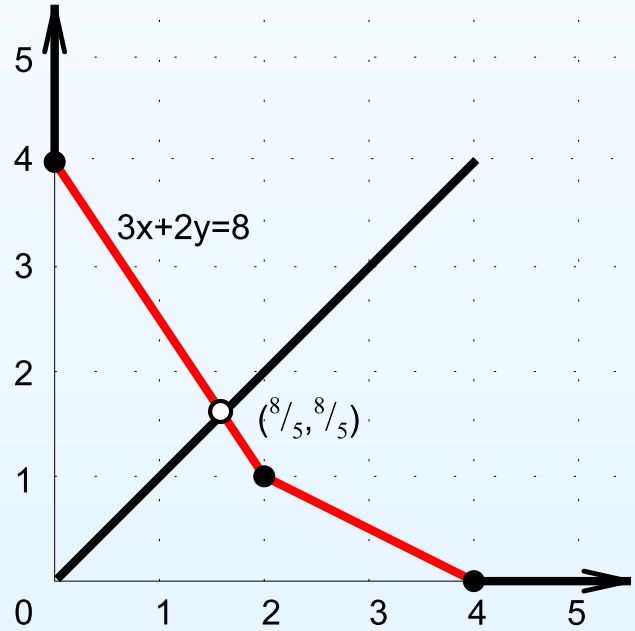
Newton Polyhedra

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

Newton polyhedron



τ-distance

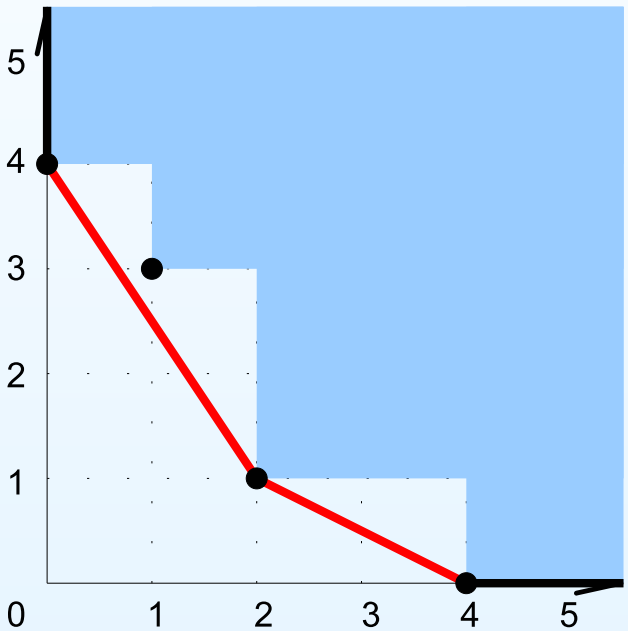


The τ-distance is $l_\tau = 8/5$ and the multiplicity is $\theta_\tau = 1$.

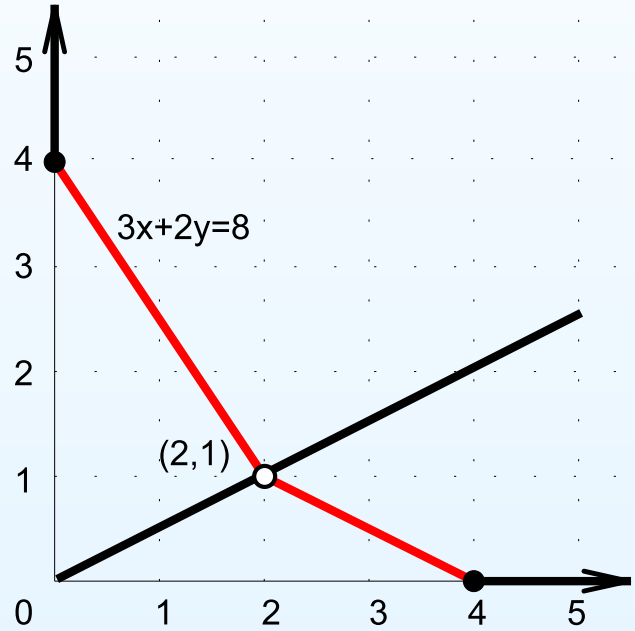
Newton Polyhedra

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

Newton polyhedron



τ-distance



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

Upper Bounds for RLCTs

Given a power series $f(\omega) \subset \mathbb{R}[\omega_1, \ldots, \omega_d]$,

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

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

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 RLCTs at the origin

If l_τ is the τ-distance of $\mathcal{P}(f)$ and θ_τ is its multiplicity, then the RLCT (λ_0, θ_0) of f with respect to $\omega^{\tau-1}d\omega$ satisfies

$$(\lambda_0, \theta_0) \leq (1/l_\tau, \theta_\tau).$$

Equality occurs when f is a sum of squares of monomials.

Thank you!

"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. 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.
4. S. LIN, B. STURMFELS AND Z. XU: Marginal likelihood integrals for mixtures of independence models, *J. Mach. Learn. Res.* **10** (2009) 1611–1631.
5. S. LIN: Algebraic methods for evaluating integrals in Bayesian statistics, PhD dissertation, Dept. Mathematics, UC Berkeley (2011).
6. S. WATANABE: *Algebraic Geometry and Statistical Learning Theory*, Cambridge Monographs on Applied and Computational Mathematics **25**, Cambridge University Press, Cambridge, 2009.

Supplementary Material

Higher Order Asymptotics

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

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

$$\sqrt{\frac{\pi}{8}} N^{-\frac{1}{2}} \log N \qquad -\sqrt{\frac{\pi}{8}} \left(\frac{1}{\log 2} - 2 \log 2 - \gamma \right) N^{-\frac{1}{2}}$$

$$-\frac{1}{4} N^{-1} \log N \qquad +\frac{1}{4} \left(\frac{1}{\log 2} + 1 - \gamma \right) N^{-1}$$

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

$$-\frac{1}{24} N^{-2} \; + \; \cdots$$

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 $q_{ij}\, q_{ji} \neq 0$.