Understanding the Curse of Singularities in Machine Learning

Shaowei Lin (UC Berkeley)
shaowei@math.berkeley.edu

31 August 2012
Autonomous Group Meeting
EECS, UC Berkeley

Sparsity Penalties

Linear Regression

Random variables $Y \in \mathbb{R}, X \in \mathbb{R}^d$ satisfy

$$Y = \omega \cdot X + \varepsilon$$

Parameters $\omega \in \mathbb{R}^d$; noise $\varepsilon \in \mathcal{N}(0,1)$; data $(Y_i, X_i), i = 1 .. N$.

- Commonly computed quantities

MLE	$\operatorname{argmin}_\omega \sum_{i=1}^N \lvert Y_i - \omega \cdot X_i \rvert^2$
Penalized MLE	$\operatorname{argmin}_\omega \sum_{i=1}^N \lvert Y_i - \omega \cdot X_i \rvert^2 + \pi(\omega)$

- Commonly used penalties

LASSO	$\pi(\omega) = \lvert\omega\rvert_1 \cdot \beta$
Bayesian Info Criterion (BIC)	$\pi(\omega) = \lvert\omega\rvert_0 \cdot \log N$
Akaike Info Criterion (AIC)	$\pi(\omega) = \lvert\omega\rvert_0 \cdot 2$

- Common applications

  Parameter estimation

  Model selection (e.g. which entries in ω are nonzero?)

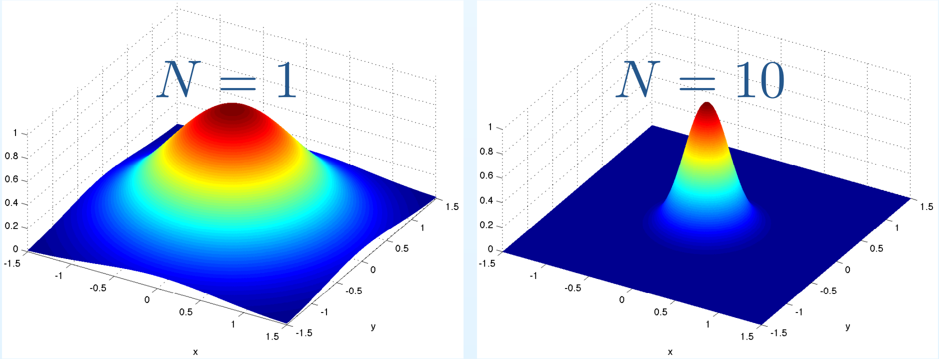
Parameter Estimation

- *Parameter estimation is a form of model selection!*
- MLE: given true parameter $u \in \mathbb{R}^d$, likelihood of data is
$$L(u) = \frac{1}{\sqrt{2\pi}} \exp(-\frac{1}{2}\sum_{i=1}^{N} |Y_i - u \cdot X_i|^2)$$
- LASSO: put Laplacian prior on the parameter space. Given true parameter $u \in \mathbb{R}^d$, likelihood of data is
$$L(u) = \frac{1}{\sqrt{2\pi}} \exp(-\frac{1}{2}\sum_{i=1}^{N} |Y_i - u \cdot X_i|^2) \exp(-\frac{1}{2}\beta |u|_1)$$
- *Integrated likelihood*: put prior $\varphi(\omega)$ on small neighborhood Ω_u of true parameter $u \in \mathbb{R}^d$. Integrated likelihood of data is
$$Z(u) = \int_{\Omega_u} \frac{1}{\sqrt{2\pi}} \exp(-\frac{1}{2}\sum_{i=1}^{N} |Y_i - \omega \cdot X_i|^2)\, \varphi(\omega) d\omega$$

Laplace Approximation

- Let $f(\omega) = \frac{1}{2N}\sum_{i=1}^{N} |Y_i - \omega \cdot X_i|^2$ so we can write
$$Z(u) = \frac{1}{\sqrt{2\pi}} \int_{\Omega_u} e^{-Nf(\omega)}\, \varphi(\omega) d\omega.$$

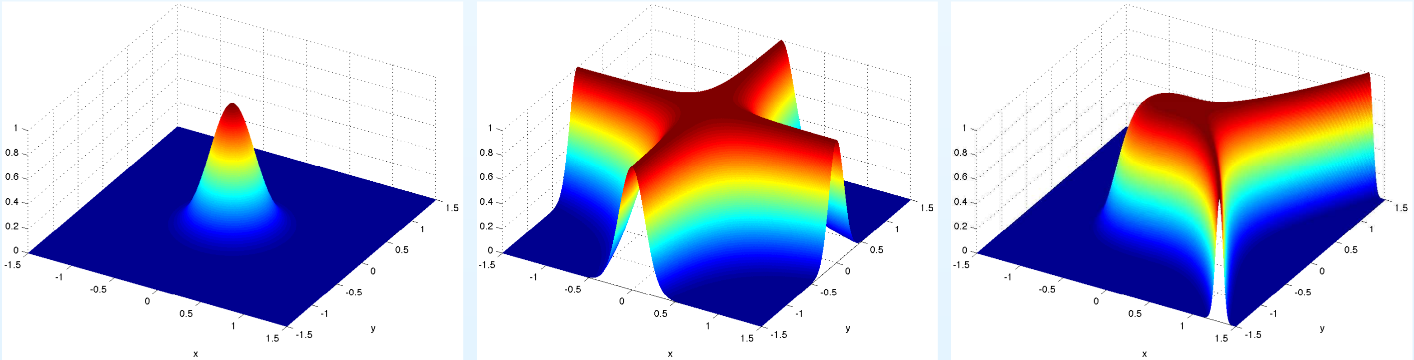
- *Laplace approximation*: If $f(\omega)$ is uniquely minimized at u and the Hessian satisfies $\det \partial^2 f(u) \neq 0$, then asymptotically
$$-\log Z(u) \approx Nf(u) + \frac{\dim \Omega_u}{2} \log N + O(1)$$
as sample size $N \to \infty$. This approximation gives us the BIC.



Graphs of $e^{-Nf(\omega)}$ for different N. Integral = Volume under graph.

Curse of Singularities

- The AIC, which is based on the *Bayes generalization error*, can also be derived using integral asymptotics.
- For *smooth* models i.e. $\det \partial^2 f(u) \neq 0$, Laplace approx works well even if parameter space $\mathbb{R}^d$ has high dimension.
- But many models in machine learning are *singular*, e.g. mixtures, neural networks, hidden variables.
- How do we study the asymptotics of integrals with singularities?

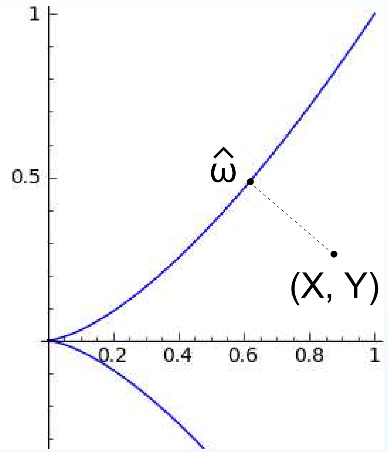


Example: Singular Model

$$X \sim \mathcal{N}(\omega^2, 1), \quad Y \sim \mathcal{N}(\omega^3, 1)$$

data $(X_i, Y_i), i = 1 \,..\, N$

parameter $\omega \in \mathbb{R}$, mean $(\bar{X}, \bar{Y})$



- MLE: $\operatorname{argmin}_\omega |\omega^2 - \bar{X}|^2 + |\omega^3 - \bar{Y}|^2$
  BIC performs poorly when MLE is close to 0.

- Put prior $\varphi(\omega)$ on small nbhd Ω_u of true parameter $u \in \mathbb{R}$.

$$Z(u) = \frac{1}{2\pi} \int_{\Omega_u} \exp\Big(-\frac{1}{2} \sum_{i=1}^{N} |\omega^2 - X_i|^2 + |\omega^3 - Y_i|^2\Big)\, \varphi(\omega) d\omega$$

- According to *Singular Learning Theory*, asymptotically

$$-\log Z(u) \approx \frac{1}{2} \sum_{i=1}^{N} (u^2 - X_i)^2 + (u^3 - Y_i)^2 + \pi(u) + O_p(1)$$

where $\pi(u) = \frac{1}{4} \log N$ if $u = 0$; otherwise $\pi(u) = \frac{1}{2} \log N$.

Higher Order Asymptotics

Higher order terms in the asymptotics of the integral can be derived by resolving the singularities. For example,

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

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

Euler-Mascheroni constant

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

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.

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

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

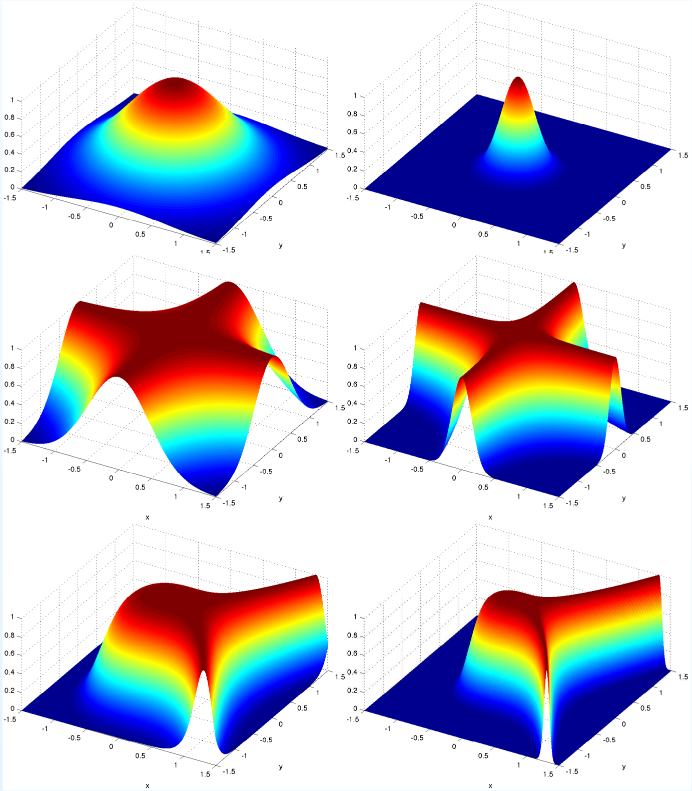
$$Z(N) = \int_{\Omega} e^{-Nf(\omega)} \varphi(\omega) d\omega \;\approx\; e^{-Nf^*} \cdot CN^{-\lambda} (\log N)^{\theta - 1}$$

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

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

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

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



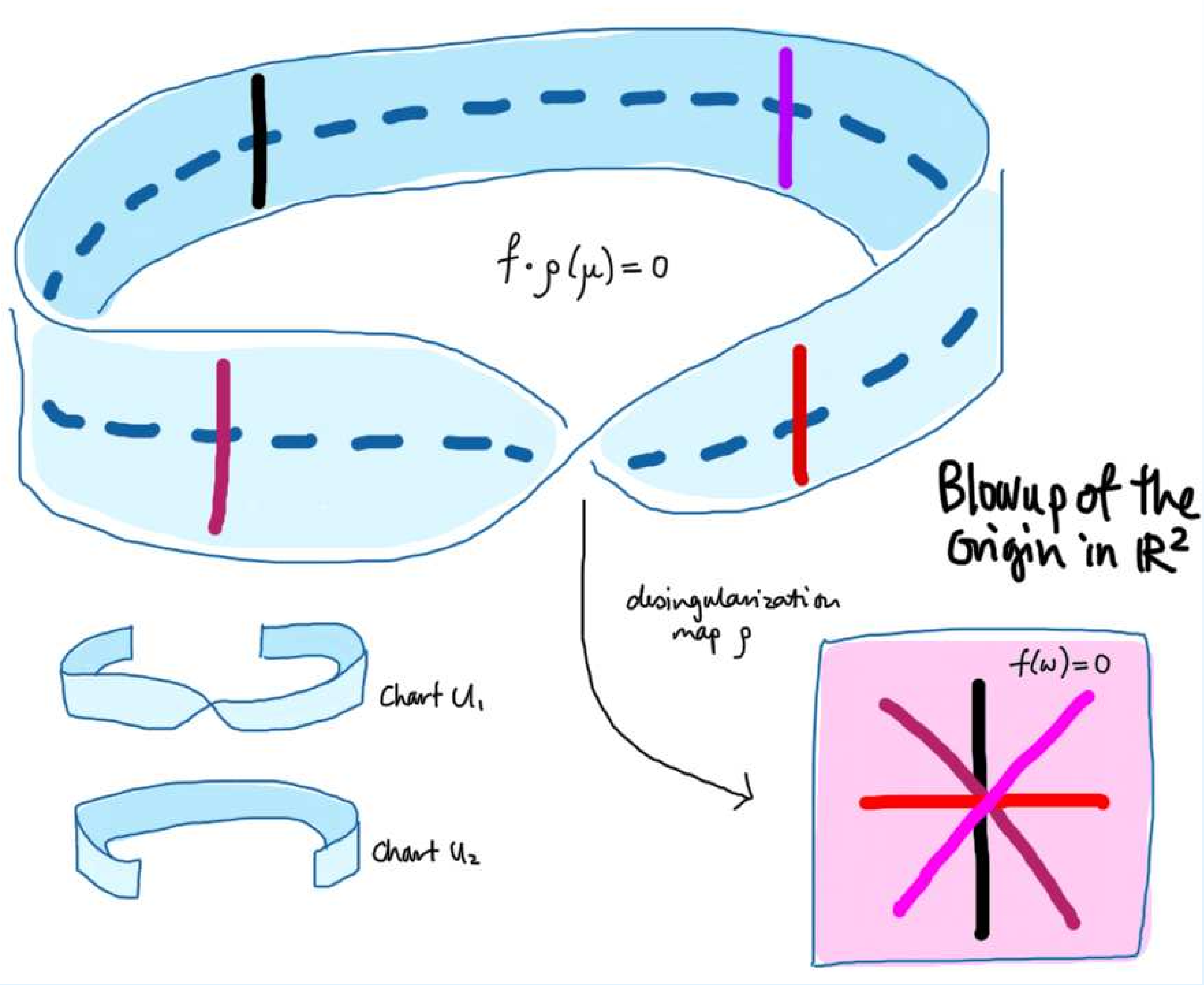
Graphs of integrand $e^{-Nf(x,y)}$ for $N = 1$ and $N = 10$

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.

Desingularizations



Algorithm for Computing RLCTs

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

$$\int_\Omega e^{-Na(\mu)\mu^\kappa} b(\mu)\mu^\tau d\mu \approx CN^{-\lambda}(\log N)^{\theta-1}$$

  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 $f(\omega)$:

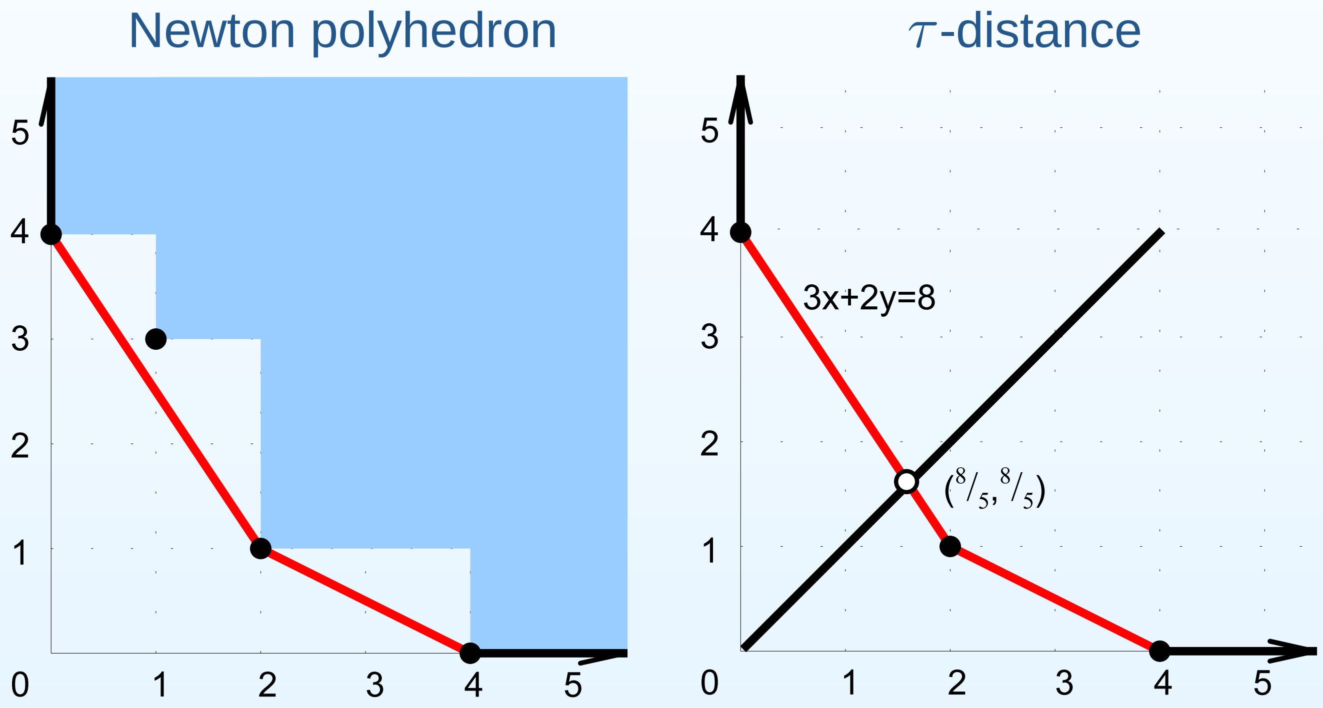
  1. Find minimum f^* of f over Ω.
  2. Find a desingularization ρ for $f - f^*$.
  3. Use AGV Theorem to find (λ_i, θ_i) on each patch U_i.
  4. $\lambda = \min\{\lambda_i\}$, $\theta = \max\{\theta_i : \lambda_i = \lambda\}$.

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

- One method for estimating RLCTs uses *Newton polyhedra*.

Newton Polyhedra

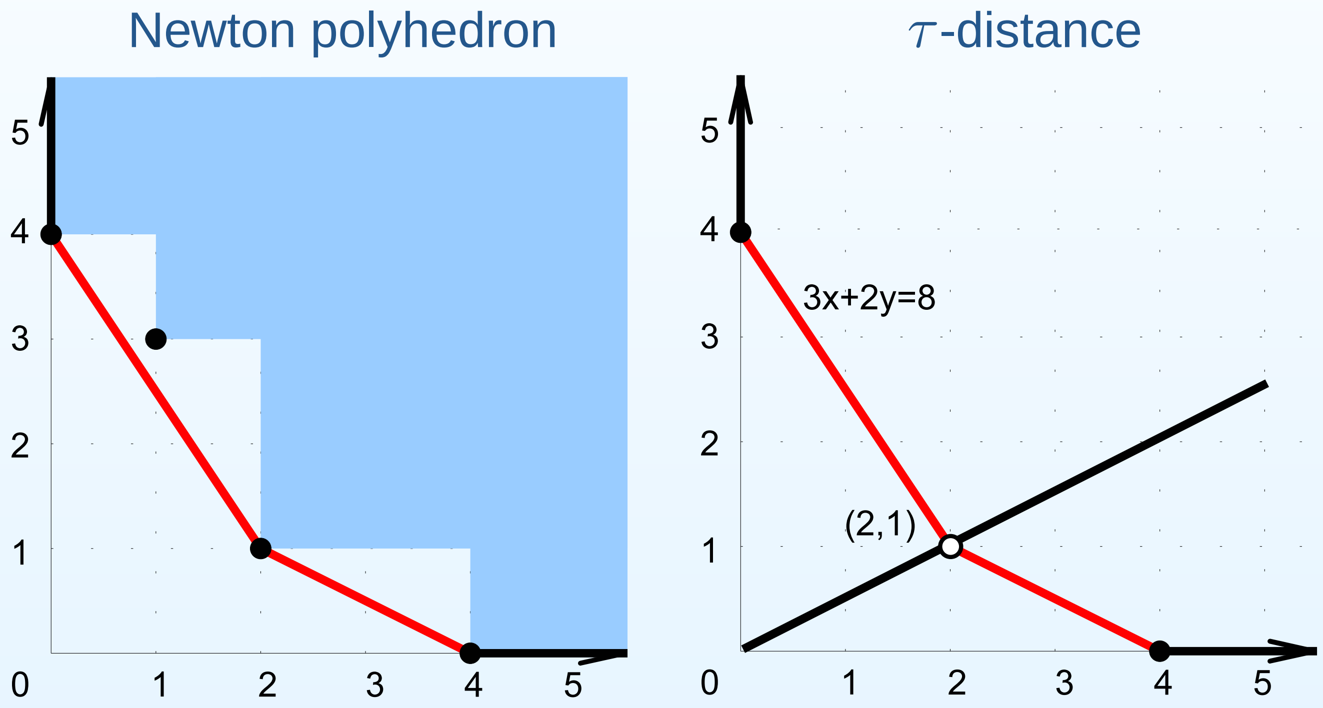
e.g. Let $f(x, y) = x^4 + x^2y + xy^3 + y^4$ and $\tau = (1, 1)$.



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 $\tau = (2,1)$.



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

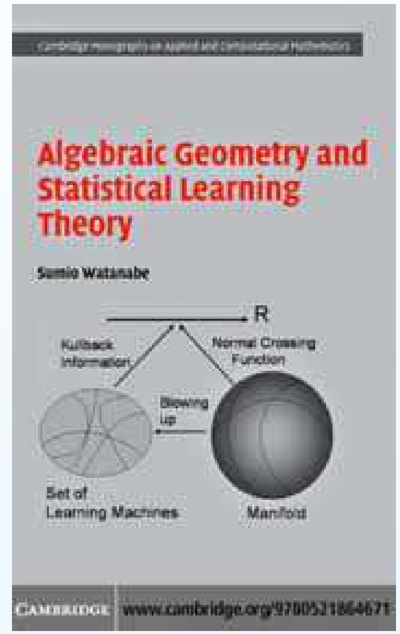
$$(\lambda_0, \theta_0) \leq (1/l_\tau, \theta_\tau).$$

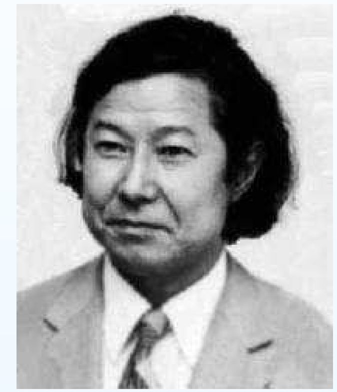
Equality occurs when f is a *sum of squares of monomials*.

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 $\mathcal{X}$

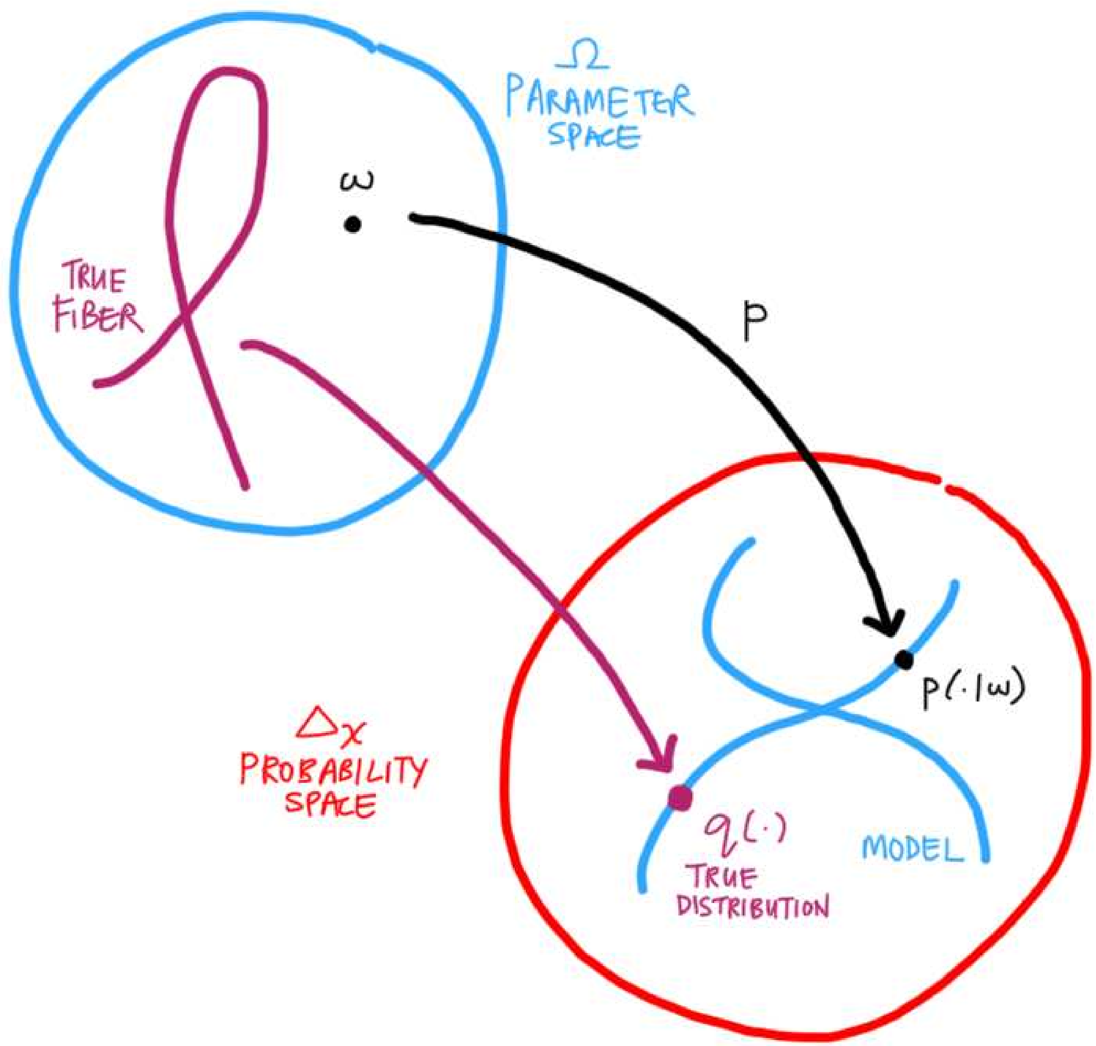
$\mathcal{M} \subset \Delta$	statistical model, image of $p : \Omega \to \Delta$
Ω	parameter space
$p(x\|\omega)dx$	distribution at $\omega \in \Omega$
$\varphi(\omega)d\omega$	prior distribution on Ω

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

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)

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

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

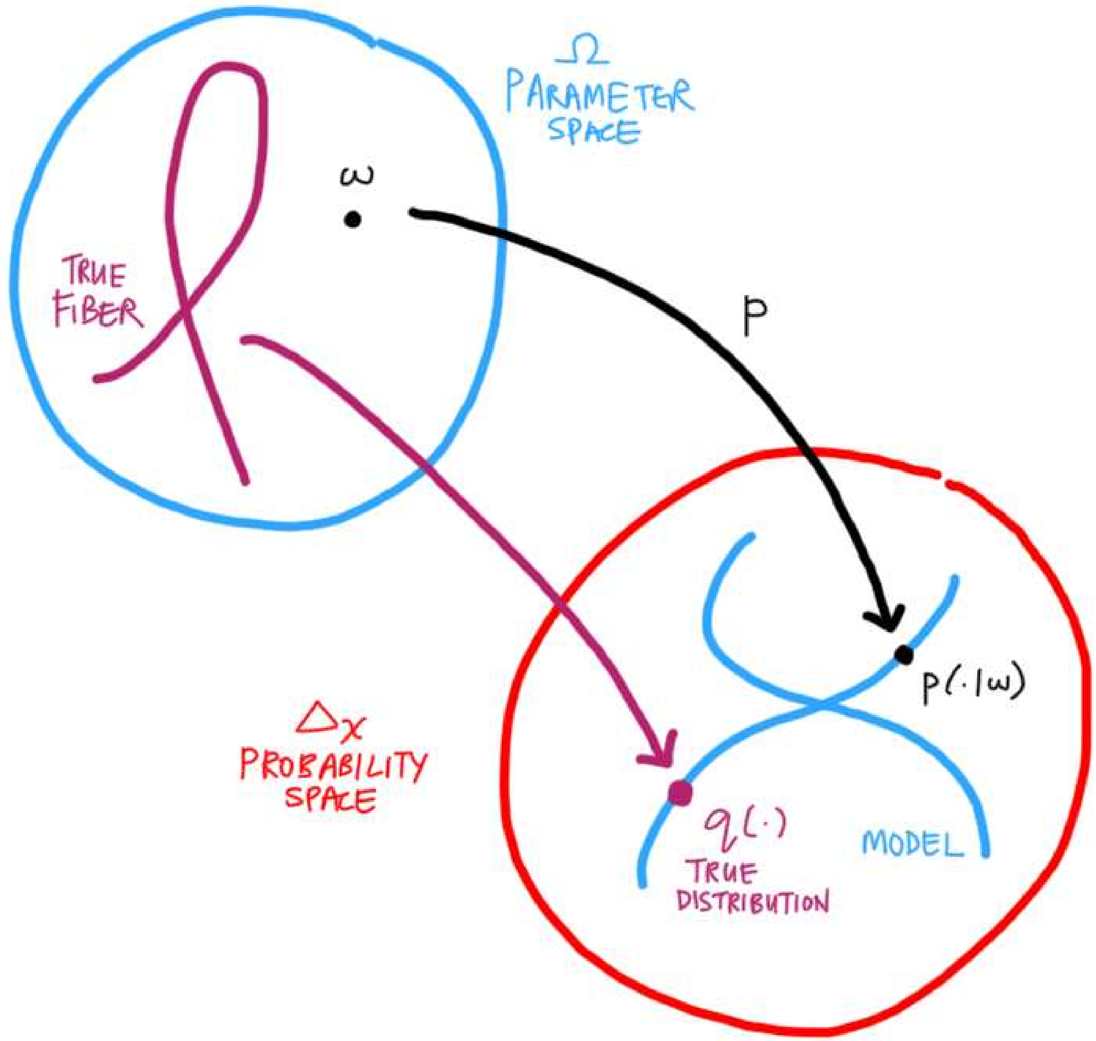
For regular models, this is the Bayesian Information Criterion.

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

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

- Akaike Information Criterion for singular models

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

Numerically, B_N can be estimated using MCMC methods.

- Deviance Information Criterion for regular models

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

- Widely Applicable Information Criterion for singular models

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

Sparsity Penalty

- *Local RLCTs*: given $u \in \Omega$, there exist a small nbhd Ω_u of u and exponents (λ_u, θ_u) such that for all smaller nbhds U,
$$\int_U e^{-Nf(\omega)}\,\varphi(\omega)d\omega \;\approx\; CN^{-\lambda_u}(\log N)^{\theta_u - 1}.$$

- *Maximum likelihood estimation*: $\operatorname{argmin}_{u\in\Omega} \ell(u)$ where
$$\ell(u) = -\sum_{i=1}^{N} \log p(X_i|u).$$

- *Sparsity penalty for MLE*: $\operatorname{argmin}_{u\in\Omega} \ell(u) + \pi(u)$ where
$$\pi(u) = \lambda_u \log N - (\theta_u - 1)\log\log N.$$

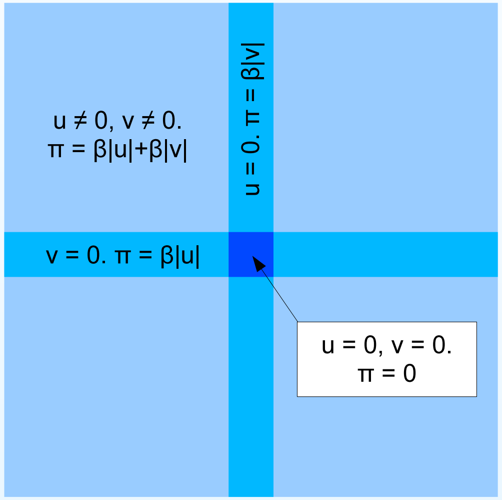
- This is a generalization of the BIC to singular models.

Open Problems

- How do we generalize LASSO to singular models?

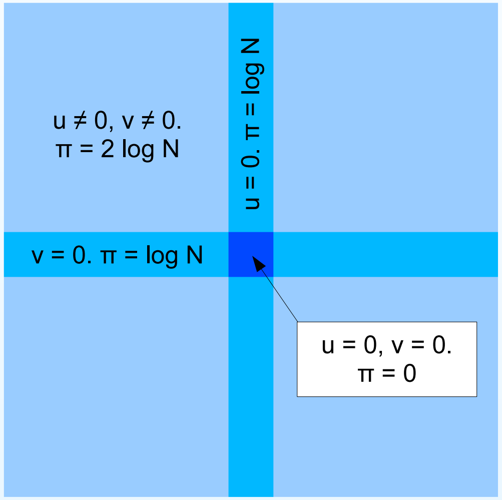
LASSO

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



Bayesian Info Criterion (BIC)

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



(Parameter space partitioned into regions with different weights.)

- How do we use RLCTs to improve MCMC techniques?

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.