

# Bayesian Statistics and Singular Learning Theory

Shaowei Lin (UC Berkeley)

`shaowei@math.berkeley.edu`

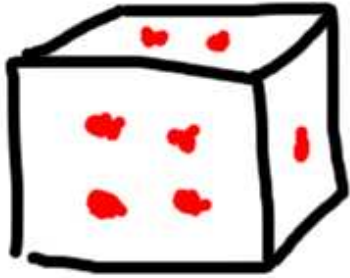
19 Oct 2011

UC Berkeley Algebraic Statistics Seminar

# **Bayesian Statistics**

Two main *interpretations*  
of probability theory:  
*Frequentist* and *Bayesian*.

These interpretations do not affect  
the *correctness* of probability theory,  
but they greatly affect  
the *statistical methodology*.

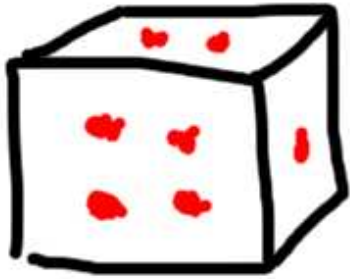


## FREQUENTIST:

each number occurs about  $N/6$  times out of  $N$  throws of the die.

## BAYESIAN:

No, not really. That's only what you BELIEVE about the die.

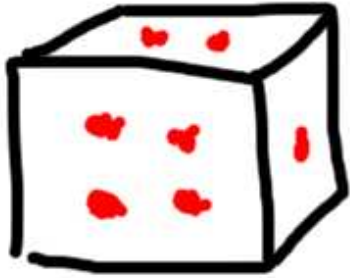


## FREQUENTIST:

surely, the die has some inherent probabilities and our purpose is to discover them!!

## BAYESIAN:

Nope! These probabilities are not inherent. A die is a die. That's it. But as we observe the die, our belief about its outcomes changes too.



FREQUENTIST:  
you are insane!

BAYESIAN:

Not really. That's just what  
you BELIEVE  $\hat{\theta}$

# Bayes' Rule

$$\boxed{P(A|B)} = \frac{P(B|A)}{P(B)} \boxed{P(A)}$$

posterior  
"new belief"

prior  
"old belief"

$$P(B) = \sum_i P(B|A_i)P(A_i) \quad \text{normalization constant}$$

# Bayes' Rule

**Example:** A coin, which we believe with

$\frac{9}{10}$  probability to be fair such that  $P(H) = \frac{1}{2}$ ,  $P(T) = \frac{1}{2}$ ,  
 $\frac{1}{10}$  probability to be biased such that  $P(H) = \frac{3}{4}$ ,  $P(T) = \frac{1}{4}$ .

After observing 8 heads and 2 tails,

$$\begin{aligned} P(\text{fair}|\text{data}) &= \frac{P(\text{data}|\text{fair})}{P(\text{data})} P(\text{fair}) \\ &= \frac{\left(\frac{1}{2}\right)^8 \left(\frac{1}{2}\right)^2 \left(\frac{9}{10}\right)}{\left(\frac{1}{2}\right)^8 \left(\frac{1}{2}\right)^2 \left(\frac{9}{10}\right) + \left(\frac{3}{4}\right)^8 \left(\frac{1}{4}\right)^2 \left(\frac{1}{10}\right)} = \frac{1024}{1753} \approx 0.584 \end{aligned}$$

Old belief = 0.900  $\longrightarrow$  New belief = 0.584

# Statistical Model

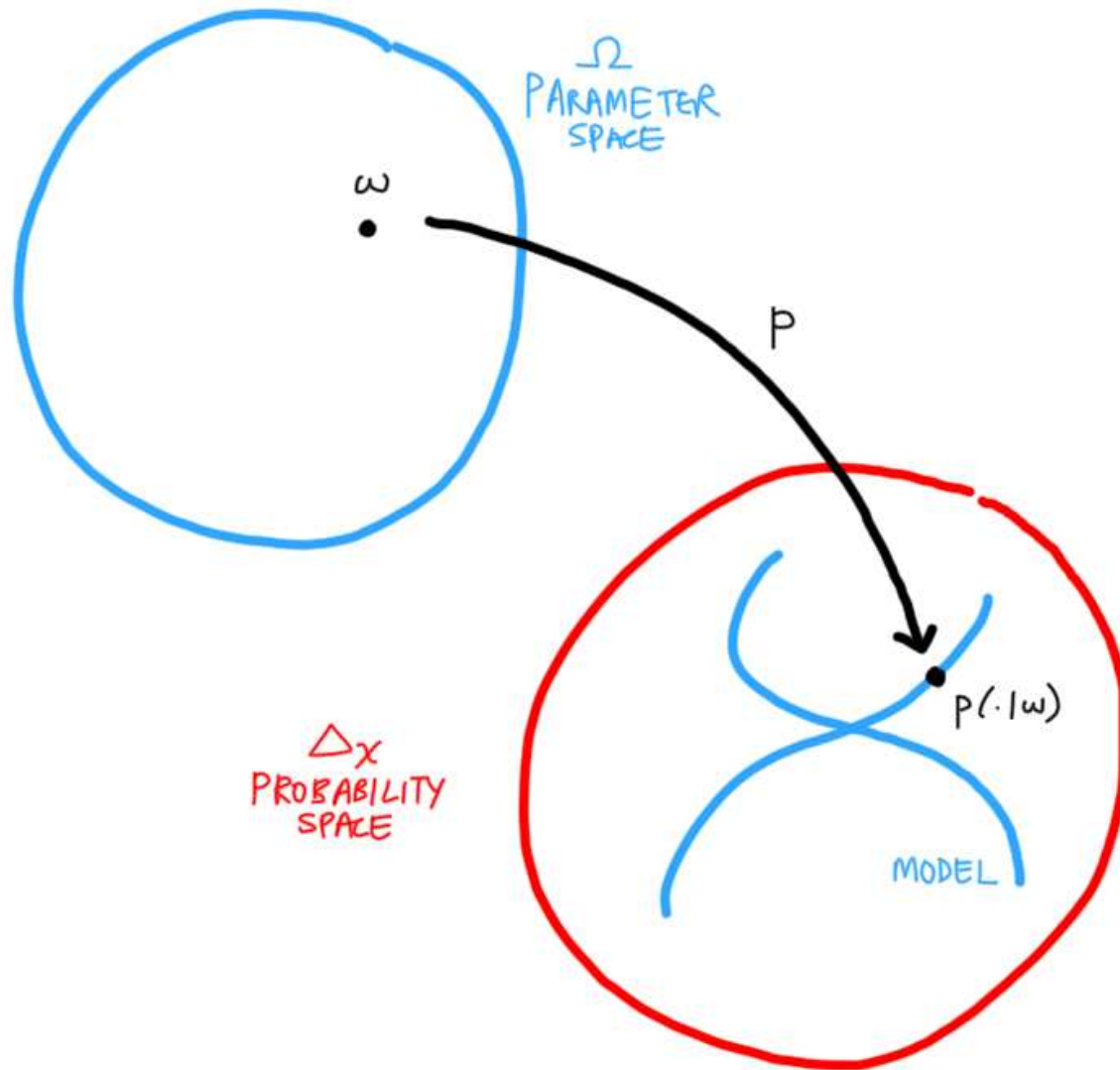
Let  $X$  be a random variable with state space  $\mathcal{X}$  (e.g.  $\{1, 2, \dots, k\}$ ,  $\mathbb{R}^k$ ).  
Let  $X_1, \dots, X_N$  be  $N$  independent random samples of  $X$ .

In the previous example, we studied *two* probability distributions, and our beliefs about each of them. More generally, let us study a *family*  $\mathcal{M}$  of probability distributions parametrized by some space  $\Omega$ . Such a family is called a *statistical model*.

In *algebraic statistics*, we think of the statistical model as a map from the *parameter space*  $\Omega$  to the *probability space*  $\Delta_{\mathcal{X}}$ .  
Let  $p(x|\omega)dx$  denote the distribution corresponding to  $\omega \in \Omega$ .

To each  $\omega \in \Omega$ , we associate a *belief* about the parameter. This belief is called the *prior distribution* and is denoted by  $\varphi(\omega)d\omega$ . Also, if we are studying more than one model, say  $\mathcal{M}_1$  and  $\mathcal{M}_2$ , we may associate *priors*  $p(\mathcal{M}_1)$  and  $p(\mathcal{M}_2)$  to each of them.

# Statistical Model



# Posterior distributions

Recall *prior* = “old belief”, *posterior* = “new belief”.

*Posterior distribution on  $\Omega$*

$$p(\omega|X_1, \dots, X_N) = \frac{\left(\prod_{i=1}^N p(X_i|\omega)\right) \varphi(\omega)}{\int_{\Omega} \left(\prod_{i=1}^N p(X_i|\omega)\right) \varphi(\omega) d\omega}$$

*Posterior distribution on models  $\mathcal{M}_1, \mathcal{M}_2$*

$$p(\mathcal{M}|X_1, \dots, X_N) = p(\mathcal{M}) \int_{\Omega} \left(\prod_{i=1}^N p(X_i|\omega)\right) \varphi(\omega) d\omega$$

# Model Selection

## **Frequentist:**

“I want to find the *best parameter*  $\omega$  which describes the data.”

**Maximum Likelihood:** Pick the model that maximizes

$$\max_{\omega} \prod_{i=1}^N p(X_i|\omega)$$

## **Bayesian:**

“Which model do I *believe in the most* after the observing data?”

**Marginal Likelihood:** Pick the model that maximizes

$$p(\mathcal{M}) \int_{\Omega} \left( \prod_{i=1}^N p(X_i|\omega) \right) \varphi(\omega) d\omega$$

An *important and difficult* problem in Bayesian statistics is the *accurate approximation* of the marginal likelihood integral.

# **Singular Learning Theory**

A statistical model is *regular* if it is identifiable and its Fisher information matrix is positive 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*.

# The True Distribution

Let  $X$  be a random variable.

In *statistical learning theory*, we are interested in using the data  $X_1, \dots, X_N$  to select a model  $\mathcal{M}$  that best describes  $X$ . For this purpose, many *model selection criteria* (e.g. maximum likelihood, marginal likelihood, AIC, BIC) have been designed.

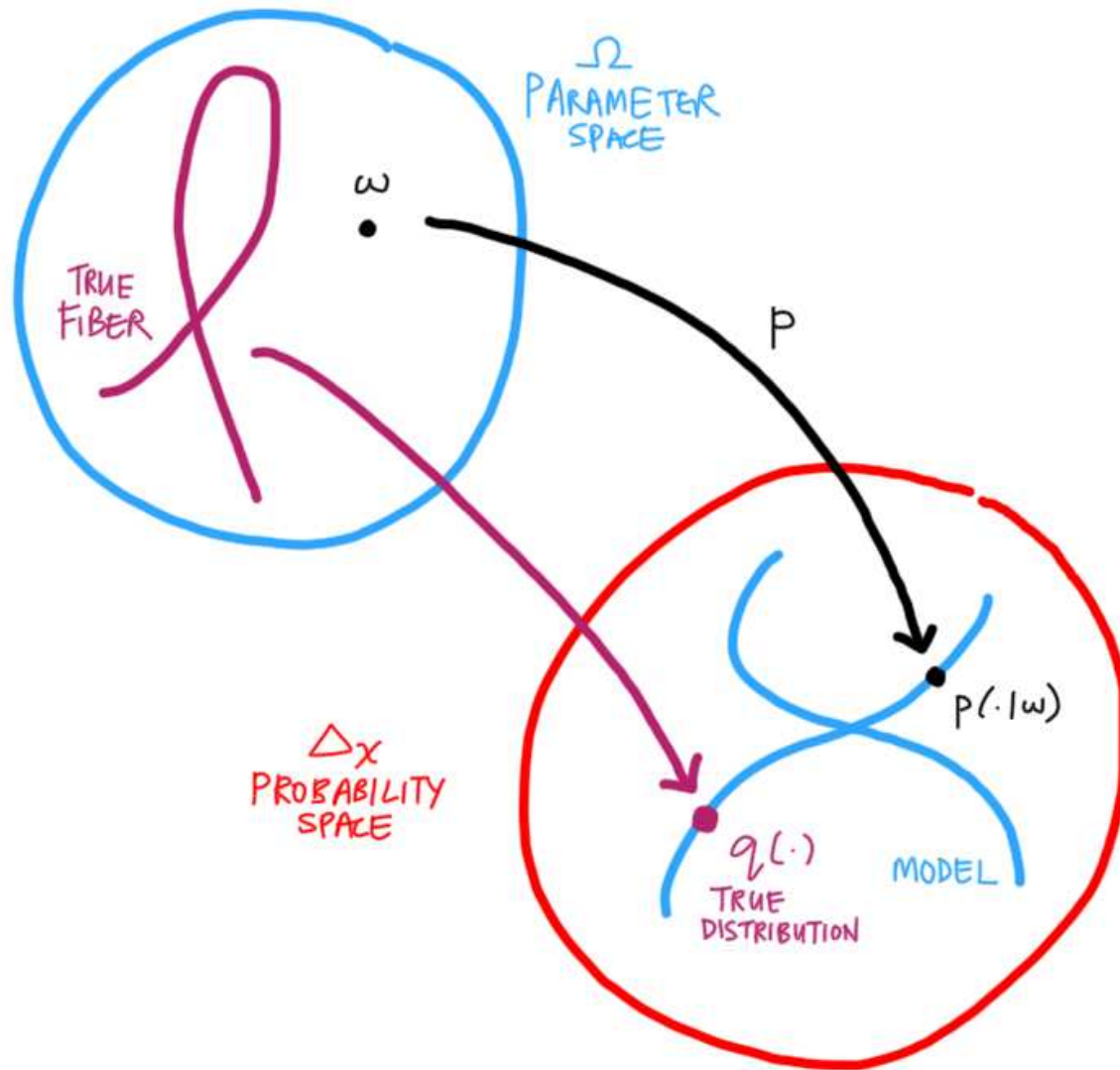
It is important to analyze how these criteria behave as the *number of samples grow large*. For this purpose, we need to assume that  $X$  has a *true distribution*  $q(x)dx$ . Given a model, let the *true fiber* be the set of all parameters  $\omega \in \Omega$  which map to the true distribution.

## Remark:

The word “true distribution” disturbs the Bayesian in us, but we disregard such philosophical objections for now.

I like to think of  $X$  as a computer (black box) producing outputs  $X_1, \dots, X_N$  according to a fixed procedure  $q(x)dx$  in some model  $\mathcal{M} \in \{\mathcal{M}_1, \dots, \mathcal{M}_K\}$ . My goal is to select the right  $\mathcal{M} \in \{\mathcal{M}_1, \dots, \mathcal{M}_K\}$  by using the outputs.

# Statistical Model



# Kullback-Leibler distance

Given a model, recall that the *likelihood* of the data is

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

To compare the model distribution with the true distribution, we have the *log likelihood ratio*

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

In fact, the expectation of  $K_N(\omega)$  over the data distribution is the *Kullback-Leibler distance*

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

In statistics, this distance is an important measure of the difference between two distributions.

# Regular and Singular Models

Suppose  $q(x)dx$  equals  $p(x|\omega_0)dx$  for some  $\omega_0 \in \Omega$ .

The model is *identifiable* at  $\omega_0$  if the true fiber has only one point.

The *Fisher information matrix*  $I(\omega_0)$  is the Hessian matrix of the KL distance  $K(\omega)$  at  $\omega_0$ . This matrix is always *positive semidefinite*.

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

# Asymptotic Behavior

To analyze the *asymptotic behavior* of model selection criteria, we often need to understand the *log likelihood ratio*  $K_N(\omega)$ .

e.g. Marginal likelihood

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

e.g. For regular models, the Bayesian Information Criterion (BIC) uses the approximation  $-\log Z_N \approx -\log L_N^* + \frac{d}{2} \log N$  for model selection. Here,  $L_N^*$  is the maximum likelihood and  $d$  the model dimension.

Watanabe showed that the *log likelihood ratio*  $K_N(\omega)$  can be put in a nice standard form if we resolve the singularities of the *Kullback-Leibler distance*  $K(\omega)$ .

# Resolution of Singularities

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

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.]

## Standard Form of Log Likelihood Ratio (Watanabe)

Given mild conditions on the model  $\mathcal{M}$ , there exists a change of variable  $\rho : \mathcal{M} \rightarrow \Omega$  such that ( $\mu^\kappa$  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  $\mathcal{M}$ .

This is the *generalized Central Limit Theorem* for singular models.

# Learning Coefficient

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  $-\log Z_N$  has the asymptotic expansion

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

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

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

This is the *generalized BIC* for singular models.

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

# Computing the Learning Coefficient

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  $\theta$  is the number of times this minimum is attained.

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

# **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  $(p_1(\omega), p_2(\omega), \dots, p_k(\omega))$

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*  $I = \langle u_i(\omega) - u_i(\omega_i^*) \text{ for all } i \rangle$ .

It is the ideal of the *true fiber*  $V = \{\omega \in \Omega \mid q(x) = p(x|\omega) \text{ for all } x\}$ .

# Real Log Canonical Thresholds

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

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

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

- This definition is independent of the choice of generators for  $I$ .
- Fix  $I, \Omega$  and  $\varphi$ . For each point  $x \in \Omega$ , there exists a sufficiently small open neighborhood  $\Omega_x$  of  $x$  in  $\Omega$  such that  $\text{RLCT}_U(I; \varphi)$  is the same for all open neighborhoods  $U$  of  $x$  contained in  $\Omega_x$ .
- We order the pairs  $(\lambda, \theta)$  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  $\lambda$  and its order  $\theta$  of the model is given by

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

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

# Newton Polyhedra

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

1. Plot  $\alpha \in \mathbb{R}^d$  for each monomial  $\omega^\alpha$  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  $I$ .

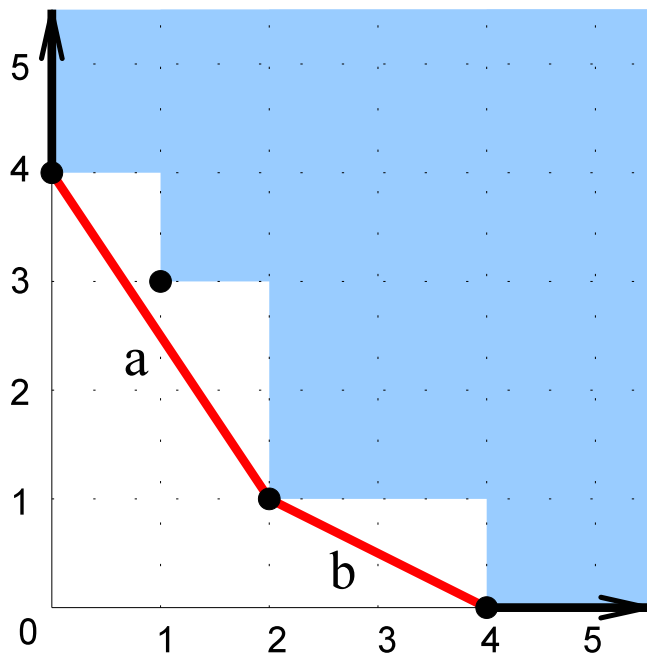
Given a vector  $\tau \in \mathbb{Z}_{\geq 0}^d$ , define

1.  *$\tau$ -distance*  $l_\tau$  : smallest  $t \geq 0$  such that  $t(\tau_1 + 1, \dots, \tau_d + 1) \in \mathcal{P}(I)$ .
2. *multiplicity*  $\theta_\tau$  : codimension of face of  $\mathcal{P}(I)$  at this intersection.

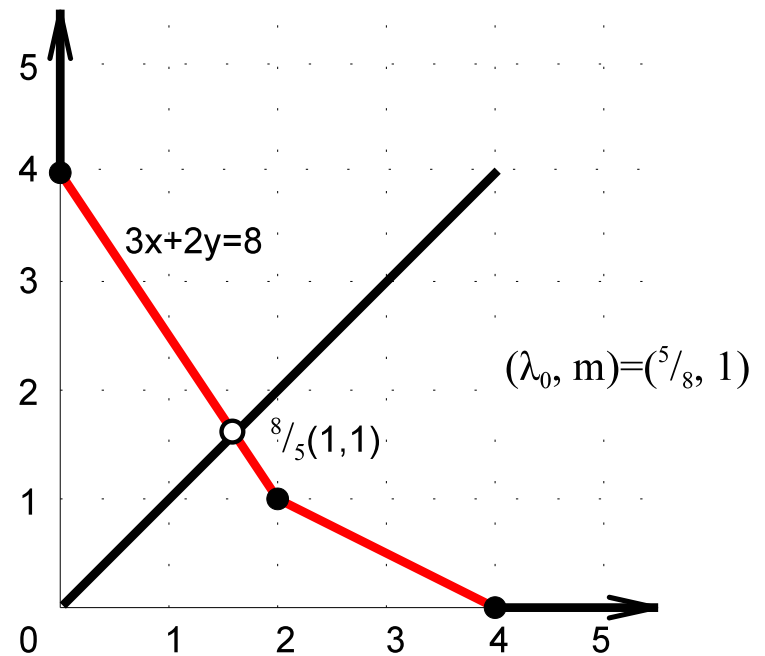
# Newton Polyhedra

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

Newton polyhedron



$\tau$ -distance



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

# Bounding the RLCT

## Theorem (L.)

Let  $I \subset \mathbb{R}[\omega_1, \dots, \omega_d]$  be a finitely generated ideal, and  $U \subset \mathbb{R}^d$  a sufficiently small nbhd of the origin. Then,

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

where  $l_\tau$  is the  $\tau$ -distance of the Newton polyhedron  $\mathcal{P}(I)$  and  $\theta_\tau$  its multiplicity.

Equality occurs when  $I$  is a monomial ideal.

Using this theorem, we can compute the RLCT of *any* ideal by monomializing the ideal.

# Examples

## Example 1: Bayesian Information Criterion

When the model is regular, the fiber ideal is  $I = \langle \omega_1, \dots, \omega_d \rangle$ .  
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).

# 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.

	$2 \leq Y < 10$	$10 \leq Y < 20$	$20 \leq Y$	<i>Totals</i>
Regularly	43	16	3	62
Rarely	6	11	10	27
Never	9	18	16	43
<i>Totals</i>	58	45	29	<b>132</b>

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

278019488531063389120643600324989329103876140805  
285242839582092569357265886675322845874097528033  
99493069713103633199906939405711180837568853737

and denominator

12288402873591935400678094796599848745442833177572204  
50448819979286456995185542195946815073112429169997801  
33503900169921912167352239204153786645029153951176422  
43298328046163472261962028461650432024356339706541132  
34375318471880274818667657423749120000000000000000.

# 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	<b>275.9144024</b>

“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 Series*. 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.