SINGULAR LEARNING THEORY

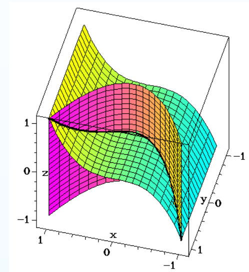
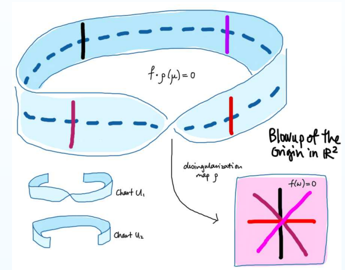
Part I: Statistical Learning

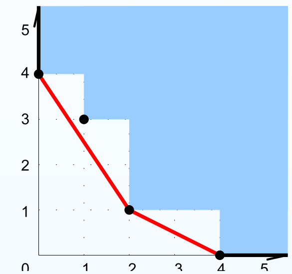
Shaowei Lin
(Institute for Infocomm Research, Singapore)

21-25 May 2013
Motivic Invariants and Singularities Thematic Program
Notre Dame University

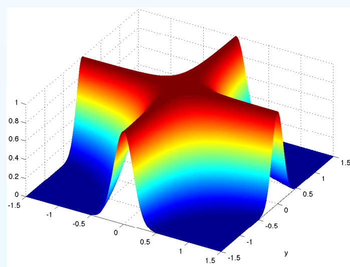
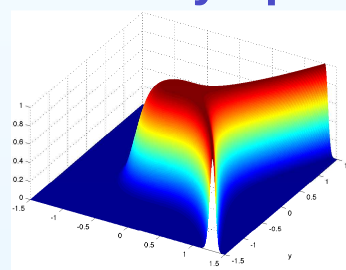
Overview

Algebraic Geometry

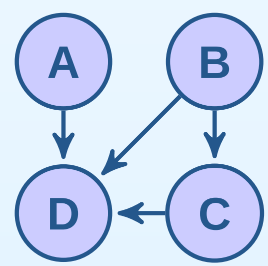


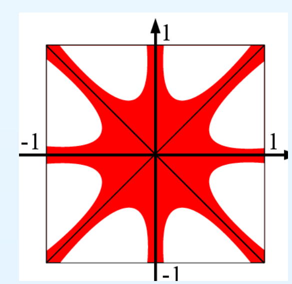
Asymptotic Theory

$$\int_{[0,1]^2}(1-x^2y^2)^{N/2}dxdy \approx \sqrt{\tfrac{\pi}{8}}N^{-\frac{1}{2}}\log N - \sqrt{\tfrac{\pi}{8}}\left(\tfrac{1}{\log 2} - 2\log 2 - \gamma\right)N^{-\frac{1}{2}} - \tfrac{1}{4}N^{-1}\log N + \tfrac{1}{4}\left(\tfrac{1}{\log 2} + 1 - \gamma\right)N^{-1} - \tfrac{\sqrt{2\pi}}{128}N^{-\frac{3}{2}}\log N + \cdots$$

Statistical Learning



Algebraic Statistics ***Singular Learning Theory***

Overview

Part I: Statistical Learning

Part II: Real Log Canonical Thresholds

Part III: Singularities in Graphical Models

Probability

Random Variables

A **probability space** $(\xi, \mathcal{F}, \mathbb{P})$ consists of

- a *sample space* ξ which is the set of all possible outcomes,
- a collection[♯] $\mathcal{F}$ of *events*, which are subsets of ξ,
- an assignment[♭] $\mathbb{P} : \mathcal{F} \to [0, 1]$ of *probabilities* to events

A (real-valued) **random variable** $X : \xi \to \mathbb{R}^k$ is

- a function[♮] from the sample space to a real vector space.
- a measurement of the possible outcomes.
- $X \sim \mathbb{P}$ means "X has the distribution given by $\mathbb{P}$".

♯ σ-algebra: closed under complement, countable union and contains $\emptyset$.
♭ probability measure: $\mathbb{P}(\emptyset) = 0$, $\mathbb{P}(\xi) = 1$, countable additivity for disjoint events.
♮ measurable function: for all $x \in \mathbb{R}$, the preimage of $\{y \in \mathbb{R}^k : y \leq x\}$ is in $\mathcal{F}$.

Example. Rolling a fair die.

$$\xi = \{⚀, ⚁, ⚂, ⚃, ⚄, ⚅\} \qquad \mathcal{F} = \{\emptyset, \{⚀\}, \{⚀, ⚁\}, \ldots\}$$
$$X \in \{1, 2, 3, 4, 5, 6\} \qquad \mathbb{P}(X = 1) = \frac{1}{6}, \ \mathbb{P}(X \leq 3) = \frac{1}{2}$$

Discrete and Continuous Random Variables

If ξ is finite, we say X is a **discrete** random variable.

- **probability mass function** $p(x) = P(X = x),\; x \in \mathbb{R}^k$.

If ξ is infinite, we define the

- **cumulative distribution function** (CDF) $F(x)$

$$F(x) = \mathbb{P}(X \leq x), \quad x \in \mathbb{R}^k.$$

- **probability density function**$^\sharp$ (PDF) $p(y)$

$$F(x) = \int_{\{y\in\mathbb{R}^k : y \leq x\}} p(y)dy, \quad x \in \mathbb{R}^k.$$

If the PDF exists, then X is a **continuous**$^\flat$ random variable.

$^\sharp$ Radon-Nikodym derivative of $F(x)$ with respect to the Lebesgue measure on $\mathbb{R}^k$.
$^\flat$ We can also define PDFs for discrete variables if we allow the Dirac delta function.

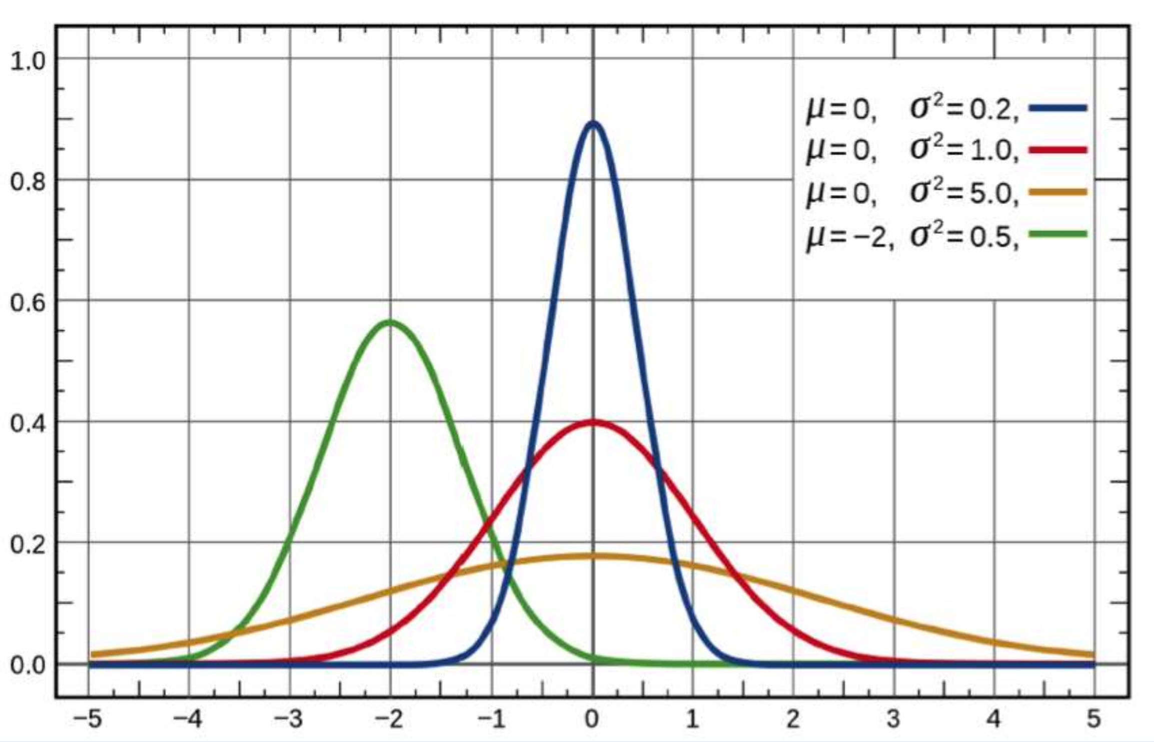
The probability mass/density function is often informally referred to as the **distribution** of X.

Gaussian Random Variables

Example. Multivariate Gaussian distribution $X \sim \mathcal{N}(\mu, \Sigma)$.

$X \in \mathbb{R}^k$, mean $\mu \in \mathbb{R}^k$, covariance $\Sigma \in \mathbb{R}^k_{\succ 0}$

$$p(x) = \frac{1}{(2\pi \det \Sigma)^{k/2}} \exp\left(-\frac{1}{2}(x-\mu)^\top \Sigma^{-1}(x-\mu)\right)$$



PDFs of univariate Gaussian distributions $X \sim \mathcal{N}(\mu, \sigma^2)$

Basic Concepts

The **expectation** $\mathbb{E}[X]$ is the integral of the random variable X with respect to its probability measure, i.e. the "average".

Discrete variables	Continuous variables
$\mathbb{E}[X] = \sum_{x\in\mathbb{R}^k} x\mathbb{P}(X=x)$	$\mathbb{E}[X] = \int_{\mathbb{R}^k} xp(x)\,dx$

The **variance** $\mathbb{E}[(X-\mathbb{E}[X])^2]$ measures the "spread" of X.

The **conditional probability**[♯] $\mathbb{P}(A|B)$ of two events $A, B \in \mathcal{F}$ is the probability that A will occur given that we know B has occurred.

- If $\mathbb{P}(B) > 0$, then $\mathbb{P}(A|B) = \mathbb{P}(A\cap B)/\mathbb{P}(B)$.

[♯] formal definition depends on the notion of conditional expectation.

Example. Weather forecast.

$\mathbb{P}(\text{Rain}|\text{Thunder})$
$= 0.2/(0.1+0.2) = 2/3$

	Rain	No Rain
Thunder	0.2	0.1
No Thunder	0.3	0.4

Independence

Let $X \in \mathbb{R}^k, Y \in \mathbb{R}^l, Z \in \mathbb{R}^m$ be random variables.

X, Y are **independent** $(X \perp\!\!\!\perp Y)$ if

- $\mathbb{P}(X \in S, Y \in T) = \mathbb{P}(X \in S)\,\mathbb{P}(Y \in T)$
  for all measurable subsets $S \subset \mathbb{R}^k, T \subset \mathbb{R}^l$.
- i.e. "knowing X gives no information about Y"

X, Y are **conditionally independent** given Z $(X \perp\!\!\!\perp Y \mid Z)$ if

- $\mathbb{P}(X \in S, Y \in T | Z = z) = \mathbb{P}(X \in S | Z = z)\,\mathbb{P}(Y \in T | Z = z)$
  for all $z \in \mathbb{R}^m$ and measurable subsets $S \subset \mathbb{R}^k, T \subset \mathbb{R}^l$.
- i.e. "any dependence between X and Y is due to Z"

Example. Hidden variables.

Favorite color $X \in \{\text{red}, \text{blue}\}$, favorite food $Y \in \{\text{salad}, \text{steak}\}$.
If X, Y are dependent, one may ask if there is a hidden variable, e.g. gender $Z \in \{\text{female}, \text{male}\}$, such that $X \perp\!\!\!\perp Y \mid Z$.

Statistics

Statistical Model

Let Δ denote the space of distributions with outcomes ξ.

Model: a family $\mathcal{M}$ of probability distributions, i.e. a subset of Δ.

Parametric model: family $\mathcal{M}$ of distributions $p(\cdot|\omega)$ are indexed by parameters ω in a space Ω, i.e. we have a map $\Omega \to \Delta$.

Example. Biased coin tosses.

Number of heads in two tosses of coin: $H \in \xi = \{0, 1, 2\}$

Space of distributions:

$$\Delta = \{p \in \mathbb{R}^3_{\geq 0} \;:\; p(0)+p(1)+p(2) = 1\}$$

Probability of getting heads: $\omega \in \Omega = [0, 1] \subset \mathbb{R}$

Parametric model for H:

$$\left.\begin{aligned} p(0|\omega) &= (1-\omega)^2 \\ p(1|\omega) &= 2(1-\omega)\omega \\ p(2|\omega) &= \omega^2 \end{aligned}\right\}$$

implicit equation

$$4\,p(0)p(2) - p(1)^2 = 0$$

Maximum Likelihood

A **sample** $X_1, \ldots, X_N$ of X is a set of independent, identically distributed (i.i.d.) random variables with the same distribution.

***Goal**: Given a statistical model $\{p(\cdot|\omega) \, : \, \omega \in \Omega\}$ and a sample, find a distribution $p(\cdot|\hat{\omega})$ that* best describes *the sample.*

A **statistic** $f(X_1, \ldots, X_N)$ is a function of the sample.
An important statistic is the **maximum likelihood estimate** (MLE).
It is a parameter $\hat{\omega}$ that maximizes the **likelihood function**

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

Example. Biased coin tosses.

Suppose the table below summarizes a sample of H of size 100.

H	0	1	2
Count	25	45	30

Then, $L(\omega) = 2^{45}\omega^{105}(1-\omega)^{95}$

$$\hat{\omega} = 105/200.$$

Kullback-Leibler Divergence

Let $X_1, \ldots, X_N$ be a sample of a *discrete* variable X. The **empirical distribution** is the function

$$\hat{q}(x) = \frac{1}{N} \sum_{i=1}^{N} \delta(x - X_i)$$

where $\delta(\cdot)$ is the Kronecker delta function.

The **Kullback-Leibler divergence** of a distribution p from q is

$$K(q||p) = \sum_{x \in \mathbb{R}^k} q(x) \log \frac{q(x)}{p(x)}.$$

Proposition. ML distributions minimize the KL divergence of $q(\cdot) = p(\cdot|\omega) \in \mathcal{M}$ from the empirical distribution $\hat{q}(\cdot)$.

$$K(\hat{q}||q) \;=\; \underbrace{\sum_{x \in \mathbb{R}^k} \hat{q}(x) \log \hat{q}(x)}_{\text{entropy}} \;-\; \frac{1}{N} \log L(\omega)$$

Kullback-Leibler Divergence

Let $X_1, \ldots, X_N$ be a sample of a *continuous* variable X. The **empirical distribution** is the *generalized function*

$$\hat{q}(x) = \frac{1}{N} \sum_{i=1}^{N} \delta(x - X_i)$$

where $\delta(\cdot)$ is the *Dirac* delta function.

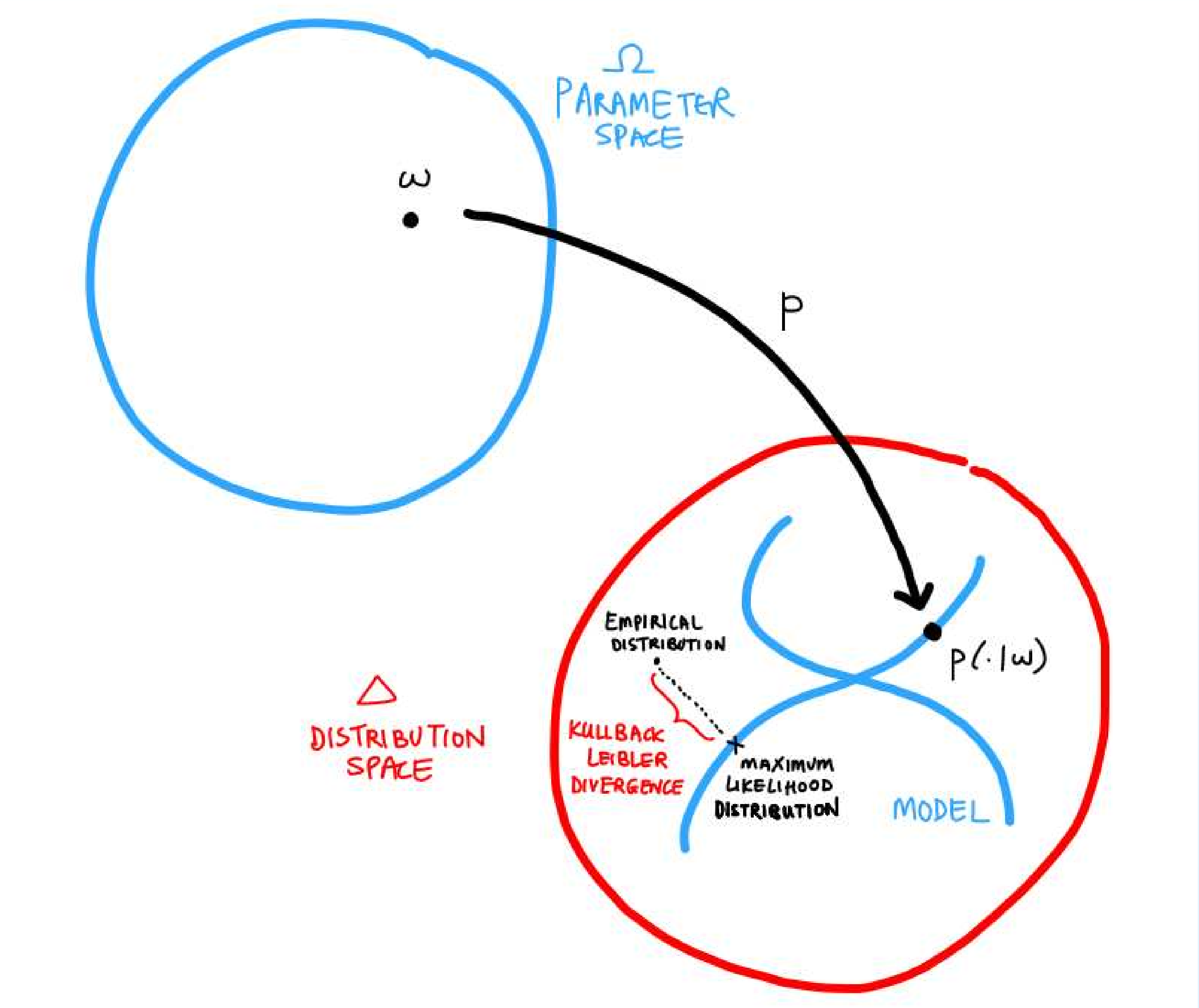
The **Kullback-Leibler divergence** of a distribution p from q is

$$K(q||p) = \int_{\mathbb{R}^k} q(x) \log \frac{q(x)}{p(x)} dx.$$

Proposition. ML distributions minimize the KL divergence of $q(\cdot) = p(\cdot|\omega) \in \mathcal{M}$ from the empirical distribution $\hat{q}(\cdot)$.

$$K(\hat{q}||q) \;=\; \underbrace{\int_{\mathbb{R}^k} \hat{q}(x) \log \hat{q}(x) dx}_{\text{entropy}} \;-\; \frac{1}{N} \log L(\omega)$$

Kullback-Leibler Divergence



Kullback-Leibler Divergence

Example. Population mean and variance.

Let $X \sim N(\mu, \sigma^2)$ be the height of a random Singaporean.
Given sample $X_1, \ldots, X_N$, estimate mean μ and variance σ^2.

Now, the Kullback-Leibler divergence is

$$K(\hat{q}||q) = \frac{1}{2\sigma^2 N}\sum_{i=1}^{N}(X_i - \mu)^2 + \frac{1}{N}\log\sigma + \text{constant.}$$

Differentiating this function gives us the MLE

$$\hat{\mu} = \frac{1}{N}\sum_{i=1}^{N} X_i, \quad \hat{\sigma}^2 = \frac{1}{N}\sum_{i=1}^{N}(X_i - \hat{\mu})^2.$$

MLE for the model mean is the sample mean.
MLE for the model variance is the sample variance.

Mixture Models

A **mixture** of distributions $p_1(\cdot), \ldots, p_m(\cdot)$ is a convex combination

$$p(x) = \sum_{i=1}^{m} \alpha_i p_i(x), \quad x \in \mathbb{R}^k$$

i.e. the **mixing coefficients** α_i are nonnegative and sum to one.

Example. Gaussian mixtures.

Mixing univariate Gaussians $\mathcal{N}(\mu_i, \sigma_i^2)$, $i = 1, \ldots, m$, produces distributions of the form

$$p(x) = \sum_{i=1}^{m} \frac{\alpha_i}{\sqrt{2\pi\sigma_i^2}} \exp\left(-\frac{(x-\mu_i)^2}{2\sigma_i^2}\right).$$

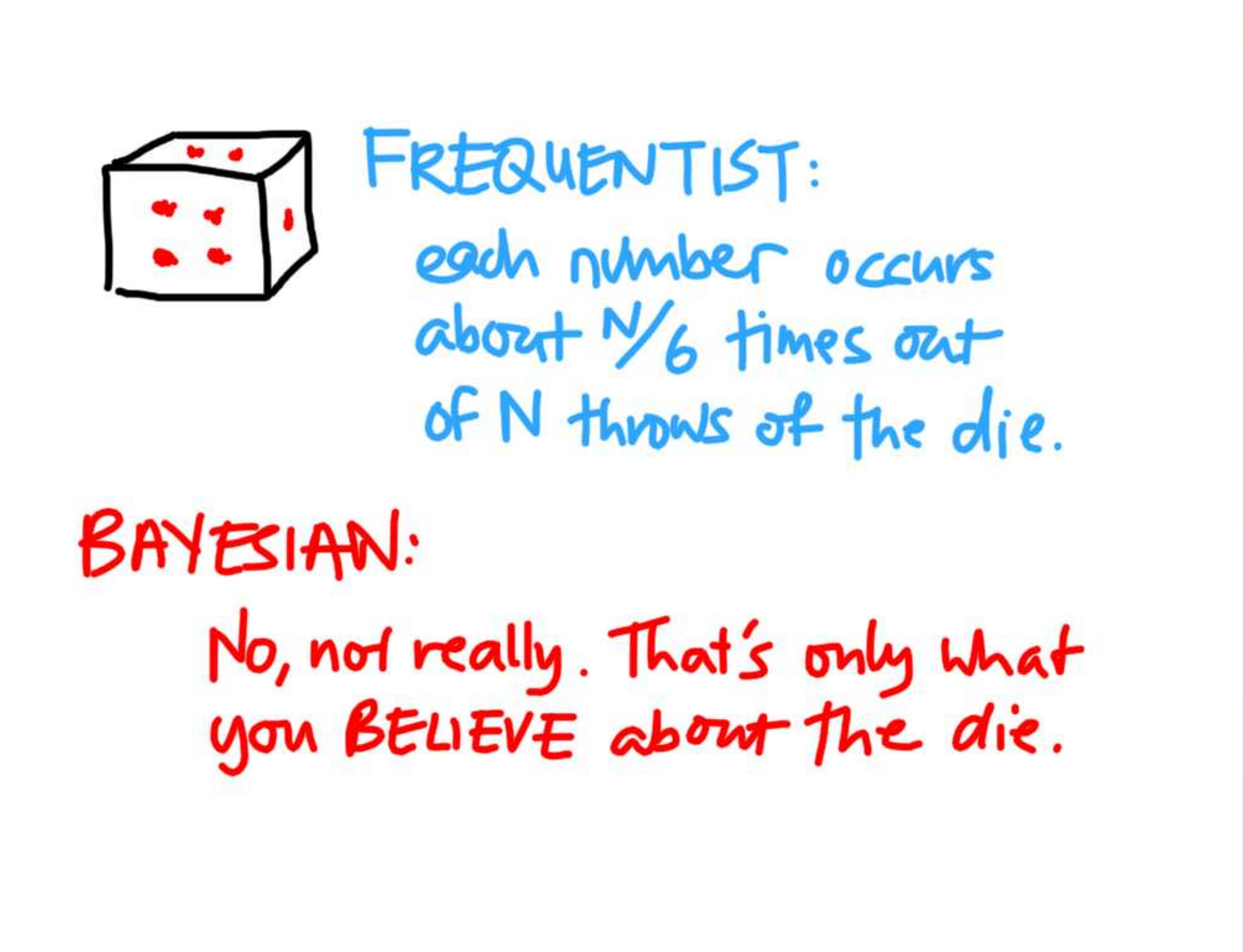
This mixture model is therefore described by parameters

$$\omega = (\alpha_1, \ldots, \alpha_m, \mu_1, \ldots, \mu_m, \sigma_1, \ldots, \sigma_m)$$

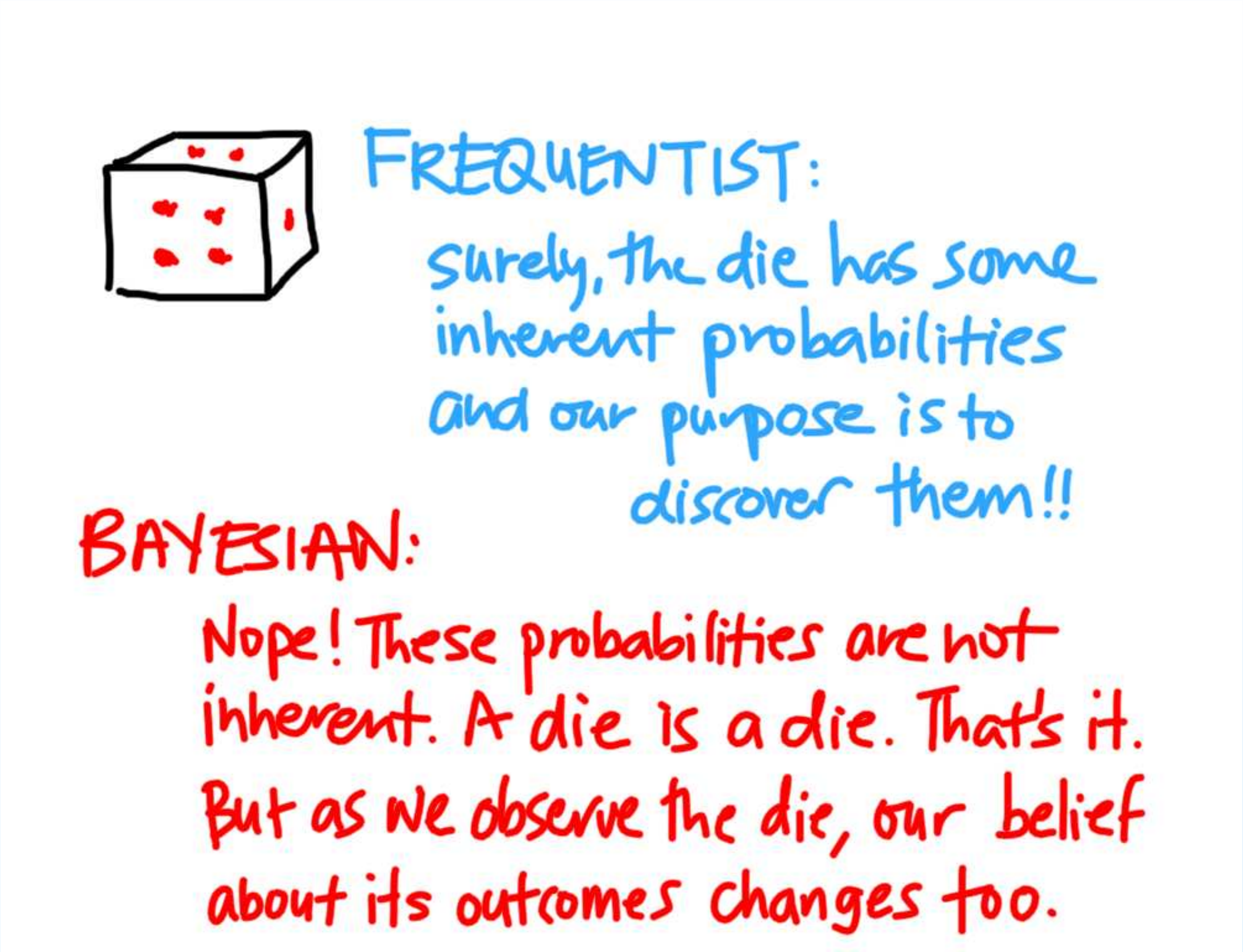
and is frequently used in cluster analysis.

Bayesian Statistics

Interpretations of Probability Theory



Interpretations of Probability Theory



Bayes' Rule

Updating our belief of event A based on an observation B.

$$\underbrace{\mathbb{P}(A|B)}_{\text{posterior (new belief)}} = \frac{\mathbb{P}(B|A)}{\mathbb{P}(B)} \underbrace{\mathbb{P}(A)}_{\text{prior (old belief)}}$$

Example. Biased coin toss.

Let θ denote $\mathbb{P}(\text{heads})$ of a coin.
Determine if the coin is fair $(\theta = \frac{1}{2})$ or biased $(\theta = \frac{3}{4})$.
Old belief: $\mathbb{P}(\text{fair}) = 0.9$
Now, suppose we observed a sample with 8 heads and 2 tails.
New belief:

$$\mathbb{P}(\text{fair}|\text{sample}) = \frac{\mathbb{P}(\text{sample}|\text{fair})}{\mathbb{P}(\text{sample})}\mathbb{P}(\text{fair})$$

$$= \frac{(\frac{1}{2})^8(\frac{1}{2})^2(0.9)^8}{(\frac{1}{2})^8(\frac{1}{2})^2(0.9)^8 + (\frac{3}{4})^8(\frac{1}{4})^2(0.1)^8} \approx 0.584$$

Parameter Estimation

Let $\mathcal{D} = \{X_1, \ldots, X_N\}$ denote a sample of X, i.e. "the data".

Frequentists: Compute maximum likelihood estimate.

Bayesians: *Treat parameters $\omega \in \Omega$ as random variables* with priors $p(\omega)$ that require updating.

Posterior distribution on parameters $\omega \in \Omega$	$p(\omega\|\mathcal{D}) = \frac{p(\mathcal{D}\|\omega)p(\omega)}{\int_\Omega p(\mathcal{D}\|\omega)p(\omega)d\omega}$
Posterior mean	$\int_\Omega \omega\, p(\omega\|\mathcal{D})d\omega$
Posterior mode	$\text{argmax}_{\omega\in\Omega}\ p(\omega\|\mathcal{D})$
Predictive distribution on outcomes $x \in \xi$	$p(x\|\mathcal{D}) = \int_\Omega p(x\|\omega)p(\omega\|\mathcal{D})d\omega$

These integrals are difficult to compute or even estimate!

Model Selection

Which model $\mathcal{M}_1, \ldots, \mathcal{M}_m$ best describes the sample $\mathcal{D}$?

Frequentists: Pick model containing ML distribution.

Bayesians: Assign priors $p(\mathcal{M}_i)$, $p(\omega|\mathcal{M}_i)$ and compute

$$\mathbb{P}(\mathcal{M}_i|\mathcal{D}) = \frac{\mathbb{P}(\mathcal{D}|\mathcal{M}_i)\mathbb{P}(\mathcal{M}_i)}{\mathbb{P}(\mathcal{D})} \propto \mathbb{P}(\mathcal{D}|\mathcal{M}_i)\mathbb{P}(\mathcal{M}_i).$$

where $\mathbb{P}(\mathcal{D}|\mathcal{M}_i)$ is the **likelihood integral**

$$\mathbb{P}(\mathcal{D}|\mathcal{M}_i) = \int_\Omega \prod_{i=1}^N p(X_i|\omega, \mathcal{M}_i)p(\omega|\mathcal{M}_i)d\omega.$$

Parameter estimation is a form of model selection!

For each $\omega \in \Omega$, we define a model $\mathcal{M}_\omega$ with one distribution.

Estimating Integrals

Generally, there are three ways to estimate statistical integrals.

1. *Exact methods*
   Compute a closed form formula for the integral, e.g.
   Baldoni·Berline·De Loera·Köppe·Vergne, 2010;
   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.
   Rewrite the likelihood integral as

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

   where $f(\omega) = -\frac{1}{N} \log L(\omega)$ and $\varphi(\omega)$ is the prior on Ω.

Bayesian Information Criterion

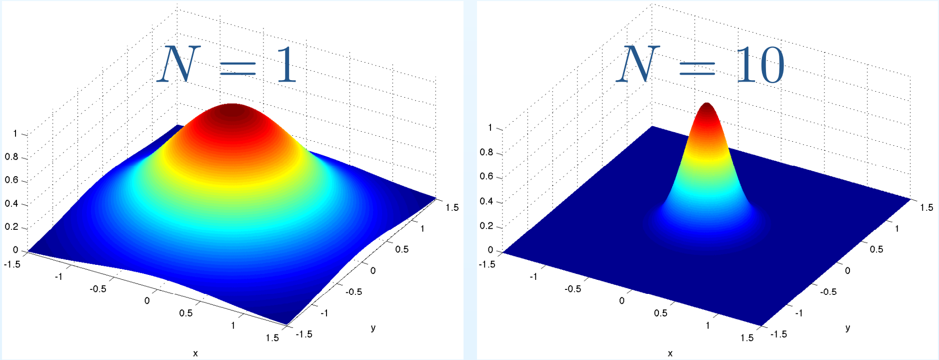
Laplace approximation: If $f(\omega)$ is uniquely minimized at MLE $\hat{\omega}$ and the Hessian $\partial^2 f(\hat{\omega})$ is full rank, then *asymptotically*

$$-\log Z_N \approx N f(\hat{\omega}) + \frac{\dim \Omega}{2} \log N + O(1)$$

as sample size $N \to \infty$.

Bayesian information criterion (BIC): Select model that maximizes

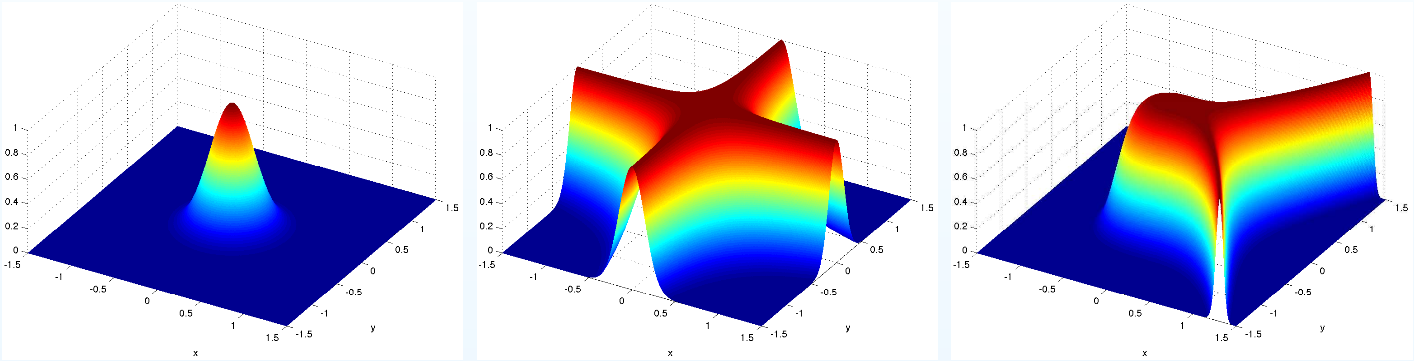
$$N f(\hat{\omega}) + \frac{\dim \Omega}{2} \log N$$



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

Singularities in Statistical Models

Informally, the model is **singular** at $\omega_0 \in \Omega$ if the Laplace approximation fails when the empirical distribution is $p(\cdot|\omega_0)$.



Formally, if we define the **Kullback-Leibler function**

$$K(\omega) = \int_\xi p(x|\omega_0) \log \frac{p(x|\omega_0)}{p(x|\omega)} dx.$$

then ω_0 is a *singularity* when the Hessian $\partial^2 K(\omega_0)$ is not full rank.

Statistical models with **hidden variables**, e.g. mixture models, often contain many singularities.

Linear Regression

Least Squares

Suppose we have random variables $Y \in \mathbb{R}, X \in \mathbb{R}^d$ that satisfy

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

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

- 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)	$\pi(\omega) = \lvert\omega\rvert_0 \cdot 2$

- Commonly asked questions

  Model selection (e.g. which factors are important?)

  Parameter estimation (e.g. how important are the factors?)

Sparsity Penalty

The best model is usually selected using a score

$$\mathrm{argmin}_{u\in\Omega}\, l(u) + \pi(u)$$

where the likelihood $l(u)$ measures the *fitting error* of the model while the penalty $\pi(u)$ measures its *complexity*.

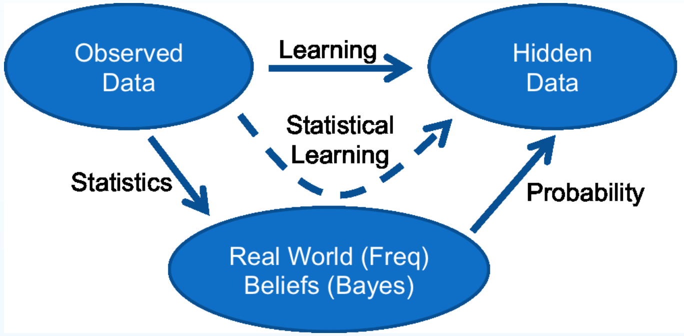
Recently, *sparse penalties* derived from statistical considerations were found to be highly effective.

Bayesian info criterion (BIC)	$\leftrightarrow$	Marginal likelihood integral
Akaike info criterion (AIC)	$\leftrightarrow$	Kullback-Leibler divergence
Compressive sensing	$\leftrightarrow$	ℓ_1-regularization of BIC

Singular learning theory plays an important role in these derivations.

Paradigms in Statistical Learning

- **Probability**
  theory of random phenomena
  Statistics
  theory of making sense of data
  Learning
  the art of prediction using data



- **Frequentists**: "True distribution, maximum likelihood"
  Bayesians: "Belief updates, maximum a posteriori"

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

- We often have to balance the *complexity* of the model with its *fitting error* via some suitable probabilistic criteria.
  The learning algorithm also needs to be *computable* to be useful.

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. V. BALDONI, N. BERLINE, J. A. DE LOERA, M. KÖPPE, M. VERGNE: *How to integrate a polynomial over a simplex*, Mathematics of Computation **80** (2010) 297–325.
3. A. BRAVO, S. ENCINAS AND O. VILLAMAYOR: A simplified proof of desingularisation and applications, *Rev. Math. Iberoamericana* **21** (2005) 349–458.
4. 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.
5. R. DURRETT: *Probability - Theory and Examples* (4th edition), Cambridge U. Press, 2010.
6. M. EVANS, Z. GILULA AND I. GUTTMAN: Latent class analysis of two-way contingency tables by Bayesian methods, *Biometrika* **76** (1989) 557–563.
7. 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.
8. S. LIN, B. STURMFELS AND Z. XU: Marginal likelihood integrals for mixtures of independence models, *J. Mach. Learn. Res.* **10** (2009) 1611–1631.
9. S. LIN: Algebraic methods for evaluating integrals in Bayesian statistics, PhD dissertation, Dept. Mathematics, UC Berkeley (2011).
10. L. MACKEY: *Fundamentals - Probability and Statistics*, (slides for CS 281A: Statistical Learning Theory at UC Berkeley, Aug 2009).
11. S. WATANABE: *Algebraic Geometry and Statistical Learning Theory*, Cambridge Monographs on Applied and Computational Mathematics **25**, Cambridge University Press, Cambridge, 2009.