SF 2930 Regression Analysis
LECTURE 11.2
Bayesian Statistics

Timo Koski

KTH Royal Institute of Technology

2023

Contents

Conditional Probability

By first definitions, assuming $P(E) > 0$

$$P(H \mid E) = \frac{P(H \cap E)}{P(E)} \Leftrightarrow P(H \cap E) = P(H \mid E) \cdot P(E)$$

and

$$P(E \mid H) = \frac{P(E \cap H)}{P(H)} = \frac{P(H \cap E)}{P(H)} \Leftrightarrow P(H \cap E) = P(E \mid H)P(H)$$

Inversion and Bayes' Rule

Hence

$$P(H \mid E) \cdot P(E) = P(E \mid H) \cdot P(A)$$

Bayes' Rule or inversion of probability

$$P(H \mid E) = \frac{P(E \mid H) \cdot P(H)}{P(E)}.$$

$$P(E) = P(E \mid H)P(H) + P\left(E \mid H^C\right) P\left(H^C\right).$$

H^C is the complement set of H.

Inversion and Bayes' Rule

Think about

$$H = \text{statement/hypothesis}$$

$$P(H) = \text{Prior probability of the statement/ hypothesis}$$

$$E = \text{evidence pertaining to the statement}$$

$P(E \mid H) =$ Probabilty or Likelihood of the Evidence given the Statem

$P(H \mid E) =$ Posterior probability of H, the statement/hypothesis given

Dictionary

- a priori: relating an argument that suggests using general principles to suggest likely effects, being without examination of facts, formed or conceived beforehand.
- a posteriori: relating to or derived by reasoning from observed facts

Bayes' Rule as a Slogan & The Iconic Portrait

$$\text{Posterior} \propto \text{Likelihood} \times \text{Prior}$$

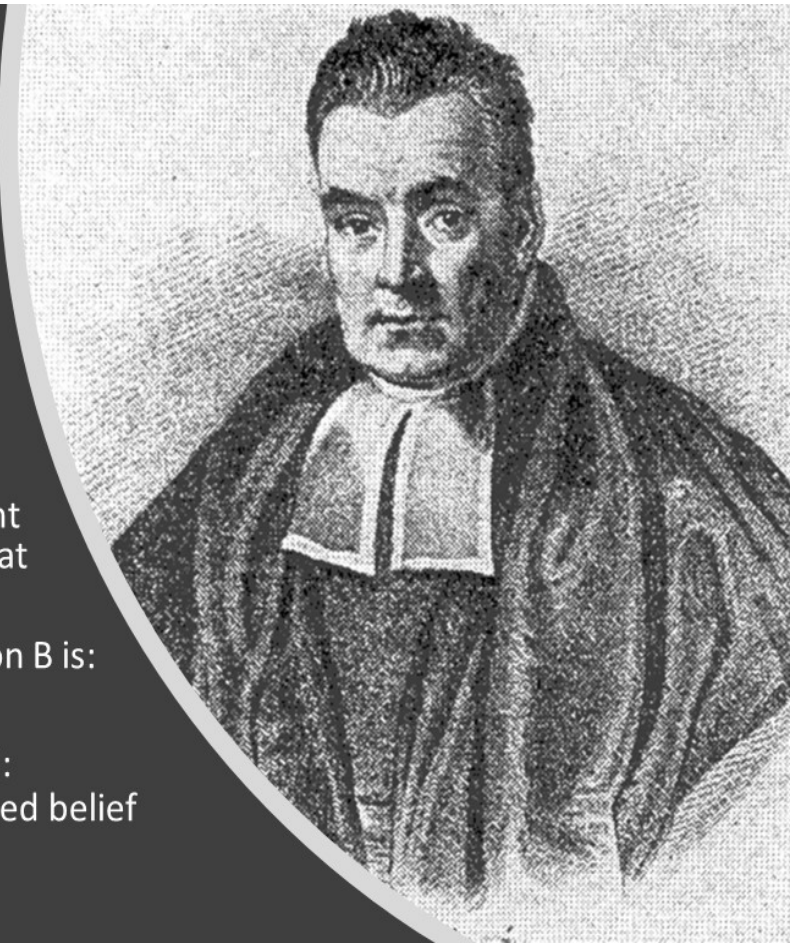
A Complete form of Bayes' Rule

$\cup_{i=1}^{k} H_j =$ the whole sample space, $H_j \cap H_i = \emptyset$.

$$P(H_j \mid E) = \frac{P(E \mid H_j) \cdot P(H_j)}{P(E)}.$$

$$P(E) = \sum_{i=1}^{k} P(E \mid A_i) P(A_i).$$

Rev. Thomas Bayes (c. 1702 – 1761)



- English theologian and mathematician
- Bayes' Theorem: the probability of an event based on prior knowledge of conditions that are related to the event
- i.e., the probability of A under the condition B is: $P(A|B) = P(B|A) \cdot P(A)/P(B)$
- i.e., "In (Conditional) Probability We Trust": Initial belief + New data = Adjusted improved belief

Bayes and Learning

There are N coins in an urn. $N-1$ of these are honest in the sense that one side is a head and the other a tail. One coin is false, it has a head on both sides.
A person A picks at random one of the coins. You are NOT permitted to inspect this selected coin. A tosses the coin k times. The outcome is k heads. Compute the probability that the selected coin is the false one?

Notations: $H_1 =$ honest coin, $H_2 =$ false coin. $E_k =$ the event of k heads in k tosses.
$P(H_1) = \frac{N-1}{N}$ the prior probability of the honest coin. $P(H_2) = \frac{1}{N}$.
$P(E_k \mid H_1) = \frac{1}{2^k}$, i.e., we assume independent tosses , each toss $\sim$ Be(1/2) given H_1. $P(E_k \mid H_2) = 1$.
Sought: the posterior probability $P(H_2 \mid E_k)$.

Bayes and Learning

H_1 = honest coin, H_2 = false coin. E_k = the evidence : k heads in k tosses.
$P(H_1) = \frac{N-1}{N}$ the prior probability of the honest coin. $P(H_2) = \frac{1}{N}$.
$P(E_k \mid H_1) = \frac{1}{2^k}$, i.e., we assume independent tosses.
$P(E_k \mid H_2) = 1$. Sought: the posterior probability $P(H_2 \mid E_k)$.

Law of total probability gives

$$P(E_k) = P(E_k \mid H_1)\, P(H_1) + P(E_k \mid H_2)\, P(H_2)$$

$$= \frac{1}{2^k}\frac{N-1}{N} + \frac{1}{N} = \frac{2^k + N - 1}{2^k N}$$

Bayes' rule entails

$$P(H_2 \mid E_k) = \frac{P(E_k \mid H_2)\, P(H_2)}{P(E_k)} = \frac{2^k}{2^k + N - 1}.$$

Bayes and Learning

H_1 = honest coin, H_2 = false coin. A_k = the event of k heads in k tosses.
$P(H_1) = \frac{N-1}{N}$ the prior probability of the honest coin. $P(H_2) = \frac{1}{N}$.
$P(A_k \mid H_1) = \frac{1}{2^k}$, i.e., we assume independent tosses.
$P(A_k \mid H_2) = 1$. Sought: $P(H_2 \mid A_k)$.

$$P(H_2 \mid A_k) = \frac{2^k}{2^k + N - 1}.$$

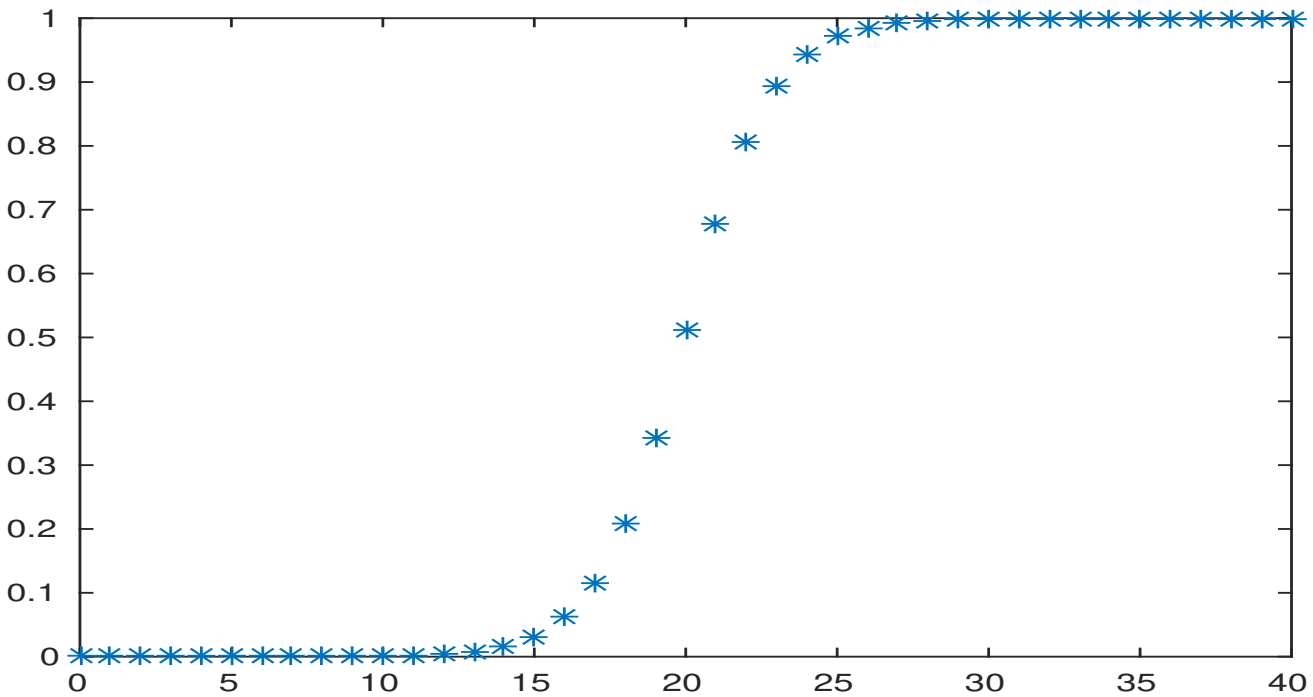
Note that

$$P(H_2 \mid A_0) = \frac{1}{N} = P(H_2).$$

Bayes and Learning

In the figure the posterior probability $P(H_2 \mid A_k)$ *is plotted with* $*$ *as a function of* k *for* $k = 0, \ldots, 40$ *and* $N = 1000000$

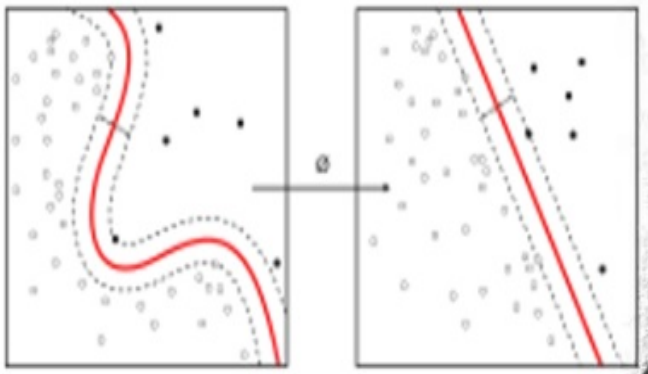
$$P(H_2 \mid A_k) = \frac{2^k}{2^k + N - 1}.$$



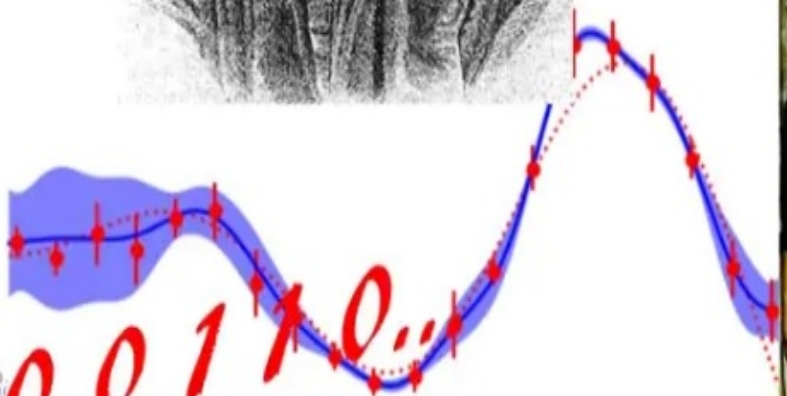
Thomas Bayes & Occam

Is there not also an Occam's razor at work here? We converge to the simplest explanation after having seen 40 tosses of coin with 40 heads as outcome.

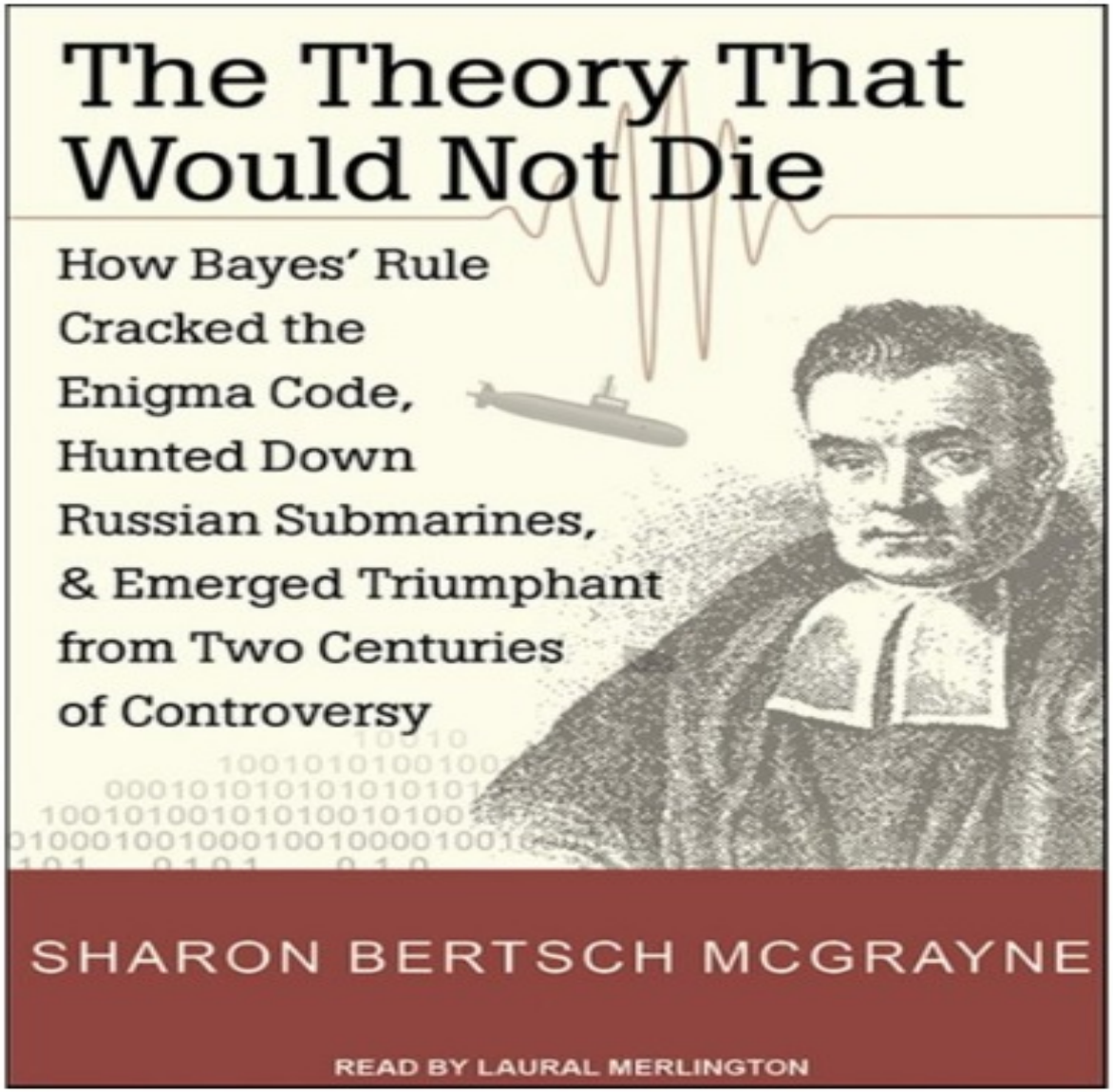
Thomas Bayes



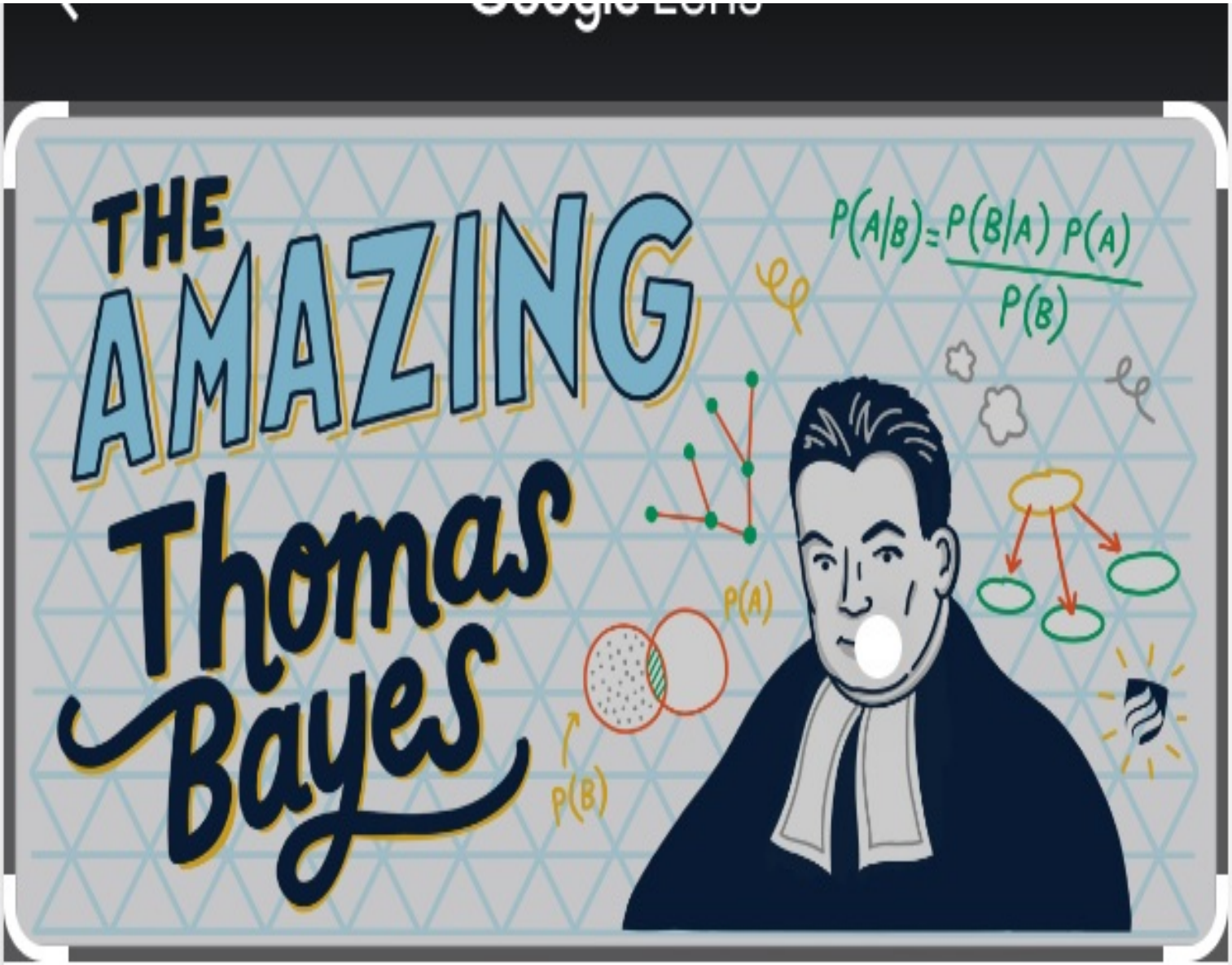
$$P(h|D) = \frac{P(D|h).P(h)}{P(D)}$$



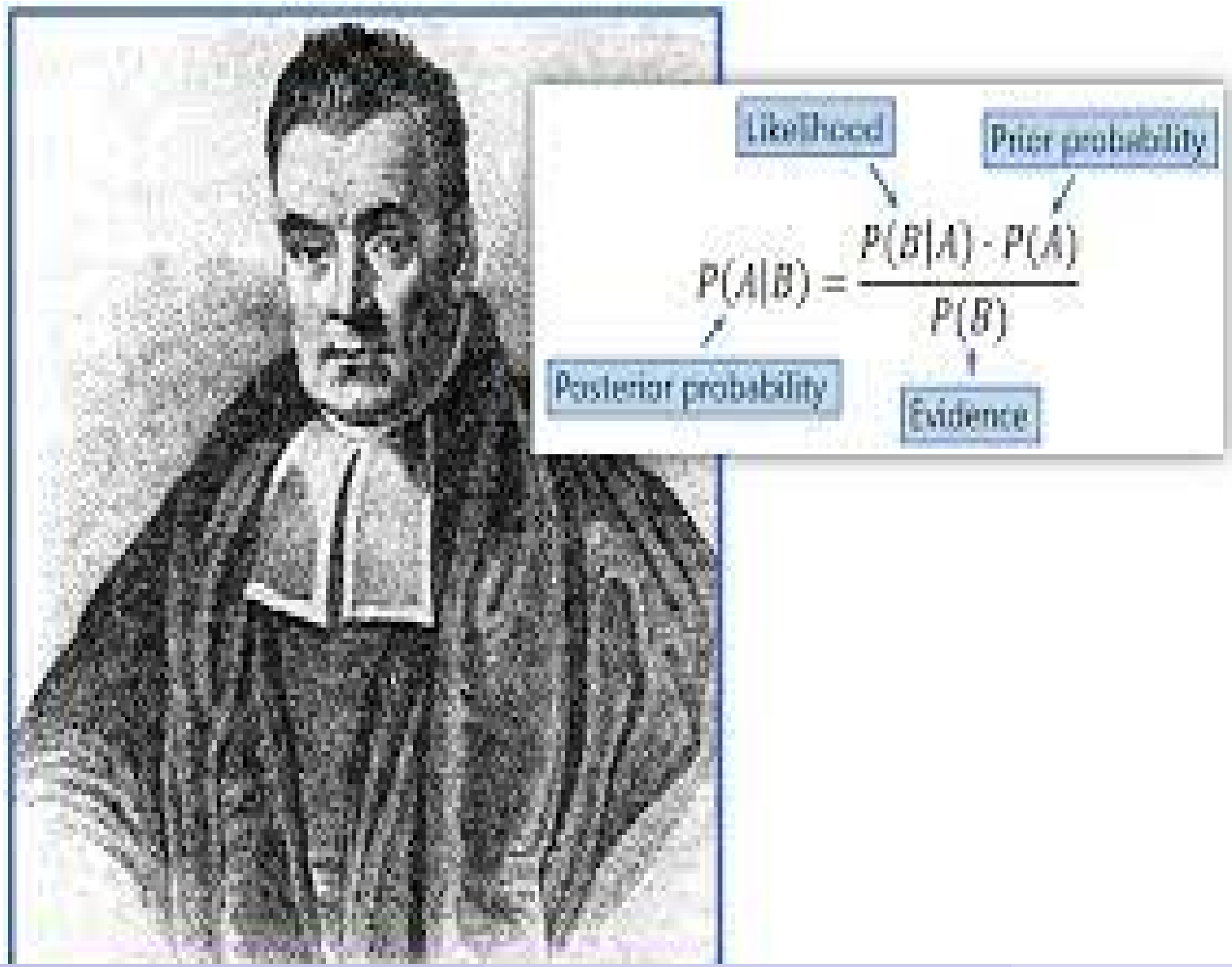
Thomas Bayes: A Cult Figure



Thomas Bayes: A Cult Figure



Thomas Bayes



A quote from[1]

> Today, Bayes formua is ubiquitous in the world of AI. We all use it on our smartphones without realizing it. In machine learning systems today, Bayesian inference is more prominent than ever.

[1] https://www.bbvaopenmind.com/en/technology/artificial-intelligence/bayesian-inference-ai-systems/

Learning/inference from data

By learning/inference from data one often means the process of inferring a general law or principle from the observations of particular instances. The general law is a piece of knowledge about the mechanism of nature that generates the data.
The intended learning in statistics is done by use of 'MODELS', which serve as the language in which the constraints predicated on the data can be described.
We deal here with parametric statistical models.

Parametric statistical model: the standard view

x is an observation of a random variable (X).

$$X \sim f(x;\theta)$$

$f(x;\theta)$ is a probability density on R^p. $f(x;\theta)$ is a known function of x and θ. θ is an unknown parameter.

- X is distributed according to $f(x|\theta)$,
- x is an observation from the distribution f.

An outcome x of the random variable (r.v.) X.

Parametric statistical model: Examples; Normal distribution

$\theta = (\mu, \sigma^2) \in \Theta = R \times (0, \infty)$.

$$f(x; \theta) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{\sigma^2}(x-\mu)^2}, -\infty < x < \infty.$$

We say that x is an observation from the normal distribution $N(\mu, \sigma^2)$.

Parametric statistical model: Examples; Bernoulli Distribution

Consider r.v. X with values $0, 1$, $0 < \theta < 1$ and

	$x = 1$	$x = 0$
$f(x;\theta)$	θ	$1 - \theta$

then we say X is distributed according to the Bernoulli distribution with the parameter θ.

$$X \sim\sim Be(\theta),$$

Learning/inference from data: Inversion

Retrieve the parameters of the probabilistic generating mechanism using x. $f(x;\theta)$ is a probabilistic generating mechanism of data, characterizes the behaviour of future observations conditional on θ, but in inference the roles of x and θ are inverted.

A Change of Notation

Allan Gut uses in *An Intermediate Course in Probability* the notation

$$f_{Y|X=x}(y)$$

to denote the conditional probability density of Y given $X = x$. In this Lecture the notation for the conditional probability density of Y given $X = x$ will be

$$f(y|x)$$

Bayes' rule

Bayes' rule extended to continuous random variables:

$$g(y|x) = \frac{f(x|y) \cdot g(y)}{\int f(x|y) \cdot g(y)\, dy},$$

Due to the standardization $g(y|x)$ is a probability density; $g(y|x) \geq 0$, $\int g(y|x)\, dy = 1$.

Bayes' rule: parametric model $f(x;\theta) \mapsto f(x|\theta)$

$$\pi(\theta|x) = \frac{f(x \mid \theta) \cdot \pi(\theta)}{\int_\Theta f(x \mid \theta) \cdot \pi(\theta)\, d\theta}$$

Terminology for Bayes' Rule:

- $\pi(\theta)$: **prior distribution** on Θ.
- $X \mid \theta$, $f(x \mid \theta)$ p.d.f: **likelihood**
- $\pi(\theta|x)$: **posterior distribution** on Θ.
- $m(x) = \int_\Theta f(x \mid \theta) \cdot \pi(\theta)\, d\theta$: **marginal distribution** of x.

Due to Pierre-Simon, marquis de Laplace 1749 – 1827

$$\pi(\theta|x) = \frac{f(x \mid \theta) \cdot \pi(\theta)}{\int_\Theta f(x \mid \theta) \cdot \pi(\theta)\, d\theta}$$

Uncertainty

Uncertainty about the unknown θ is modeled by a probability distribution $\pi(\theta)$, and $\pi(\theta|x)$ expresses the uncertainty about the unknown θ after the observation of x.
Hence this analysis does not regard θ as a random variable. $\pi(\theta)$ expresses our subjective a priori opinion of where in Θ the unknown might lie. This was/is controversial.

Mathematically: the unknown θ is dealt with as a random variable. (x, θ) will have a joint distribution.

Distributions

$$\pi(\theta|x) = \frac{f(x \mid \theta) \cdot \pi(\theta)}{m(x)},$$

Terminology:

- $m(x) = \int_\Theta f(x \mid \theta) \cdot \pi(\theta)\, d\theta$
- $\phi(x, \theta)$: **joint distribution** of (x, θ).
- $$\pi(\theta|x)\, m(x) = \phi(x, \theta) = f(x|\theta)\, \pi(\theta)$$

Notation

The notation

$$\int_{\Theta} f(x \mid \theta) \cdot \pi(\theta)\, d\theta$$

is imprecise by intent, as it can mean both a single integral and a multiple integral.

Bayesian Parametric Statistical Model

A Bayesian parametric statistical model consists of

- a prior distribution

$$\theta \sim \pi(\theta)$$

- a parametric model

$$x|\theta f\left(x|\theta\right)$$

The quantity of interest

$$\theta|x \sim \pi\left(\theta|x\right)$$

Prior Density

Any function $\pi(\cdot)$ such that

$$\pi(\theta) \geq 0,$$

and

$$\int_{\Theta} \pi(\theta)\, d\theta = 1,$$

can technically serve as a prior distribution.

Improper Prior Densities

But even functions with the properties

$$\pi(\theta) \geq 0,$$

and

$$\int_{\Theta} \pi(\theta)\, d\theta = \infty,$$

are also invoked as priors, and are called improper priors.

An Example

$X_i \mid M = m \sim \mathbf{N}\left(m, \sigma_0^2\right)$, $M \sim \mathbf{N}\left(\mu, s^2\right)$. $x^{(n)} = (x_1, \ldots, x_n)$ a sample of I.I.D. X_i, $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$.

$$M \mid (X_1, \ldots, X_n) \in \mathbf{N}\left(\frac{n\overline{x}/\sigma_0^2 + \mu/s^2}{n/\sigma_0^2 + 1/s^2}, \frac{1}{n/\sigma_0^2 + 1/s^2}\right)$$

i.e., $\pi\left(m|x^{(n)}\right)$ is the density of this normal distribution. Here μ and s^2 are *hyperparameters*.

Confidence interval

- $P(a(x) \leq \theta \leq b(x)) = \int_{a(x)}^{b(x)} \pi(\theta|x)\, d\theta$
- $$\underbrace{P(a(x) \leq \theta \leq b(x))}$$

  This is a probability, not a degree of confidence

Confidence interval: Example

Take the example above

$$\mathbf{N}\left(\frac{n\overline{x}/\sigma_0^2+\mu/s^2}{n/\sigma_0^2+1/s^2},\frac{1}{n/\sigma_0^2+1/s^2}\right)$$

Let $s\to\infty$ (the prior becomes improper). Then

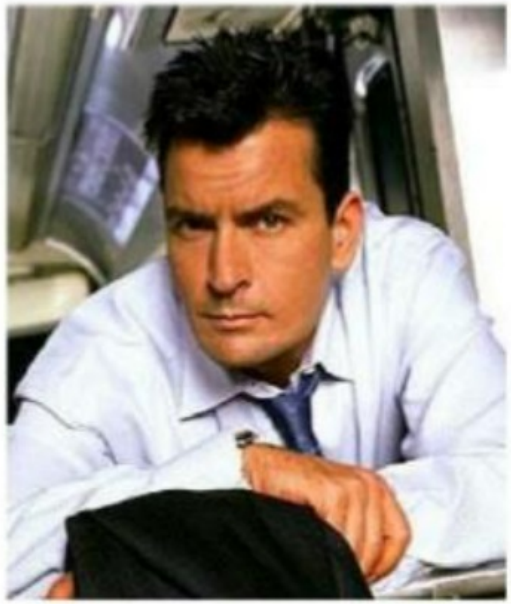
$$\mathbf{N}\left(\frac{n\overline{x}/\sigma_0^2+\mu/s^2}{n/\sigma_0^2+1/s^2},\frac{1}{n/\sigma_0^2+1/s^2}\right)\to\mathbf{N}\left(\overline{x},\frac{\sigma_0^2}{n}\right)$$

But then the Bayesian confidence interval
$P(a(x)\le\theta\le b(x))=0.95$ becomes the familiar $\overline{x}\pm\lambda_{0.025}\frac{\sigma_0}{\sqrt{n}}$, and the common mistaken but natural interpretation of the confidence interval is correct !

Slide from CS 109 Stanford University

Important fact that of which all CS109 students have historically been made aware, and of which I will now make you aware:

The Rev. Thomas Bayes (1702-1761) bore an uncanny resemblance to actor Charlie Sheen.



General Aspects of Bayesian Inference

- inference is based on the observed x, not on an unobserved sample space.
- $\pi(\theta|x)$ is the only quantity evaluated for inference about θ.

On the other hand , evaluation of $\pi(\theta|x)$ is not in general possible by explicit means of integral calculus. This is where statistical inference needs Markov chain Monte Carlo (McMC).

Likelihood

The distribution $f(x \mid \theta)$ regarded as a function of θ is known as the *likelihood function*

$$l(x;\theta) = f(x \mid \theta).$$

The likelihood function $l(\theta|x)$ thus compares the plausibilities of different parameter values for given x.

Bayes' rule: posterior $\propto$ likelihood $\times$ prior

$$\pi(\theta|x) = \frac{f(x \mid \theta) \cdot \pi(\theta)}{\int_\Theta f(x \mid \theta) \cdot \pi(\theta)\, d\theta},$$

$$= \frac{l(x;\theta) \cdot \pi(\theta)}{\int_\Theta l(x \mid \theta) \cdot \pi(\theta)\, d\theta}$$

Hence Likelihood Principle is satisfied by Bayesian inference. There are ways of implementing the likelihood principle: MLE and MAP $\Rightarrow$

The Maximum Likelihood Estimate (MLE)

The **maximum likelihood estimate** MLE, $\widehat{\theta}_{\mathrm{ML}}$ of θ, is defined by

$$\widehat{\theta}_{\mathrm{ML}} = \mathrm{argmax}_{\theta \in \Theta} f(x \mid \theta)$$

$$= \mathrm{argmin}_{\theta \in \Theta} ll(x; \theta)$$

MLE *is a parameter value that gives the observed x the highest possible probability*.

The Maximum A Posterior Estimate (MAP)

The **maximum a posterior estimate** MAP $\widehat{\theta}_{\mathrm{MAP}}$ of θ is defined by

$$\widehat{\theta}_{\mathrm{MAP}} = \mathrm{argmax}_{\theta \in \Theta} \pi(\theta \mid x)$$

Definition: Conjugate Family of Priors

A family $\mathcal{F}$ of probability distributions on Θ is said to be **conjugate** or **closed under sampling** for a likelihood function

$$l(x;\theta) = f(x \mid \theta).$$

if for every $\pi \in \mathcal{F}$, the posterior distribution $\pi(\theta|x)$ also belongs to $\mathcal{F}$. In practical terms this means that one can inegrate $m(x) = \int_\Theta f(x \mid \theta) \cdot \pi(\theta)\, d\theta$ explicitly. Above we have seen an example with the normal prior density on the mean of a normal. There are several additional examples of computation of $m(x)$ (and $\pi(\theta|x)$ in Allan Gut uses in *An Intermediate Course in Probability* Chapter 2.3-2.4, and exercises 2.30–2.35.

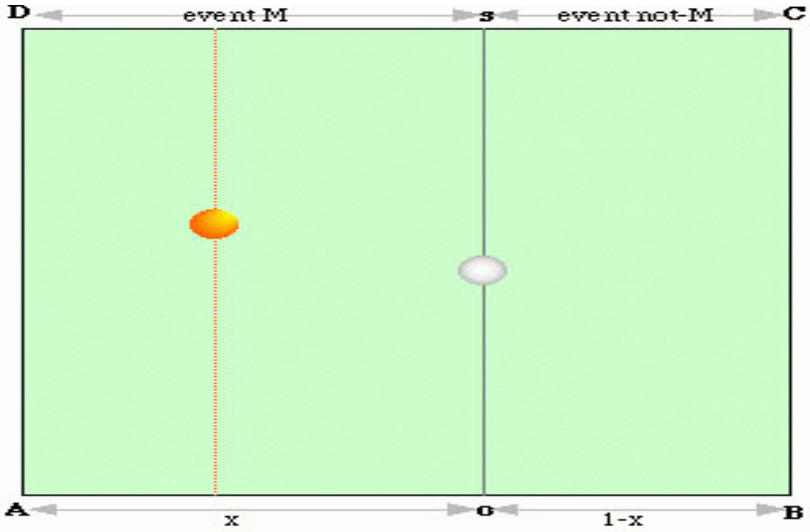
Definition: Conjugate Family of Priors

An intuitive way of understanding conjugate priors is that with conjugate priors the prior knowledge can be translated into equivalent sample information. See, e.g.,

$$\mathrm{N}\left(\frac{n\overline{x}/\sigma_0^2 + \mu/s^2}{n/\sigma_0^2 + 1/s^2}, \frac{1}{n/\sigma_0^2 + 1/s^2}\right).$$

Next we reconsider another problem with a conjugate family of priors.

Bayes' Billiard Ball



The square table is made in such a way that, if the White or Orange ball be thrown on it, the probability that it rests on any part of the plane is the same. First, the White ball is thrown, and suppose it rests on line **os**; then the Orange ball is thrown n times. We define event M as any event the Orange ball rests between **A** and **o**, and not-M as its resting between **o** and **B**. What this means is this:

The first throw of the White ball determines the value of probability x (i.e. the probability of an unknown event) from a uniform distribution between 0 and 1; and then a series of trials, with probability x of success (i.e., M) is generated (this provides the data on which to infer the correct value of x).

Then, thanks to the geometrical representation of the problem in the Figure, we can obtain the solution to the initial problem, by calculating integration. Although we have omitted mathematical formulas, the preceding is the central idea.

Bayes' Billiard Ball

A billiard ball W is rolled on a line of length one, with a uniform probability of stopping anywhere. It stops at p, not disclosed to us. A second ball O is rolled n times under the same assumptions and X denotes the number of times O stops to the left of W.
Given $X = x$, what inference can we make on p ?
(In the figure above $x \leftrightarrow p$.)

Modeling and Learning for Bayes' Billiard Ball

We let $\mathbf{P}$ be a random variable, whose values are denoted by p, $0 \leq p \leq 1$.
Parametric statistical model for rolls of Bayes' Billiard Ball O:
Conditional on $\mathbf{P} = p$, the rolls are outcomes of I.I.D $\mathrm{Be}(p)$ R.V's.

Modeling and Learning for Bayes' Billiard Ball

Hence for $x = 0, 1, 2, \ldots, n$,

$$f(x|p) = P(X = x \mid \mathbf{P} = p)$$

$$= \binom{n}{x} p^x \cdot (1-p)^{n-x},$$

(the Binomial distribution)

The Posterior Density

Bayes' rule

$$\pi(p \mid x) = \frac{f(x \mid p) \cdot \pi(p)}{\int_0^1 f(x \mid p) \cdot \pi(p)\, dp}, 0 \leq p \leq 1$$

and zero elsewhere. The marginal distribution of x is

$$m(x) = \int_0^1 f(x \mid p) \cdot \pi(p)\, dp.$$

The Posterior Density

The posterior $\pi(p \mid x)$ expresses our updated uncertainty of the 'true' position of W given the data $X = x$.

The Posterior Density

One way to get further from here is to use an explicit expression for $\pi(p)$. There are many possible choices (some more systematic choices outlined below), some have straightforward analytical advantages. Laplace assumed that $p \sim \mathrm{U}(0,1)$. i.e.,

$$\pi(p) = \begin{cases} 1 & 0 \leq p \leq 1 \\ 0 & \text{elsewhere,} \end{cases}$$

The marginal distribution of X: uniform prior

$$m(x) = \int_0^1 f(x \mid p) \cdot \pi(p)\, dp$$

$$= \binom{n}{x} \int_0^1 p^x \cdot (1-p)^{n-x}\, dp.$$

We use the Beta integral:

The Beta Integral

$$\int_0^1 p^{\alpha-1}(1-p)^{\beta-1}dp = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}.$$

Recall also that $\Gamma(x+1) = x!$, if x is a positive integer. $\alpha = \beta = 1$ gives the distribution $U(0,1)$. We set

$$B(\alpha,\beta) := \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}.$$

The Beta Density

$$\pi(p) = \begin{cases} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} p^{\alpha-1}(1-p)^{\beta-1} & 0 < p < 1 \\ 0 & \text{elsewhere.} \end{cases}$$

is a probability density $\mathcal{B}\mathrm{eta}(\alpha, \beta)$. $\alpha > 0$ and $\beta > 0$ are hyperparameters.

The marginal distribution of X: uniform prior

$$m(x) = \int_0^1 f(x \mid p) \cdot \pi(p)\, dp$$

$$= \binom{n}{x} \frac{x!(n-x)!}{(n+1)!}$$

The marginal distribution of x, $p \in U(0,1)$

$$m(x) = \int_0^1 f(x \mid p) \cdot dp = \left(\begin{array}{c} n \\ x \end{array} \right) \frac{x!(n-x)!}{(n+1)!}$$

$$= \frac{n!}{x!(n-x)!} \frac{x!(n-x)!}{(n+1)!} = \frac{1}{(n+1)}$$

There is an interpretation of Bayes' work claiming that the problem really attacked and solved by Bayes was: What should $\pi(p)$ be so that

$$\int_0^1 f(x \mid p) \cdot \pi(p) dp = \frac{1}{(n+1)}$$

holds for the Billiard Balls.

The Posterior Density for n rolls of Bayes' Orange Ball

$$\pi(p \mid x) = \frac{\binom{n}{x} p^x \cdot (1-p)^{n-x}}{m(x)}$$

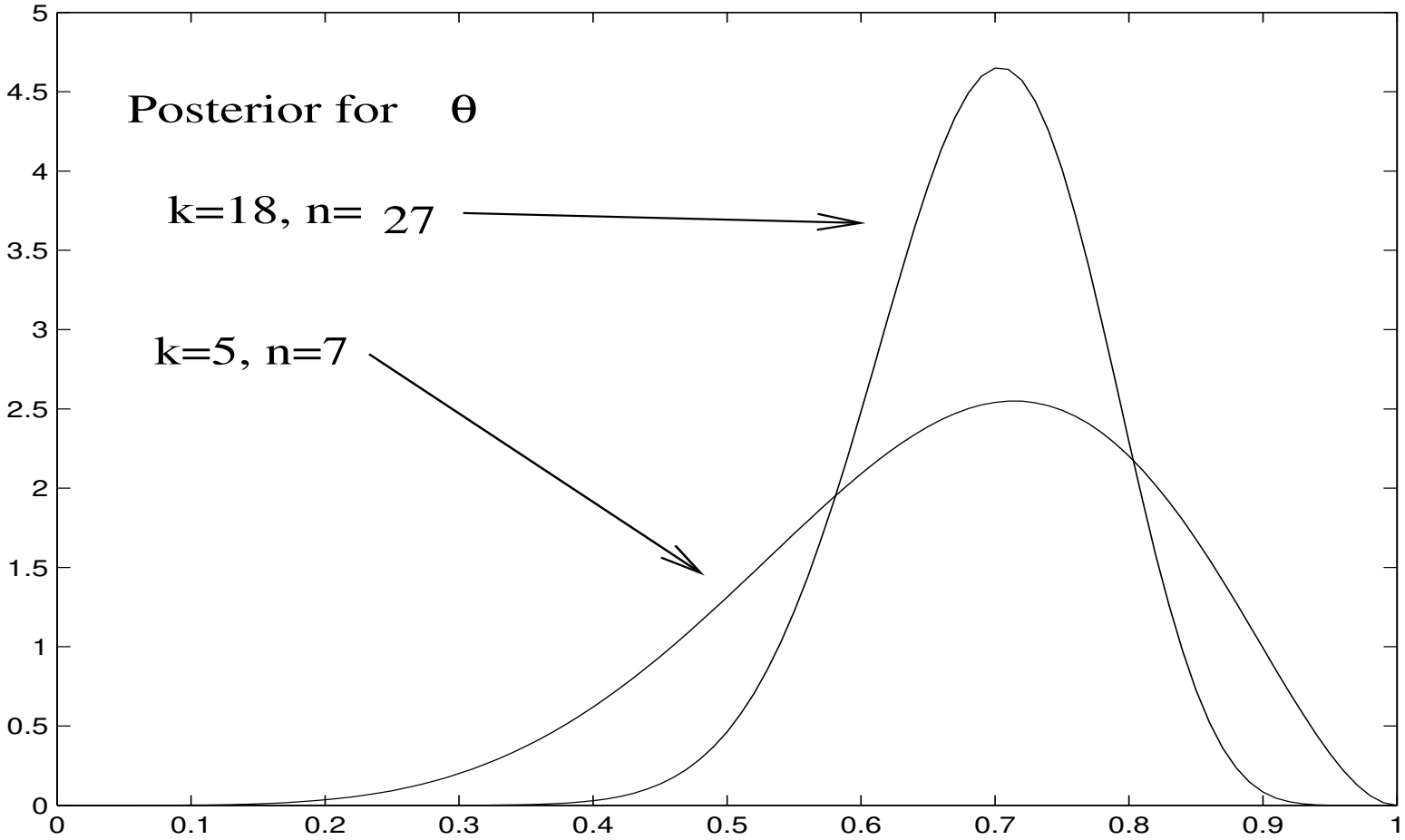
$$= \begin{cases} \frac{(n+1)!}{x!(n-x)!} \cdot p^k (1-p)^{n-k} & 0 \leq p \leq 1 \\ 0 & \text{elsewhere.} \end{cases}$$

This is again a Beta density, i.e., we have used a conjugate family of priors.

The Posterior Density for n rolls of Bayes' Orange Ball

$$\frac{(n+1)!}{x!(n-x)!} = \frac{\Gamma(n+2)}{\Gamma(x+1)\Gamma(n-x+1)} = \frac{1}{B(x+1, n-x+1)}.$$

The Posterior Density for n rolls of Bayes' Orange Ball



The densities in the picture do not have the proper constants.

The Posterior Density with n rolls of Bayes Ball, $p \in \mathcal{B}\mathrm{eta}(\alpha, \beta)$

$$\pi(p \mid x) = \begin{cases} \frac{1}{B(x+\alpha, n-x+\beta)} \cdot p^{x+\alpha-1}(1-p)^{\beta+n-x-1} & 0 \leq p \leq 1 \\ 0 & \text{elsewhere.} \end{cases}$$

This is the Beta density $\mathcal{B}\mathrm{eta}(\alpha + x, \beta + n - x)$.

The Maximum Likelihood Estimate of p in Bayes' Billiard

$$\widehat{p}_{\mathrm{ML}} = \mathrm{argmin}_{0\leq p\leq 1} ll\,(p \mid x)$$

$$= \mathrm{argmin}_{0\leq p\leq 1} \left[-\log \binom{n}{x} - x \log p - (n-x) \log (1-p) \right].$$

$$= \mathrm{argmin}_{0\leq p\leq 1} \left(-x \log p - (n-x) \log (1-p) \right).$$

$$\Rightarrow$$

$$\widehat{p}_{\mathrm{ML}} = \frac{x}{n}$$

If you observed $x = 0$, would you belive in the estimate $\widehat{p} = 0$ for all future purposes ?

MLE and Predictive Probability in Bayes' Billiard

The predictive probability

$$\frac{x+1}{n+2} = \frac{x+1}{(n+1)+1} = \int_0^1 p\pi(p|X=x)dp \tag{1}$$

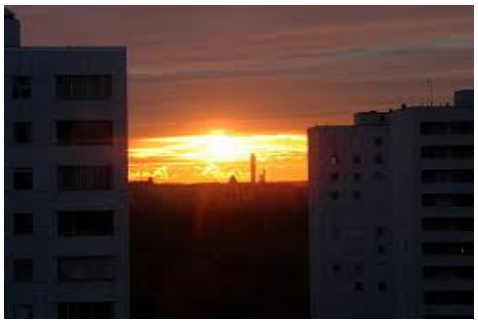
is a maximum likelihood estimate of p when $n+1$ rolls of the ball O and the first roll of the ball W are included in the data.

Sun Rises Tomorrow

Laplace exemplified the previous formula by the following.

$$P(\text{“sun rises tomorrow”}) = \frac{n+1}{n+2}$$

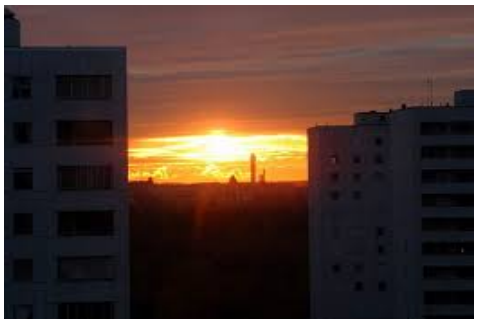
Laplace estimated the number sunsrises so far by finding the age of the universe from the Bible, and converting that to the number of days $= n$. Then in (1) $x = n$, as sun has risen every day so far. By a reading of the Bible the universe is 6000 yrs old, then $P(\text{“sun rises tomorrow”}) = 0.99999954$.



Sun Rises Tomorrow

Laplace is said to have added

> But this number (the probability of the sun coming up tomorrow) is far greater for him who, seeing in the totality of phenomena the principle regulating the days and seasons, realizes that nothing at present moment can arrest the course of it.



CS 547: Sensing and Planning in Robotics Gaurav S. Sukhatme Computer Science Robotic Embedded Systems Laboratory University of Southern California.

Used in Spring 2013

1. Examples of using probabilistic ideas in robotics
2. Reverend Bayes and review of probabilistic ideas
3. Introduction to Bayesian AI
4. Simple example of state estimation – robot and door to pass
5. Simple example of modeling actions
6. Bayes Filters.
7. Probabilistic Robotics

Q: How do we choose $\pi(\theta)$?

- Assessment (by Questionnaries)
- Conjugate prior
- Non-informative or reference prior
  - Laplace's prior
  - Jeffreys' prior
- Maximum entropy prior

Assessment of prior knowledge

(One form of) Bayesian statistics relies upon a **personalistic theory of probability** for quantification of prior knowledge. In such a theory

- probability measures the confidence that a particular individual (assessor) has in the truth of a particular proposition
- no attempt is made to specify which assessments are correct
- personal probabilities should satisfy certain postulates of coherence.
- A. O'Hagan: Eliciting Expert Beliefs in Substantial Practical Applications. *The Statistician* , 47, pp. 21–35, 1998.
- R.L. Keeney & D. von Winterfeldt: Eliciting Probabilities in Complex Technical Problems. *IEEE Transactions on Engineering Management*, 38, pp.191–201, 1991.

Choice of prior distributions by assessment

- C-A. S. Stael von Holstein: *Assessment and Evaluation of Subjective Probability Distributions*. 1970, Stockholm School of Economics.

Parametric Statistical Model, n I.I.D. $|\theta$ rv's

$$x_i|\theta \sim f(x|\theta), \text{ I.I.D. },$$

or **independent, identically, distributed conditional on** θ

$$x^{(n)} = (x_1, x_2, \ldots, x_n) \in \mathcal{X}^n$$

$f(x|\theta)$ is a probability density on R^p. $f(x|\theta)$ is a known function of x and θ. θ is an unknown parameter $\in \Theta =$ a vector space of finite dimension.

Asymptotic Shape of the Posterior

Asymptotically, for large n,

$$\pi(\theta|x^{(n)}) \approx f\left(x^{(n)}|\theta\right)$$

The influence of the prior vanishes.

Asymptotic Shape of the Posterior

Let us assume that $f(x|\theta)$ is a density with a scalar parameter (for simplicity of notation) , and that $f(x|\theta)$ is some $k \geq 2$ times differentiable in θ. We let $\widehat{\theta}_{ML}$ be the maximum likelihood estimate of θ. We expand the log likelihood function around $\widehat{\theta}_{ML}$

$$\log f\left(x^{(n)}|\theta\right) =$$

$$\log f\left(x^{(n)}|\widehat{\theta}_{ML}\right) + \left(\theta - \widehat{\theta}_{ML}\right) \frac{d}{d\theta} \log f\left(x^{(n)}|\widehat{\theta}_{ML}\right)$$

$$+\frac{1}{2}\left(\theta - \widehat{\theta}_{ML}\right)^2 \frac{d^2}{d\theta^2} \log f\left(x^{(n)}|\widehat{\theta}_{ML}\right) + R_n(\theta)$$

Asymptotic Shape of the Posterior

But here $\widehat{\theta}_{ML}$ is a solution of the equation

$$\frac{d}{d\theta} \log f\left(x^{(n)}|\widehat{\theta}_{ML}\right) = 0$$

Hence

$$\log f\left(x^{(n)}|\theta\right) =$$

$$\log f\left(x^{(n)}|\widehat{\theta}_{ML}\right) + \frac{1}{2}\left(\theta - \widehat{\theta}_{ML}\right)^2 \frac{d^2}{d\theta^2} \log f\left(x^{(n)}|\widehat{\theta}_{ML}\right) + R_n(\theta)$$

Asymptotic Shape of the Posterior: Law of Large Numbers

We have by assumption of I.I.D. data

$$\frac{d^2}{d\theta^2} \log f\left(x^{(n)}|\theta\right) = \sum_{l=1}^{n} \frac{d^2}{d\theta^2} \log f\left(x_l|\theta\right)$$

We set $Y_l = \frac{d^2}{d\theta^2} \log f\left(x_l|\theta\right)$. Then the Law of Large Numbers says that

$$\frac{1}{n} \sum_{l=1}^{n} Y_l \to E\left[Y\right], n \to \infty$$

where

$$E\left[Y\right] = \int_{\mathcal{X}} \frac{d^2}{d\theta^2} \log f\left(x|\theta\right) f\left(x \mid \theta\right) dx$$

Asymptotic Shape of the Posterior: Fisher Information

The integral

$$I(\theta) = -\int_{\mathcal{X}} \frac{d^2}{d\theta^2} \log f(x|\theta)\, f(x \mid \theta)\, dx$$

is called *Fisher information*.

Asymptotic Shape of the Posterior: Fisher Information

Then we may feel inclined to believe that

$$\frac{d^2}{d\theta^2} \log f\left(x^{(n)}|\widehat{\theta}_{ML}\right) = \sum_{l=1}^{n} \frac{d^2}{d\theta^2} \log f\left(x_l|\widehat{\theta}_{ML}\right) \approx -n \cdot I\left(\widehat{\theta}_{ML}\right)$$

Note that even $\widehat{\theta}_{ML}$ depends on n.

Asymptotic Shape of the Posterior

This gives

$$\log f\left(x^{(n)}|\theta\right) \approx$$

$$\log f\left(x^{(n)}|\widehat{\theta}_{ML}\right) - \frac{1}{2}\left(\theta - \widehat{\theta}_{ML}\right)^2 n \cdot I\left(\widehat{\theta}_{ML}\right)$$

The first term does not involve θ.

Asymptotic Shape of the Posterior

Then

$$f\left(x^{(n)}|\theta\right) \approx e^{-\frac{n}{2}\left(\theta-\widehat{\theta}_{ML}\right)^2 \cdot I\left(\widehat{\theta}_{ML}\right)}$$

The interpretation of the relation is that the likelihood function can be for large n be approximated by a normal density for which the mean is $\widehat{\theta}_{ML}$ and the variance is $\frac{1}{nI\left(\widehat{\theta}_{ML}\right)}$.