ECE295, Data Assimilation and Inverse Problems, Spring 2015

1 April, Intro; Linear discrete Inverse problems (Aster Ch 1 and 2) Slides

8 April, SVD (Aster ch 2 and 3) Slides

15 April, Regularization (ch 4)

22 April, Sparse methods (ch 7.2-7.3), radar

29 April, more on Sparse

6 May, Bayesian methods and Monte Carlo methods (ch 11), Markov Chain Monte Carlo

13 May, Introduction to sequential Bayesian methods, Kalman Filter (KF)

20 May, Gaussian Mixture Model (Nima)

27 May, Ensemple Kalman Filer (EnKF)

3 June, EnKF, Particle Filter,

Homework:

Just email the code to me (I dont need anything else).

Call the files LastName_ExXX.

Homework is **due 8am on Wednesday.**

8 April: Hw 1: Download the matlab codes for the book (cd_5.3) from this website

15 April: SVD analysis:

SVD homework. You can also try replacing the matrix in the Shaw problem with the beamforming sensing matrix. The sensing matrix is available here .

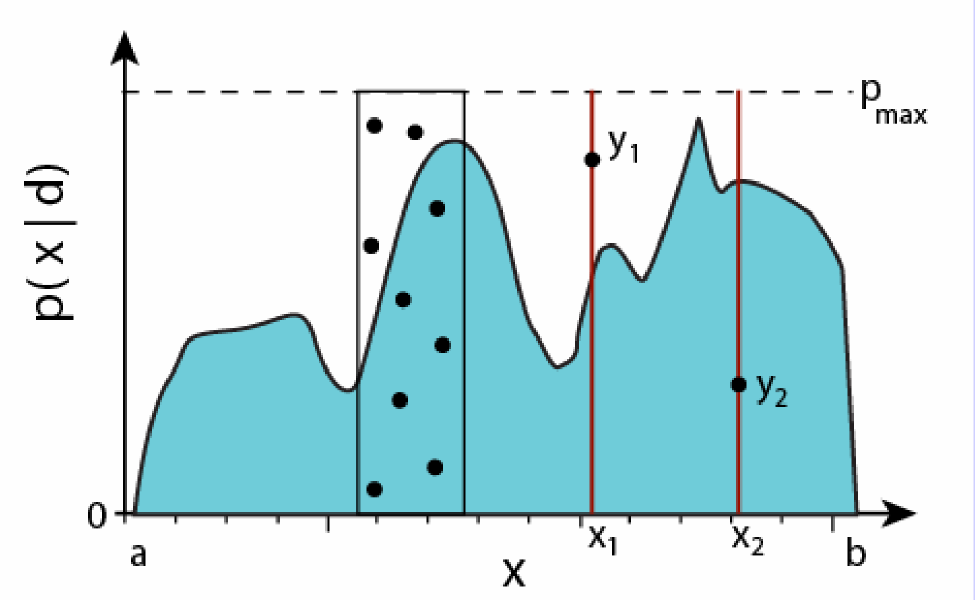
22 April

Late April: Beamforming

May: Ice-flow from GPS

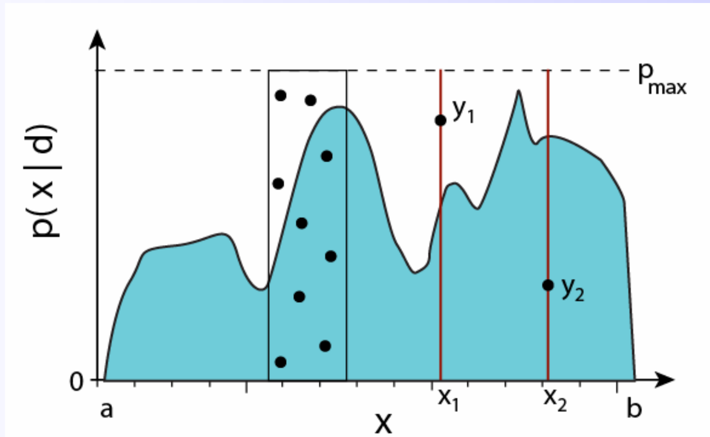
Generating samples from an arbitrary posterior PDF

- The rejection method



Generating samples from the posterior PDF

- Rejection method



But this requires us to know P_{max}

Step 1: generate a uniform random variable, x_i between a and b

$$p(x_i) = \frac{1}{(b-a)}, \quad a \leq x_i \leq b$$

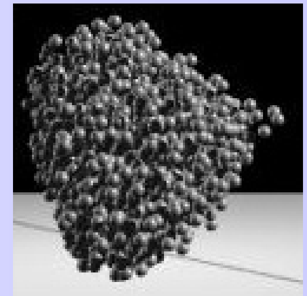
Step 2: generate a second uniform random variable, y_i

$$p(y_i) = \frac{1}{p_{max}}, \quad 0 \leq y_i \leq p_{max}$$

Step 3: accept x_i if $y_i \leq p(x_i|d)$ otherwise reject

Step 4: go to step 1

Monte Carlo integration



Consider any integral of the form

$$I = \int_{\mathcal{M}} f(\boldsymbol{m}) p(\boldsymbol{m}|d) d\boldsymbol{m}$$

Given a set of samples m_i (i=,...,N_s) with sampling density $h(m_i)$, the Monte Carlo approximation to I is given by

$$I \approx \sum_{i=1}^{N_s} \frac{f(\boldsymbol{m}_i) p(\boldsymbol{m}_i|d)}{h(\boldsymbol{m}_i)}$$

Only need to know p(m | d) to a multiplicative constant

If the sampling density is proportional to $p(m_i \mid d)$ then,

$$h(\boldsymbol{m}) = N_s \times p(\boldsymbol{m}|\boldsymbol{d})$$

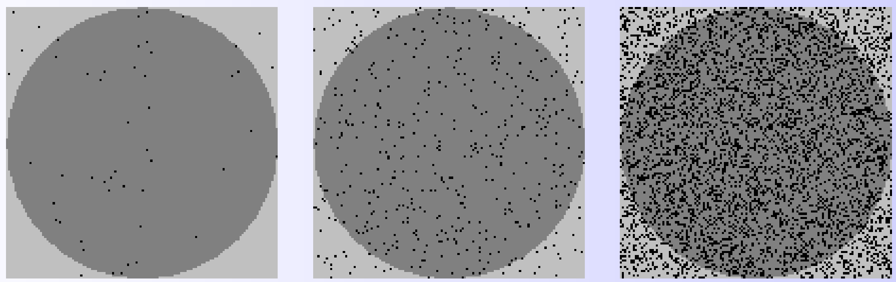
$$\Rightarrow \quad I \approx \frac{1}{N_s} \sum_{i=1}^{N_s} f(\boldsymbol{m}_i)$$

The variance of the $f(m_i)$ values gives the numerical integration error in I

Example: Monte Carlo integration

Finding the area of a circle by throwing darts

$$I = \int_A f(\boldsymbol{m}) d\boldsymbol{m}$$



$$f(\boldsymbol{m}) = \begin{cases} 1 & \boldsymbol{m} \text{ inside circle} \\ 0 & \text{otherwise} \end{cases}$$

$$h(\boldsymbol{m}) = \frac{N_s}{A}$$

$$I \approx \frac{1}{N_s} \sum_{i=1}^{N_S} f(\boldsymbol{m}_i)$$

$$\approx \frac{\text{Number of points inside the circle}}{\text{Total number of points}}$$

Monte Carlo integration

We have

$$I = \int_{\mathcal{M}} f(\boldsymbol{m}) p(\boldsymbol{m}|d) d\boldsymbol{m} \approx \sum_{i=1}^{N_s} \frac{f(\boldsymbol{m}_i) p(\boldsymbol{m}_i|d)}{h(\boldsymbol{m}_i)} \approx \frac{1}{N_s} \sum_{i=1}^{N_s} f(\boldsymbol{m}_i)$$

The variance in this estimate is given by

$$\sigma_I^2 = \frac{1}{N_s} \left\{ \frac{1}{N_s^2} \sum_{i=1}^{N_s} f^2(\boldsymbol{m}_i) - \left(\frac{1}{N_s} \sum_{i=1}^{N_s} f(\boldsymbol{m}_i) \right)^2 \right\}$$

- To carry out MC integration of the posterior we ONLY NEED to be able to evaluate the integrand up to a multiplicative constant.
- As the number of samples, N_s, grows the error in the numerical estimate will decrease with the square root of N_s.
- In principal any sampling density h(m) can be used but the convergence rate will be fastest when h(m) $\propto$ p(m | d).

What useful integrals should one calculate using samples distributed according to the posterior p(m | d)?

In low dimensions, these volume and radius formulas simplify to the following:

Dimension	Volume of a ball of radius *R*	Radius of a ball of volume *V*
0	1	All balls have volume 1
1	$2R$	$V/2$
2	πR^2	$\frac{V^{1/2}}{\sqrt{\pi}}$
3	$\frac{4}{3}\pi R^3$	$\left(\frac{3V}{4\pi}\right)^{1/3}$
4	$\frac{\pi^2}{2}R^4$	$\frac{(2V)^{1/4}}{\sqrt{\pi}}$
5	$\frac{8\pi^2}{15}R^5$	$\left(\frac{15V}{8\pi^2}\right)^{1/5}$
6	$\frac{\pi^3}{6}R^6$	$\frac{(6V)^{1/6}}{\sqrt{\pi}}$
7	$\frac{16\pi^3}{105}R^7$	$\left(\frac{105V}{16\pi^3}\right)^{1/7}$
8	$\frac{\pi^4}{24}R^8$	$\frac{(24V)^{1/8}}{\sqrt{\pi}}$
9	$\frac{32\pi^4}{945}R^9$	$\left(\frac{945V}{32\pi^4}\right)^{1/9}$
10	$\frac{\pi^5}{120}R^{10}$	$\frac{(120V)^{1/10}}{\sqrt{\pi}}$

The volume of a N-dim cube
2^N

For N=2
3.14/2^2=3/4

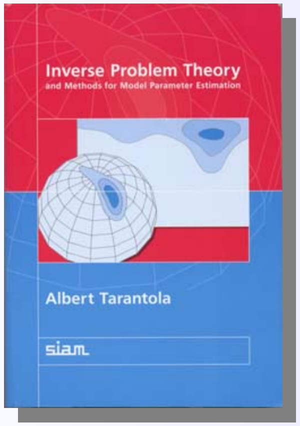
For N=10
3.14^5/120/2^10=2/1000

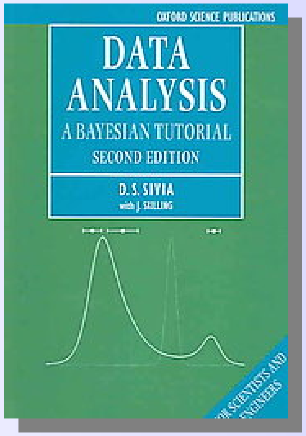
This will be hard!

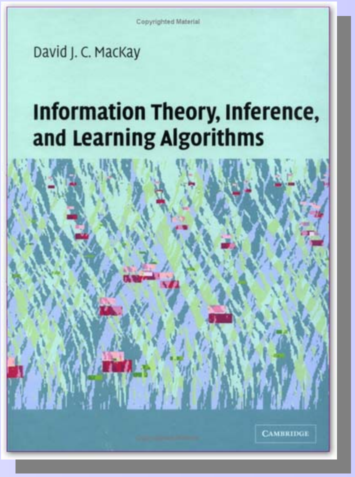
Probabilistic inference

Bayes theorem and all that....

Books

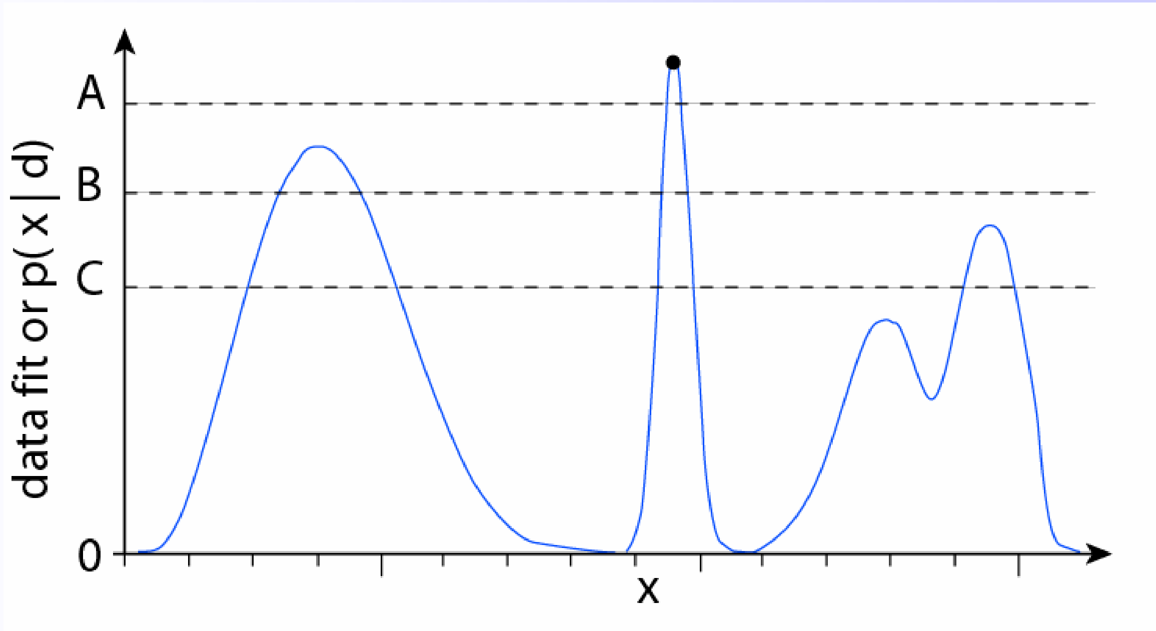

Inverse Problem Theory
and Methods for Model Parameter Estimation
Albert Tarantola
siam


OXFORD SCIENCE PUBLICATIONS
DATA ANALYSIS
A BAYESIAN TUTORIAL
SECOND EDITION
D. S. SIVIA
with J. SKILLING
FOR SCIENTISTS AND ENGINEERS


Copyrighted Material
David J. C. MacKay
Information Theory, Inference, and Learning Algorithms
CAMBRIDGE

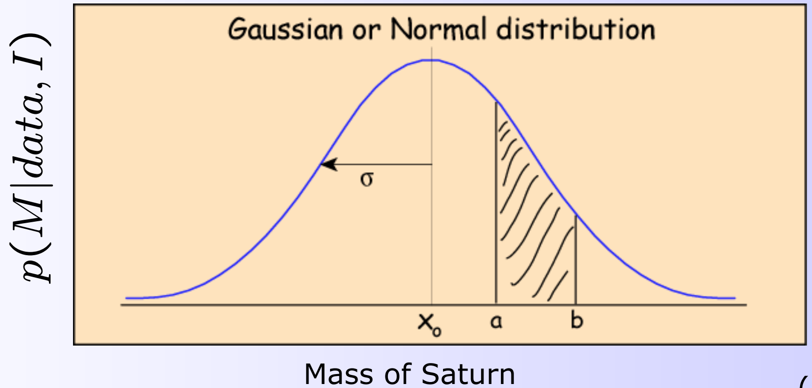
Highly nonlinear inverse problems

Multi-modal data misfit/objective function



What value is there in an optimal model ?

Probabilistic inference: History



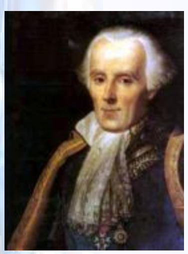
$$\int_{-\infty}^{\infty} p(x)dx = 1$$

(Laplace 1812)

$$Pr(x : a \leq x \leq b) = \int_a^b p(x)dx$$

We have already met the concept of using a probability density function p(x) to describe the state of a random variable.

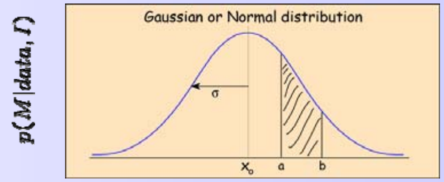
In the probabilistic (or *Bayesian*) approach, probabilities are also used to describe *inferences* (or *degrees of belief)* about x even if x itself is not a random variable.



Laplace (1812) rediscovered the work of Bayes (1763), and used it to constrain the mass of Saturn. In 150 years the estimate changed by only 0.63% !

But Laplace died in 1827 and then the arguments started...

Bayesian or Frequentist: the arguments



Mass of Saturn

Some thought that using probabilities to describe degrees of belief was too subjective and so they redefined probability as the *long run relative frequency* of a random event. This became the *Frequentist* approach.

To estimate the mass of Saturn the frequentist has to relate the mass to the data through a *statistic*. Since the data contain \`random' noise probability theory can be applied to the statistic (which becomes the random variable !). This gave birth to the field of statistics !

But how to choose the statistic ?

\`.. a plethora of tests and procedures without any clear underlying rationale'

(D. S. Sivia)

\`Bayesian is subjective and requires too many guesses'

A. Frequentist

\`Frequentist is subjective, but BI can solve problems more completely'

A. Bayesian

For a discussion see Sivia (2005, pp 8-11).

Data Analysis: A Bayesian Tutorial' 2nd Ed. D. S. Sivia with J. Skilling, O.U.P. (2005)

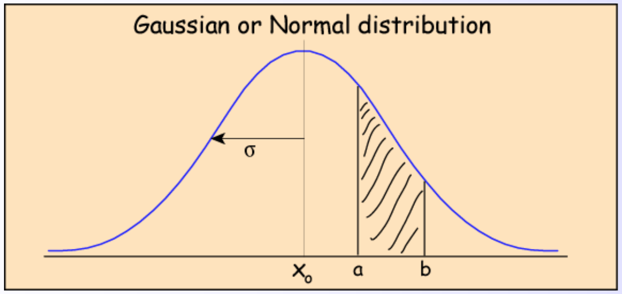
Probability theory: Joint probability density functions

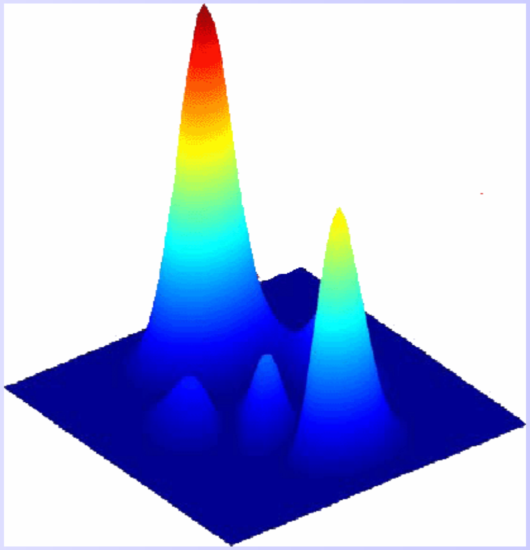
A PDF for variable x

$$p(x)$$

Joint PDF of x and y

$$p(x,y)$$





Probability is proportional to
area under the curve or surface

If x and y are independent their joint PDF is separable

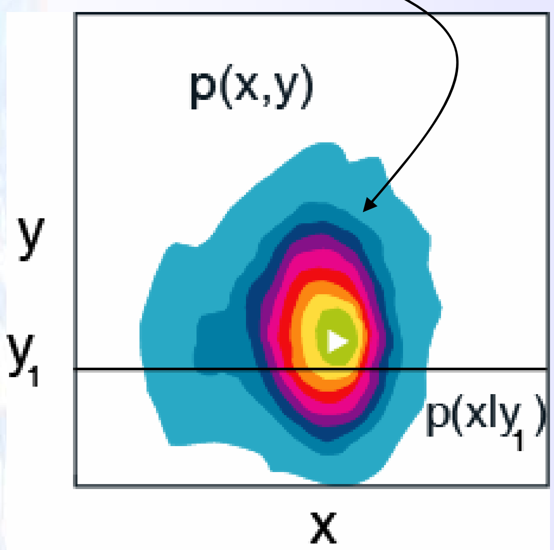
$$p(x,y) = p(x) \times p(y)$$

Probability theory: Conditional probability density functions

Joint PDF of x and y

"The PDF of x and y taken together"

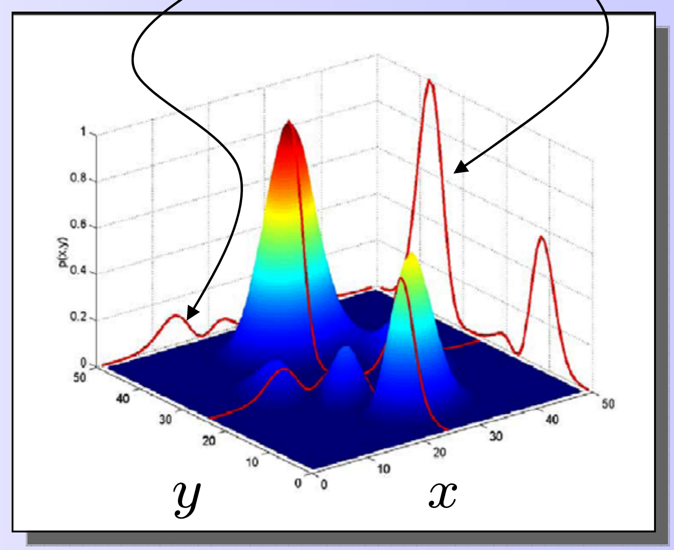
$$p(x, y)$$



Conditional PDFs

"The PDF of x given a value for y"

$$p(x|y) \qquad p(y|x)$$



Relationship between joint and conditional PDFs

$$p(x, y) = p(x|y) \times p(y)$$

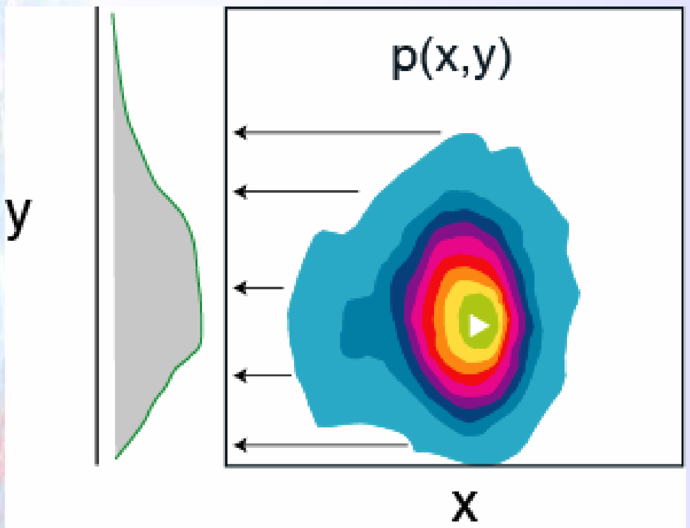
Probability theory: Marginal probability density functions

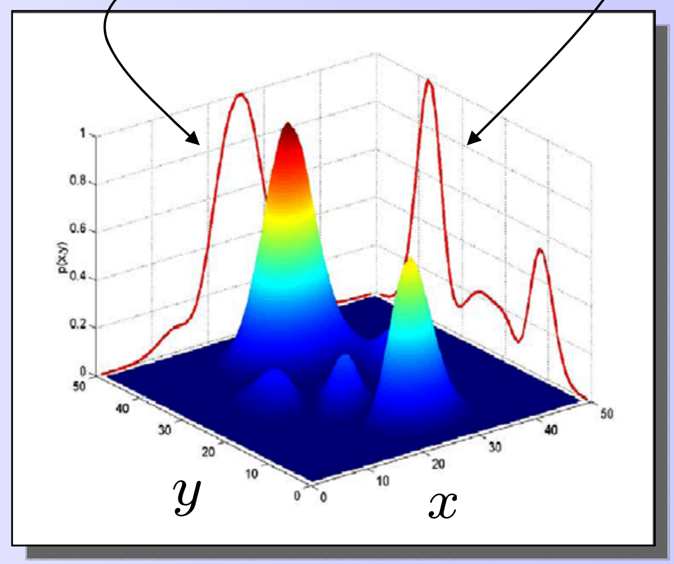
Marginal PDFs

A marginal PDF is a summation of probabilities

$$p(y) = \int p(x,y)dx$$

$$p(x) = \int p(x,y)dy$$



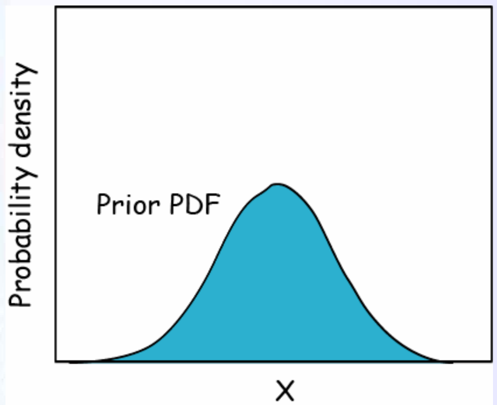


Relationship between joint, conditional and marginal PDFs

$$p(x,y) = p(x|y) \times p(y)$$

Prior probability density functions

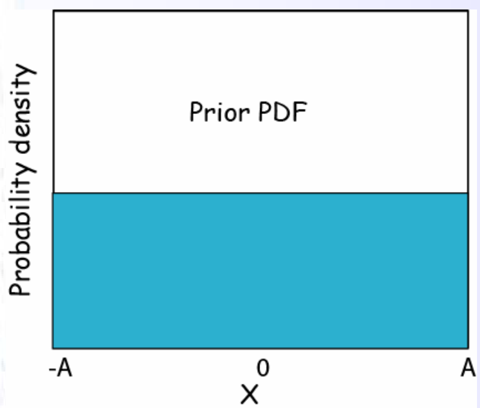
What we know from previous experiments, or what we guess...



$$p(x) = k \exp\left\{-\frac{(x-x_o)^2}{2\sigma^2}\right\}$$

$$p(\boldsymbol{m}) = k \exp\left\{-\frac{1}{2}(\boldsymbol{m}-\boldsymbol{m}_o)^T C_m^{-1}(\boldsymbol{m}-\boldsymbol{m}_o)\right\}$$

Beware: there is no such thing as a non-informative prior

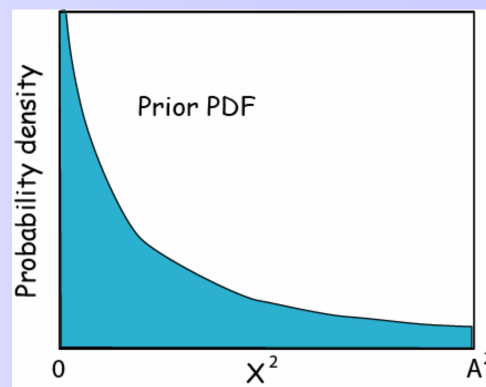


$$p(x)dx = p(y)dy$$

$$p(y) = p(x)\frac{dx}{dy}$$

$$p(x) = C$$

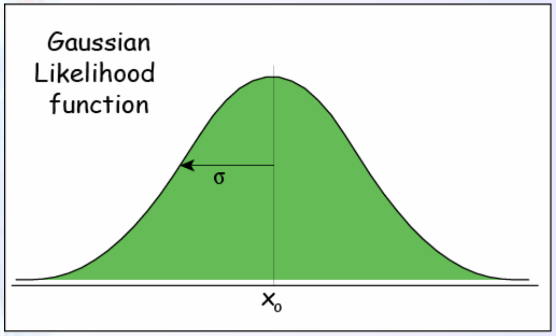
$$p(x^2) = \frac{C}{2x}$$



As A $\rightarrow \infty$ this is not proper !

Likelihood functions

The likelihood that the data would have occurred for a given model



$$p(d_i|x) = \exp\left\{-\frac{(x - x_{o,i})^2}{2\sigma_i^2}\right\}$$

$$p(\boldsymbol{d}|\boldsymbol{m}) = \exp\left\{-\frac{1}{2}(\boldsymbol{d} - G\boldsymbol{m})^T C_D^{-1}(\boldsymbol{d} - G\boldsymbol{m})\right\}$$

Maximizing likelihoods is what Frequentists do. It is what we did earlier.

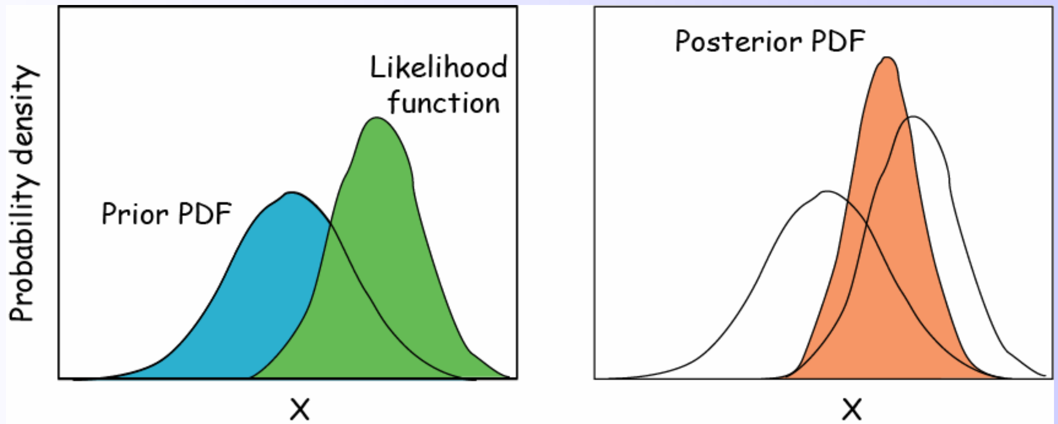
$$\max_{\boldsymbol{m}} \; p(\boldsymbol{d}|\boldsymbol{m}) = \min_{\boldsymbol{m}} - \ln(p(\boldsymbol{d}|\boldsymbol{m}))$$

$$= \min_{\boldsymbol{m}} \; (\boldsymbol{d} - G\boldsymbol{m})^T C_D^{-1}(\boldsymbol{d} - G\boldsymbol{m})$$

Maximizing the likelihood = minimizing the data prediction error

Bayes' theorem

All information is expressed in terms of probability density functions



Bayes' rule (1763)

1702-1761

Conditional PDFs   Assumptions

$$p(\boldsymbol{m}|\boldsymbol{d}, I) \propto p(\boldsymbol{d}|\boldsymbol{m}, I) \times p(\boldsymbol{m}|I)$$

data   model parameters

Posterior probability density $\propto$ Likelihood x Prior probability density

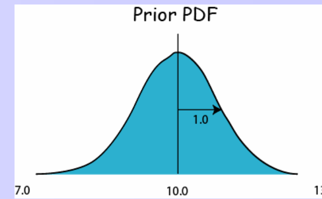
What is known after the data are collected

Measuring fit to data

What is known before the data are collected

Example: Measuring the mass of an object

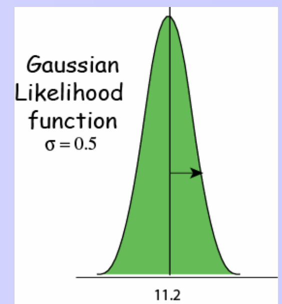
If we have an object whose mass, *m*, we which to determine. Before we collect any data we believe that its mass is approximately $10.0 \pm 1\mu$g. In probabilistic terms we could represent this as a Gaussian prior distribution



prior

$$p(m) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(m-10.0)^2}$$

Suppose a measurement is taken and a value 11.2 μ g is obtained, and the measuring device is believed to give Gaussian errors with mean 0 and σ = 0.5 μ g. Then the likelihood function can be written



$$p(d|m) = \frac{1}{0.5\sqrt{2\pi}} e^{-2(m-11.2)^2}$$

Likelihood

$$p(m|d) = \frac{1}{\pi} e^{-\frac{1}{2}(m-10.0)^2-2(m-11.2)^2}$$

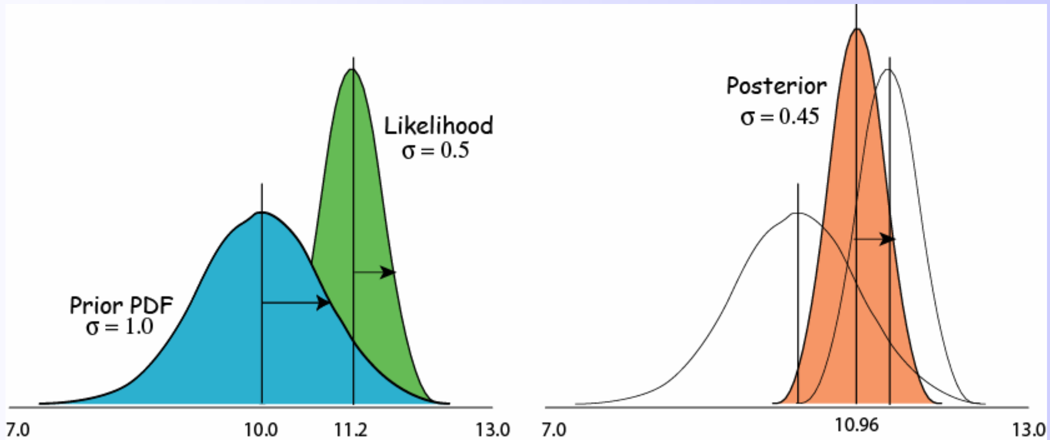
Posterior

$$p(m|d) \propto e^{\frac{-\frac{1}{2}(m-10.96)^2}{1/5}}$$

The posterior PDF becomes a Gaussian centred at the value of 10.96 μ g with standard deviation $\sigma = (1/5)^{1/2} \approx 0.45$.

Example: Measuring the mass of an object

The more accurate new data has changed the estimate of *m* and decreased its uncertainty



One data point problem

For the general linear inverse problem we would have

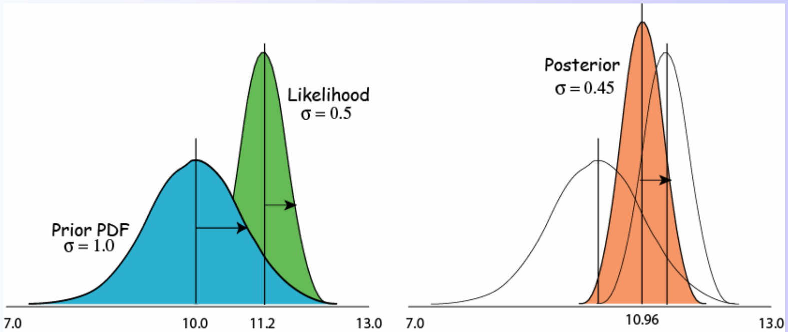
Prior: $$p(\boldsymbol{m}) \propto \exp\left\{-\frac{1}{2}(\boldsymbol{m}-\boldsymbol{m}_o)^T C_m^{-1}(\boldsymbol{m}-\boldsymbol{m}_o)\right\}$$

Likelihood: $$p(\boldsymbol{d}|\boldsymbol{m}) \propto \exp\left\{-\frac{1}{2}(\boldsymbol{d}-G\boldsymbol{m})^T C_d^{-1}(\boldsymbol{d}-G\boldsymbol{m})\right\}$$

Posterior PDF

$$\propto \exp\left\{-\frac{1}{2}[(\boldsymbol{d}-G\boldsymbol{m})^T C_d^{-1}(\boldsymbol{d}-G\boldsymbol{m}) + (\boldsymbol{m}-\boldsymbol{m}_o)^T C_m^{-1}(\boldsymbol{m}-\boldsymbol{m}_o)]\right\}$$

Product of Gaussians=Gaussian:



One data point problem

For the general linear inverse problem we would have

Prior: $$p(\boldsymbol{m}) \propto \exp\left\{-\frac{1}{2}(\boldsymbol{m}-\boldsymbol{m}_o)^T C_m^{-1}(\boldsymbol{m}-\boldsymbol{m}_o)\right\}$$

Likelihood: $$p(\boldsymbol{d}|\boldsymbol{m}) \propto \exp\left\{-\frac{1}{2}(\boldsymbol{d}-G\boldsymbol{m})^T C_d^{-1}(\boldsymbol{d}-G\boldsymbol{m})\right\}$$

Posterior PDF

$$\propto \exp\left\{-\frac{1}{2}[(\boldsymbol{d}-G\boldsymbol{m})^T C_d^{-1}(\boldsymbol{d}-G\boldsymbol{m}) + (\boldsymbol{m}-\boldsymbol{m}_o)^T C_m^{-1}(\boldsymbol{m}-\boldsymbol{m}_o)]\right\}$$

$$\propto \exp\left\{-\frac{1}{2}[\mathbf{m}-\hat{\mathbf{m}}]^T \mathbf{S}^{-1}[\mathbf{m}-\hat{\mathbf{m}}]\right\}$$

$$\mathbf{S}^{-1} = \mathbf{G}^T\mathbf{C}_d^{-1}\mathbf{G} + \mathbf{C}_m^{-1}$$

$$\hat{\mathbf{m}} = \left(\mathbf{G}^T\mathbf{C}_d^{-1}\mathbf{G} + \mathbf{C}_m^{-1}\right)^{-1}\left(\mathbf{G}^T\mathbf{C}_d^{-1}\mathbf{d} + \mathbf{C}_m^{-1}\mathbf{m}_0\right)$$

$$= \mathbf{m}_0 + \left(\mathbf{G}^T\mathbf{C}_d^{-1}\mathbf{G} + \mathbf{C}_m^{-1}\right)^{-1}\mathbf{G}^T\mathbf{C}_d^{-1}\left(\mathbf{d} - \mathbf{G}\mathbf{m}_0\right)$$

The biased coin problem

Suppose we have a suspicious coin and we want to know if it is biased or not ?

Let α be the probability that we get a head.

$$0 \le \alpha \le 1$$

$\alpha = 1$: means we always get a head.

$\alpha = 0$: means we always get a tail.

$\alpha = 0.5$: means equal likelihood of head or tail.

We can collect data by tossing the coin many times

$$\{H, T, T, H, \ldots\}$$

We seek a probability density function for α given the data

$$p(\alpha|\boldsymbol{d}, I) \propto p(\boldsymbol{d}|\alpha, I) \times p(\alpha|I)$$

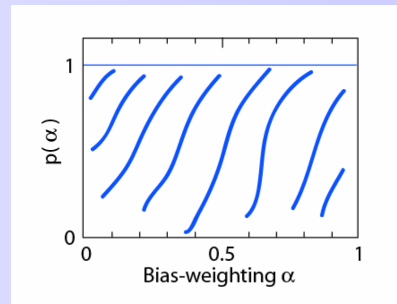
Posterior PDF $\propto$ Likelihood x Prior PDF

The biased coin problem

What is the prior PDF for α ?

Let us assume that it is uniform

$$p(\alpha|I) = 1, \quad 0 \leq \alpha \leq 1$$



What is the Likelihood function ?

The probability of observing R heads out of N coin tosses is

$$p(\boldsymbol{d}|\alpha, I) \propto \alpha^{R}(1-\alpha)^{N-R}$$

$$p(\alpha|\boldsymbol{d}, I) \propto p(\boldsymbol{d}|\alpha, I) \times p(\alpha|I)$$

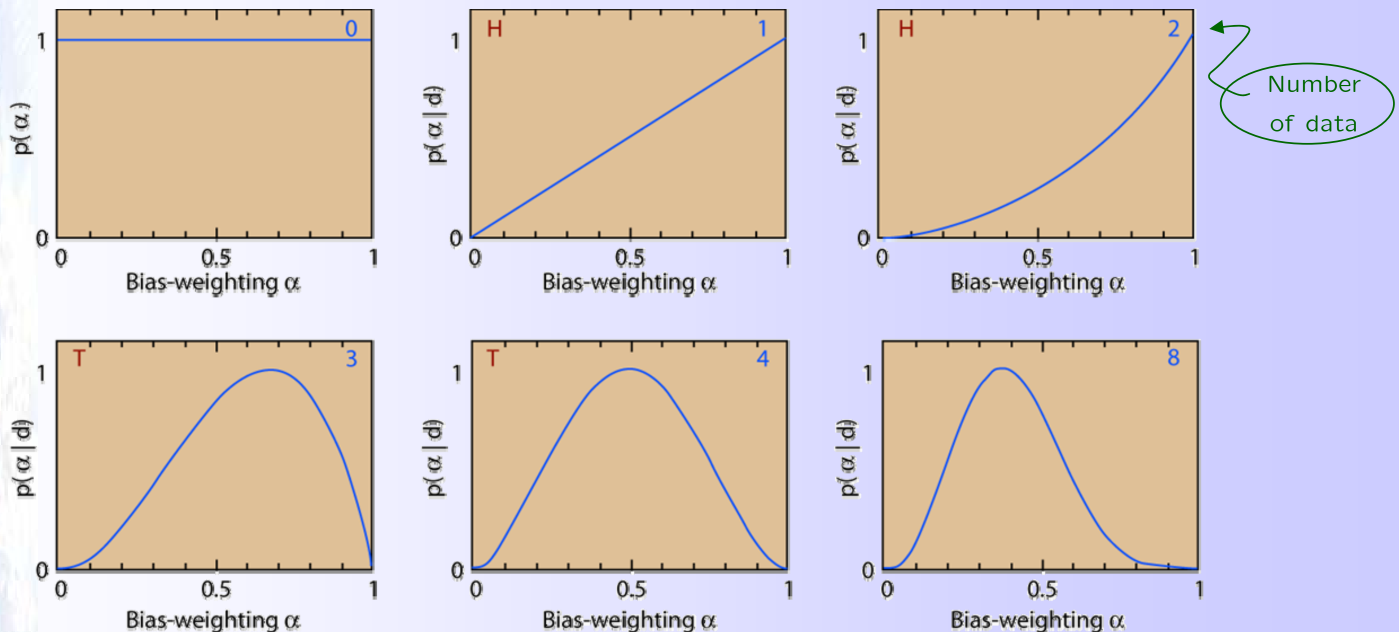
Posterior PDF $\propto$ Likelihood x Prior PDF

The biased coin problem

We have the posterior PDF for α given the data and our prior PDF

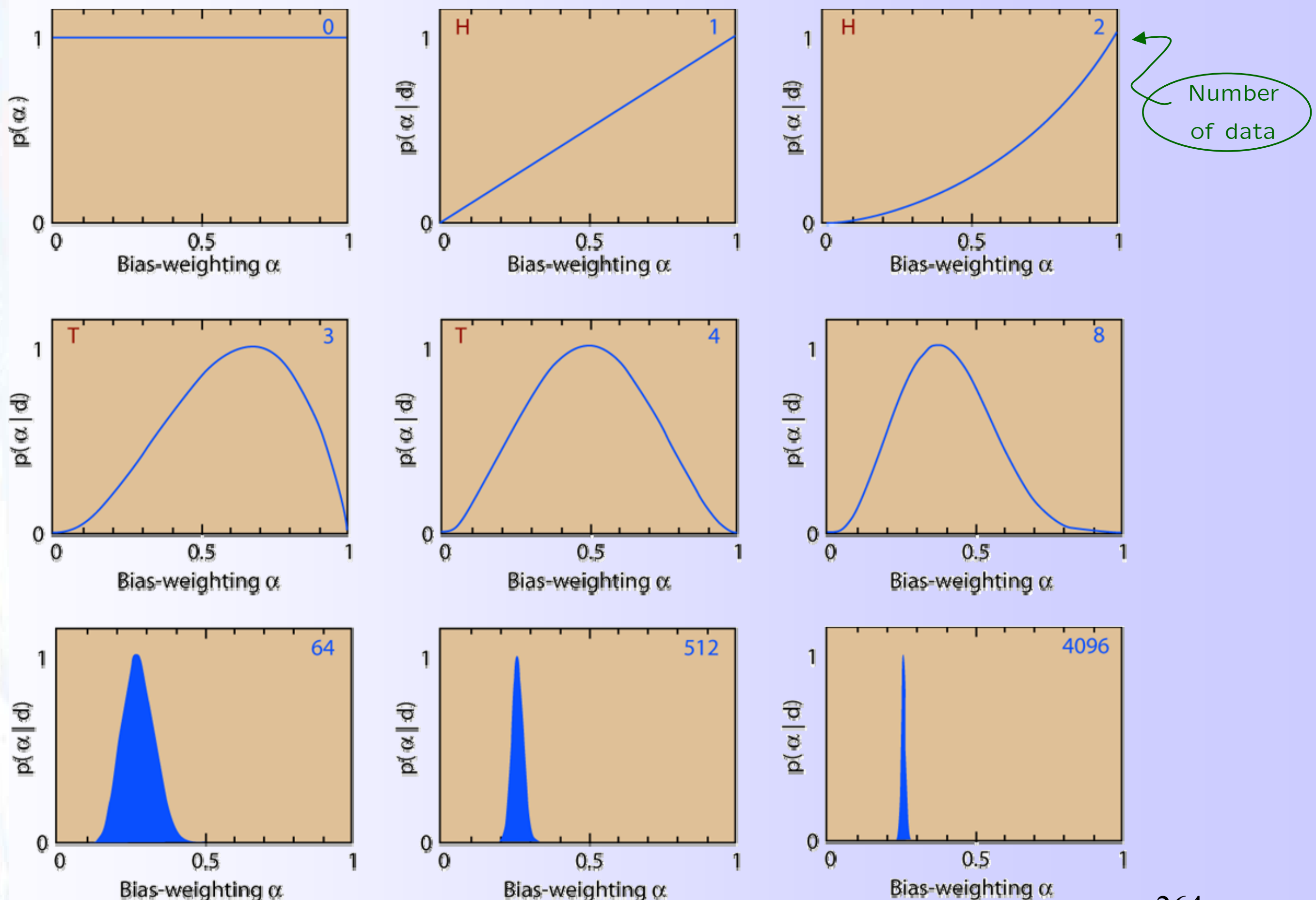
$$p(\alpha|\boldsymbol{d}, I) \propto \alpha^{R}(1-\alpha)^{N-R}$$

After N coin tosses let R = number of heads observed. Then we Can plot the probability density for p(α | d) as data are collected



The biased coin problem

$$p(\alpha|\boldsymbol{d}, I) \propto \alpha^{R}(1-\alpha)^{N-R}$$



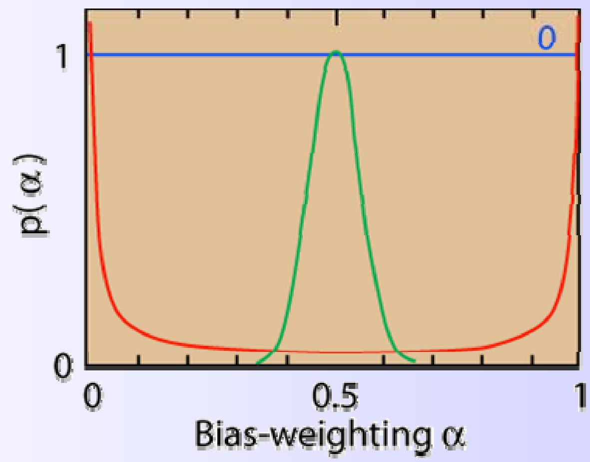
The biased coin problem

But what if three people had different opinions about the coin prior to collecting the data ?

Dr. Blue knows nothing about the coin.

Dr. Green thinks the coin is likely to be almost fair.

Dr. Red thinks the coin is either highly biased to heads or tails.



$$p(\boldsymbol{d}|\alpha, I) \propto \alpha^R (1-\alpha)^{N-R}$$

The biased coin problem