Bayesian Statistics

Introduction to Scientific Reasoning

Baback Moghaddam

baback @ jpl.nasa.gov

Machine Learning Group

Acknowledgements

Roughly 70% of these slides are from

Aaron Hertzmann
University of Toronto

SIGGRAPH 2004 Tutorial

Intro to Bayesian Learning

Some of the original slides are “retooled” a bit

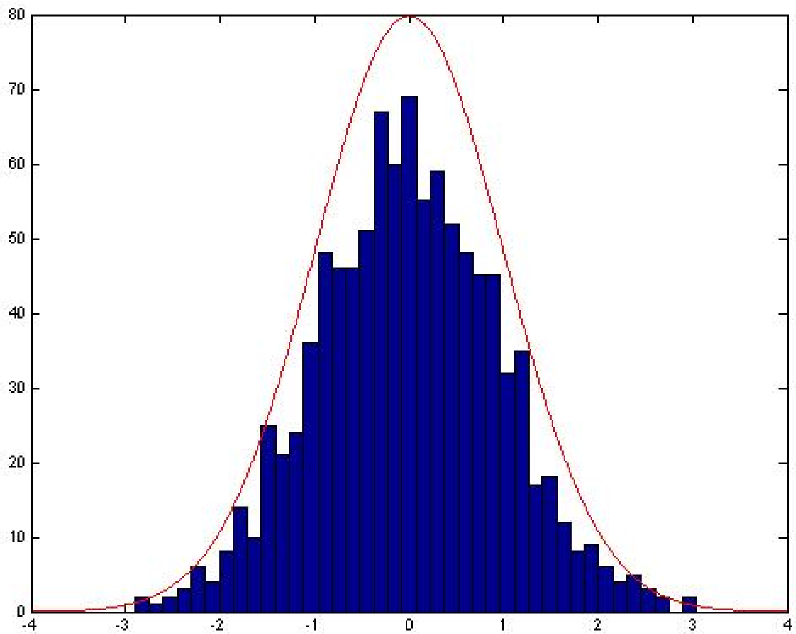
Key problems

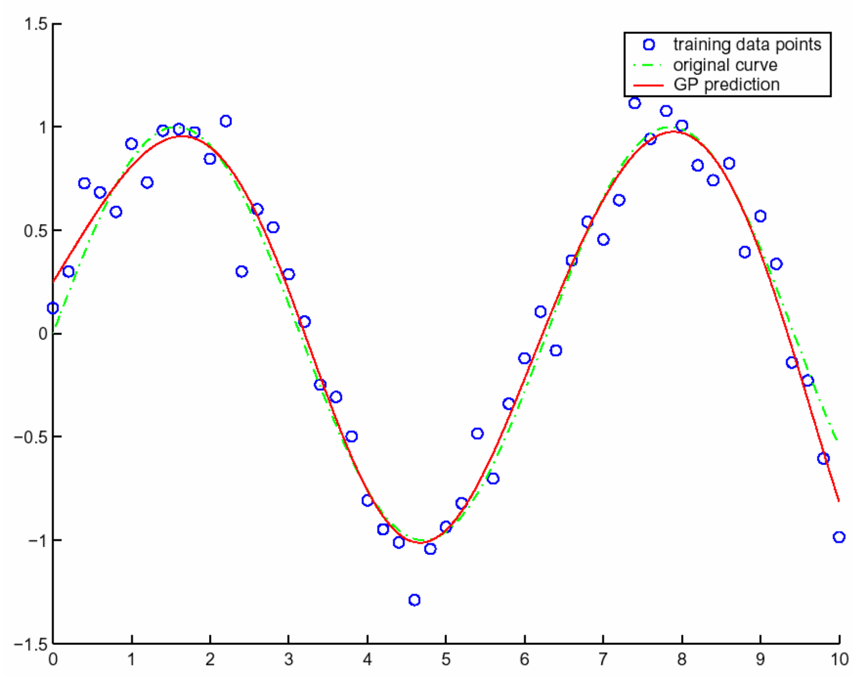
- **How do you fit a model to data?**
  - How do you choose weights and thresholds?
  - How do you incorporate prior knowledge?
  - How do you merge multiple sources of info?
  - How do you model uncertainty?

Bayesian reasoning provides solutions

Bayesian reasoning is …

Probability, statistics, data-fitting





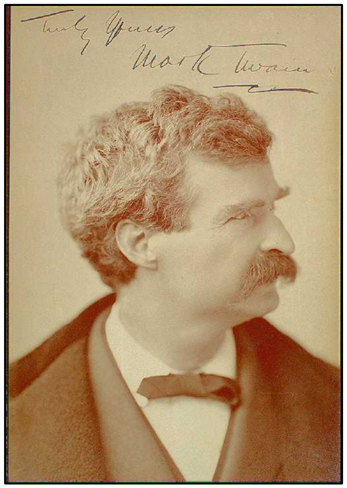
Applications

- Data mining
- Robotics
- Signal processing
- Document Analysis
- Marketing
- Bioinformatics
- Astronomy, *etc*

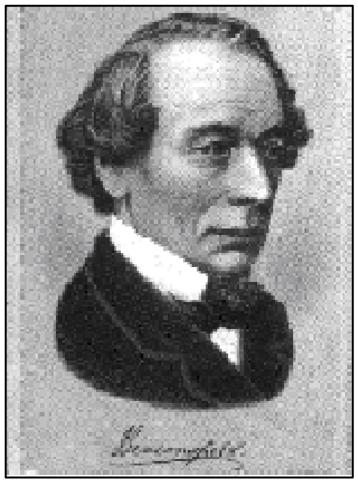
In fact, it applies to *all* data-driven fields

Statistics: A Bad Rap

Mark Twain



Benjamin Disraeli



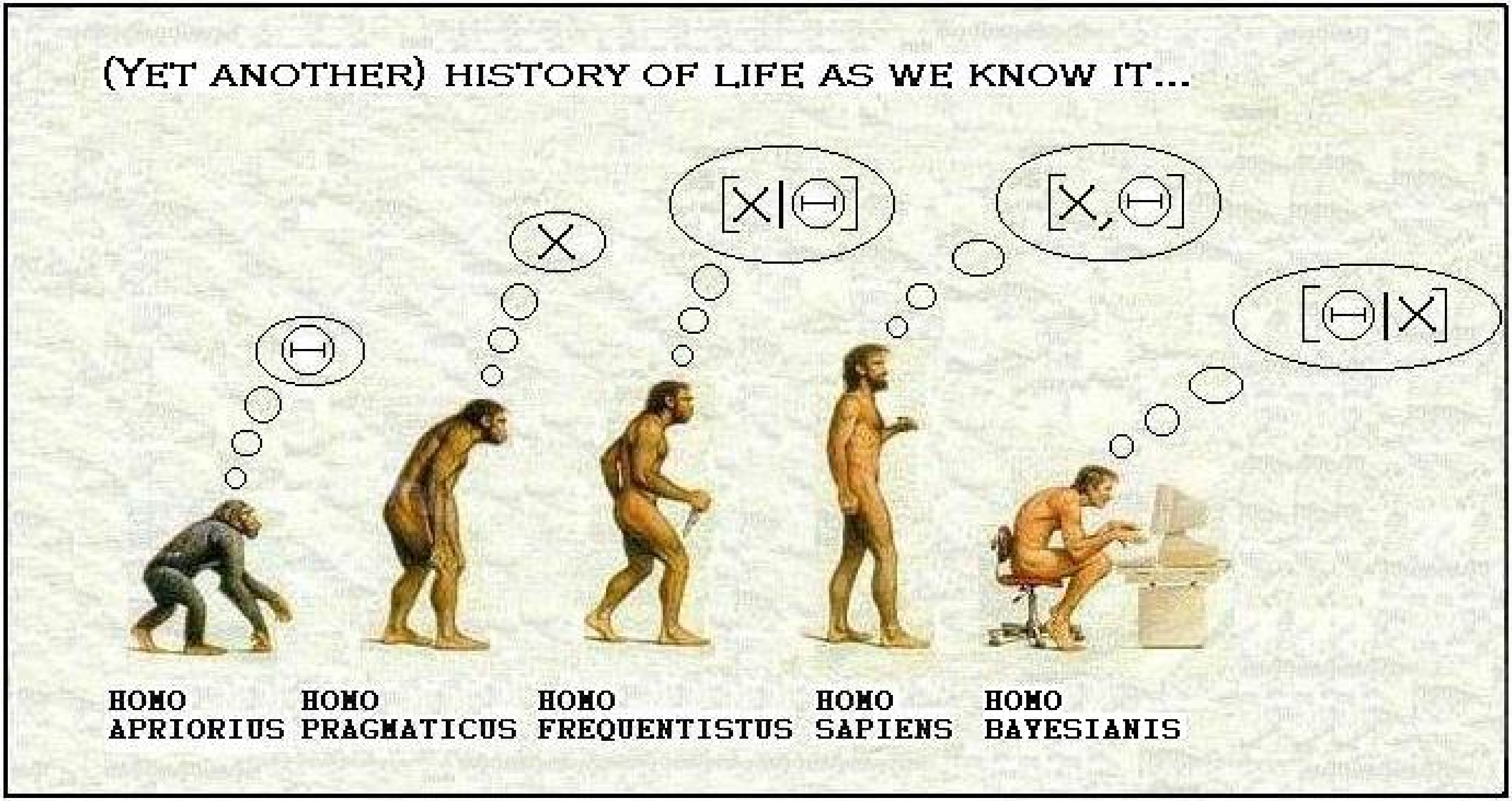
There are 3 types of lies:

1. Lies
2. Damned lies
3. and Statistics !

June 3rd 2004 **The Economist**

"… two researchers at the University of Girona in Spain, have found that 38% of a [random] *sample of papers in* ***Nature*** *contained one or more statistical errors …"*

Evolution of Inference



Bayesian *vs*. Frequentist

- Frequentist Statistics
  - *a.k.a.* "orthodox statistics" ("classical theory")
  - Probability as *frequency* of occurrences in ∞ # of trials
  - Historically arose from study of populations
  - Based on repeated trials and *future* datasets
  - *p*-values, *t*-tests, ANOVA, *etc*. (a cookbook of hacks!)

- This debate has been long & acrimonious

- 18th – 19th century was mostly (already) Bayesian, the 20th century was dominated by Frequentists, and now looks like the **21st is back to Bayesics!**

Bayesian *vs*. Frequentist

"In academia, the Bayesian revolution is on the verge of becoming the majority viewpoint, which would have been unthinkable 10 years ago."

from The New York Times, January 20th 2004

- Bradley P. Carlin,
  Mayo Professor of Public Health
  Head of Division of Biostatistics
  University of Minnesota

Bayesian *vs*. Frequentist

- Pathologies of Freq Statistics are finally being acknowledged

- Tests of *statistical significance* are now increasingly Bayesian

- Many journals discourage p-values
  - American J. of Public Health
  - Medical J. of Australia
  - The British Heart Journal
  - The Lancet
  - and even more generally by the Int'l Committee of Medical Journal Editors

The Earliest "Bayesian" ?

Herodotus
(c. 500 BC)

"A decision was **wise**, even though it led to disastrous consequences, **if the evidence at hand indicated that it was the best one to make**

And a decision was **foolish**, even though it led to the happiest possible consequences, **if it was unreasonable to expect those consequences**"

A Pre-Bayesian Minimalist

William of Occam
(1288 – 1348 AD)

Occam's Razor :

"Frustra fit per plura, quod fieri potest per pauciora."

"It is vain to do with more what can be done with less."

Everything else being equal, one should favour the simpler model

Bayesian model selection automatically implements a form of Occam's Razor (*i.e.* automatic complexity control)

The Founding Founders

Blaise Pascal

(1623-1662, France)

Pierre Fermat

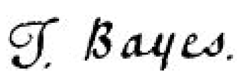
(1601-1665, France)

They laid the foundations of Probability Theory
in a correspondence about a *game of dice*.

The Reverend Bayes

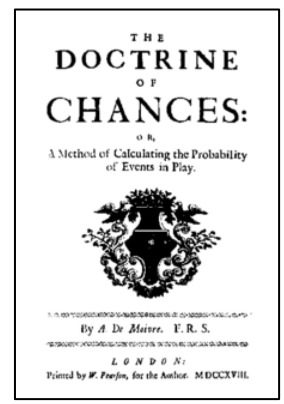
Thomas Bayes

(1702 - 1761, UK)



His manuscript "***An essay towards solving a problem in The Doctrine of Chances***" was found by a friend after his death and (given due special consideration) was published in the *Philosophical Transactions of the Royal Society of London* in 1764.

** The Doctrine of Chances: A Method for Calculating Probability of Events in Play*
is a book by *Abraham de Moivre* (1718)



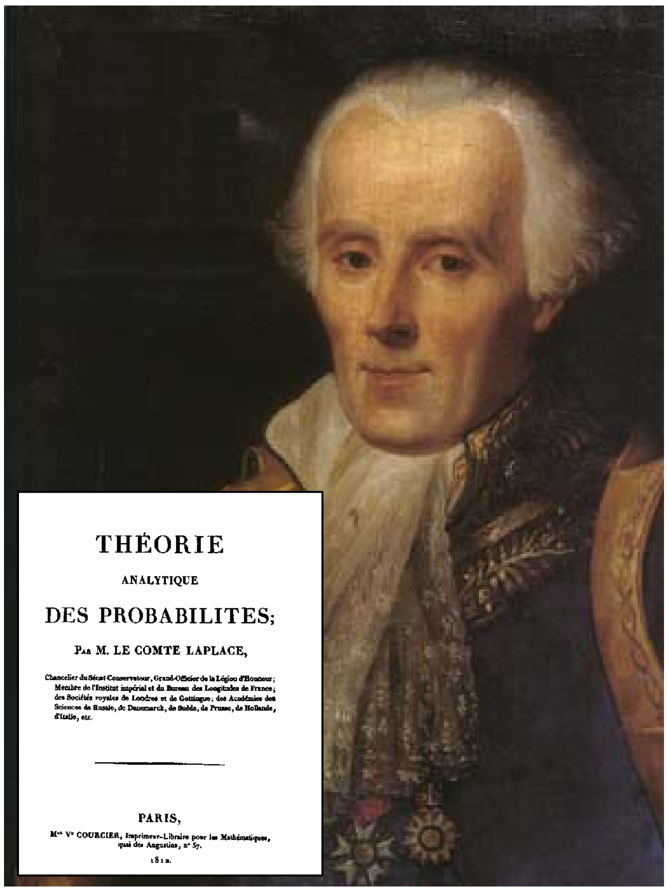
- Bayes was first to tackle ***Inverse Probability*** :

  going from effects (observations)
  to their causes (models/parameters)

The Prince of Probability

Pierre-Simon Laplace

(1749 – 1827)



"Probability theory is nothing but common sense reduced to calculation"

- Mathematical Physicist & Astronomer
- A shrewd self-promoter (but truly gifted)
- Independently discovered Bayes' rule (but he later acknowledged Bayes' *role*)
- Laplace argued in favor of *uniform* priors
- Solved many applied inverse-probability problems in physics and astronomy
- The term **Bayesian** may very well be replaced by **Laplacian**, in Statistics

The Father of Orthodoxy

Ronald Fisher
(1890 – 1962)

Cambridge Geneticist & Biologist
(also a key proponent of *eugenics* in the 1930s)

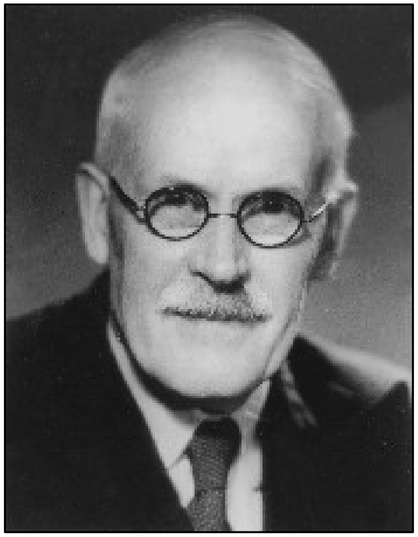
- Fisher misunderstood Laplace's work
- He found Bayesian integrals/math too hard
- Re-invented statistical inference as being solely *likelihood-based* (and called it "fiducial")
- By most accounts Fisher was a harsh, rigid, egotistical and vindictive man [Jaynes 2003]

"So long as you avoided a handful of subjects like inverse probability that would turn Fisher in the briefest possible moment from extreme urbanity into a boiling cauldron of wrath, you got by …"

– Fred Hoyle, Cambridge Astronomer

The Gentle Revivalist

Harold Jeffreys
(1891 – 1989)



- Mathematician, Statistician, Astronomer
- A contemporary of Fisher, who had more than a few disagreements with Fisher
- Revived Bayes-Laplace style of inference
- Derived *invariant* uninformative priors
- Pointed out some fallacies of Frequentists

" What the use of the p-value **[significance level]** *implies, therefore, is that a hypothesis that may be true may be rejected because it has not predicted observable results that have not occurred. "*

— Harold Jeffreys, Theory of Probability (1939)

The Hardcore Crusader

Edwin Jaynes

(1922 – 1998)

- Physicist, Statistician
- Modern proselytizer of Bayes-Laplace view
- Probability Theory as Extended Logic
- Statistical Mechanics & Information Theory
- Devised "Maximum-Entropy" (MaxEnt) priors
- Pointed out endless flaws of Orthodox Statistics

"This may seem like an inflexible, cavalier attitude;* [however] *I am convinced that nothing short of it can ever remove the ambiguity of* [the problem] *that has plagued probability theory for two centuries"

— Ed Jaynes, Probability Theory: The Logic of Science (2003)

A Frequentist's *Mea Culpa*

Jerzy Neyman
(1894 – 1981)

- Founder of *Hypothesis Testing*
- Co-Inventor of *Confidence Intervals*
- Inventor of *Random Sampling*
- Emphasis on repeated randomized trials
- *Neyman-Pearson Lemma* (with his advisor)

***" The trouble is that what we* [statisticians] *call modern* [orthodox] *statistics was developed under strong pressure on the part of biologists*. *As a result, there is practically nothing done by us which is directly applicable to problems of astronomy."* -- Jerzy Neyman (years later)**

Bayesian *vs*. Frequentist

So leave these assumptions behind:

- “A probability is a frequency”

- “Probability theory only applies to large populations”

- “Probability theory is arcane and boring”

Fundamentals

What is Reasoning?

- How do we infer properties of the world?
  - we want *inductive* reasoning
  - we must account for *all* uncertainty
    - due to our own ignorance (about the world)
    - inherent "noise/chance" (intrinsic to the world)
- How should computers do it?

Aristotelian (Deductive) Logic

- If A is true, then B is true
- A is true
- Therefore, B is true

A: patient has AIDS
B: patient is HIV +

Note: *if-then* is *not* always causation

Real-World is Uncertain

Problems with pure (Boolean) logic:

- Don't have perfect information
- Don't really know the model
- Pure logic is deterministic
  - No way to have "chance" outcomes
  - No way to capture noise, uncertainty, *etc*

So let's build a logic of uncertainty!

Beliefs

Let bel(A) = "belief that A is **true**"

bel(¬A) = "belief that A is **false**"

e.g., A = "Mars has microbial life"

bel(A) = "belief in Martian microbial life"

Reasoning with Beliefs

Cox Axioms [Cox 1946]

1. Ordering exists
   - *e.g.*, bel(A) > bel(B) > bel(C)
2. Negation function exists
   - $\text{bel}(\neg A) = \mathbf{f}(\text{bel}(A))$ for some function $\mathbf{f}$
3. Product function exists
   - $\text{bel}(A \wedge Y) = \mathbf{g}(\text{bel}(A|Y), \text{bel}(Y))$

     for some function $\mathbf{g}$

This is all we need!

Reasoning with Beliefs

The 3 Cox Axioms ***uniquely*** define a ***complete*** system of reasoning

which is … **Probability Theory** **!**

* Any other framework will therefore have to be *incomplete*, *incoherent* and/or *sub-optimal* and can lead to *paradoxes*

Principle #1:

"Probability theory is nothing more than common sense reduced to calculation."

-- Laplace (1814)

Definitions

P(A) = "probability A is true"

= bel(A) = "belief A is true"

P(A) is a real value in [0,1]

P(A) = 1 iff "A is true"

P(A) = 0 iff "A is false"

P(A|B) = "probability of A if we knew B"

P(A, B) = "probability of A and B"

Examples

A: "patient has a concussion"

B: "patient has a headache"

$$P(A) = 0.11$$

$$P(B) = 0.53$$

$$P(B \mid A) = 0.92$$

$$P(A \mid B) = 0.05$$

Basic Rules

Sum rule:

$$P(A) + P(\neg A) = 1$$

Example:

A: "**spacecraft will survive EDL**"

$P(A) = 0.9$ thus $P(\neg A) = 0.1$

Basic Rules

Sum rule:

$$\sum_i P(A_i) = 1$$

when exactly one of the A_i must be true

Basic Rules

Product rule:

$$P(A,B) = P(A|B)\,P(B)$$

$$= P(B|A)\,P(A)$$

Basic Rules

Conditioning

Product Rule

$$P(A, B) = P(A|B)\, P(B)$$

$$\rightarrow \quad P(A, B|C) = P(A|B,C)\, P(B|C)$$

Sum Rule

$$\sum_i P(A_i) = 1 \quad \rightarrow \quad \sum_i P(A_i|B) = 1$$

Basic Rules

Product rule	$P(A,B) = P(A\|B)\,P(B)$
Sum rule	$\sum_i P(A_i) = 1$

All derivable from Cox axioms;
obey rules of common sense

From these we can derive new rules

Example

A = "patient loses weight over the next 2 weeks"

B = "patient watches diet and does exercise"

¬B = "patient takes some OTC weight-loss pill"

Model: $P(B) = 0.7$

$P(A|B) = 0.8$

$P(A|\neg B) = 0.5$

what is $P(A)$?

Example, continued

Model: $P(B) = 0.7$, $P(A|B) = 0.8$, $P(A|\neg B) = 0.5$

$1 = P(B) + P(\neg B)$ **Sum rule**

$1 = P(B|A) + P(\neg B|A)$ **Conditioning**

$P(A) = P(B|A)P(A) + P(\neg B|A)P(A)$

$= P(A,B) + P(A,\neg B)$ **Product rule**

$= P(A|B)P(B) + P(A|\neg B)P(\neg B)$ **Product rule**

$= 0.8 \times 0.7 + 0.5 \times (1 - 0.7) =$ **0.71**

Basic Rules

Marginalizing

$$P(A) = \sum_i P(A, B_i)$$

for mutually-exclusive B_i

for example,

$$P(A) = P(A,B) + P(A, \neg B)$$

Syllogisms Revisited

A -> B

A

Therefore B

$P(B|A) = 1$

$P(A) = 1$

$$
\begin{aligned}
P(B) &= P(B,A) + P(B, \neg A) \\
&= P(B|A)P(A) + P(B|\neg A)P(\neg A) \\
&= 1
\end{aligned}
$$

More than 2 Variables

1. **Knowing $P(A,B,C)$ is equivalent to:**
   - $P(A,B|C)\ P(C)$
   - $P(A|C)\ P(B|A,C)$
   - $P(B|C)\ P(A|B,C)$

(Cox's Theorem)

Principle #2:

Given a complete model, we can derive any other probability

The joint probability of all the unknowns is the "full recipe" or description of our model.

All inferential goals derive from that joint probability, using the Sum/Product rules

Inference

Model: $P(B) = 0.7,\ P(A|B) = 0.8,\ P(A|\neg B) = 0.5$

Given observation A (patient lost some weight)

what is $P(B|A)$? (patient did diet/exercise)

$P(A,B) = P(A|B)\,P(B) = P(B|A)\,P(A)$ Product Rule

$$\mathbf{P(B|A)} = \frac{\mathbf{P(A|B)\,P(B)}}{\mathbf{P(A)}} = \frac{(0.8)(0.7)}{(\mathbf{0.71})} = \mathbf{0.79}$$

this is the *controversial* Bayes' Rule !

Inference

<u>Bayes Rule</u>

Likelihood

Prior

$$P(\theta|D) = \frac{P(D|\theta)\,P(\theta)}{P(D)}$$

Posterior

Frequentists *accept* this formula (it's irrefutable!)

But they object to *using* priors (as being subjective)

Principle #3:

Setup your model of the world and then compute probabilities of the unknowns given the observations

$P($ parameters $|$ data$)$ ***estimation***

$P($ *new* data $|$ data$)$ ***prediction***

$P($ model $|$ data$)$ ***model selection***

$P(H_0 \mid$ data , model$)$ ***hypothesis tests***

One unified framework for multiple tasks!

Principle #3a:

Use Bayes' Rule to infer the unknown **X** from the observed **O**

Likelihood

Prior

$$\mathbf{P(X|O)} = \frac{\mathbf{P(O|X)\ P(X)}}{\mathbf{P(O)}}$$

Posterior

Independence

Definition:

A and B are independent iff

$$P(A,B) = P(A)\,P(B)$$

Example: Diagnosis

Jo takes a blood test for a certain disease

Test result is either "positive" (T) or "negative" (¬T)

The test is 95% reliable

1% of people in Jo's demographic have the disease

If the test result is "positive" (T)

does Jo have the disease? [MacKay 2003]

Example: Diagnosis

Model: $P(D) = 0.01$ $\qquad P(T|D) = 0.95$

$$P(\neg T|\neg D) = 0.95$$

$$P(D|T) = \frac{P(T|D)\,P(D)}{P(T)} \approx 0.16 \text{ or } \mathbf{16\ \%}$$

since $P(T) = P(T|D)\,P(D) + P(T|\neg D)\,P(\neg D)$

$$= 0.95 \times 0.01 + (1 - 0.95) \times 0.99 = \mathbf{0.059}$$

Example: Diagnosis

What if we tried different tests ?

99.9% reliable test gives $P(D|T_2) \approx$ **91 %**

70% reliable test gives $P(D|T_3) \approx$ **2 %**

The posterior combines all available information – so could use *multiple* tests, e.g., $P(D|T_2, T_3)$

Discrete Variables

Probabilities over discrete variables

$C \in \{ \textit{Heads}, \textit{Tails} \}$

$P(C = \textit{Heads}) = 0.5$ (but *why ?*)

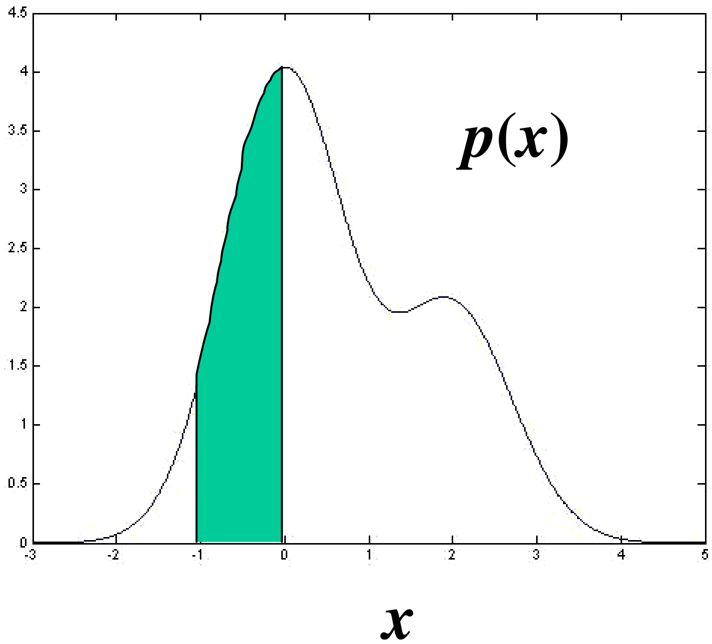
Sum Rule:

$P(C = \textit{Heads}) + P(C = \textit{Tails}) = 1$

Continuous Variables

Probability Density Function (PDF)

measures *concentration* of probability "mass"



$$P(a \le x \le b) = \int_a^b p(x)\,dx$$

Notation:

$P(x)$ is probability *distribution* function (cumulative) whereas $p(x)$ is local probability density so $Prob(x = 2)$ is zero !

Continuous Variables

Probability Density Function (PDF)

Let $x \in R$

$p(x)$ is any *non-negative* function s.t.

$$\int p(x)dx = 1$$

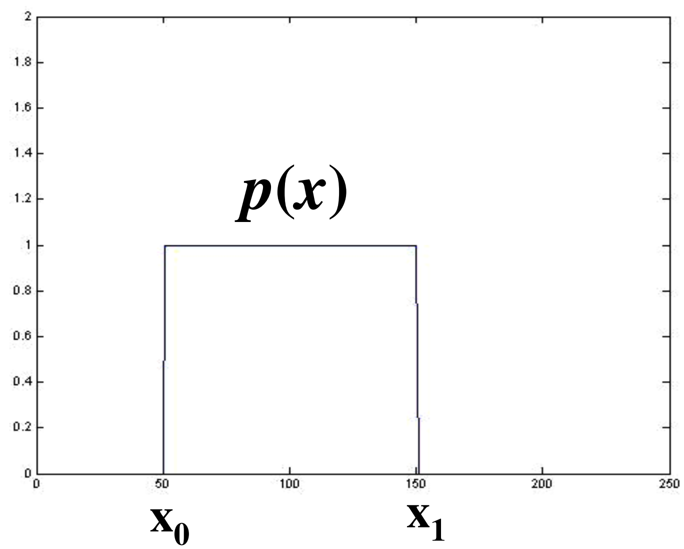
$$P(a \leq x \leq b) = \int_a^b p(x)dx$$

Uniform Distribution

$$x \sim U(x_0, x_1)$$

$$p(x) = 1/(x_1 - x_0) \quad if \quad x_0 \le x \le x_1$$

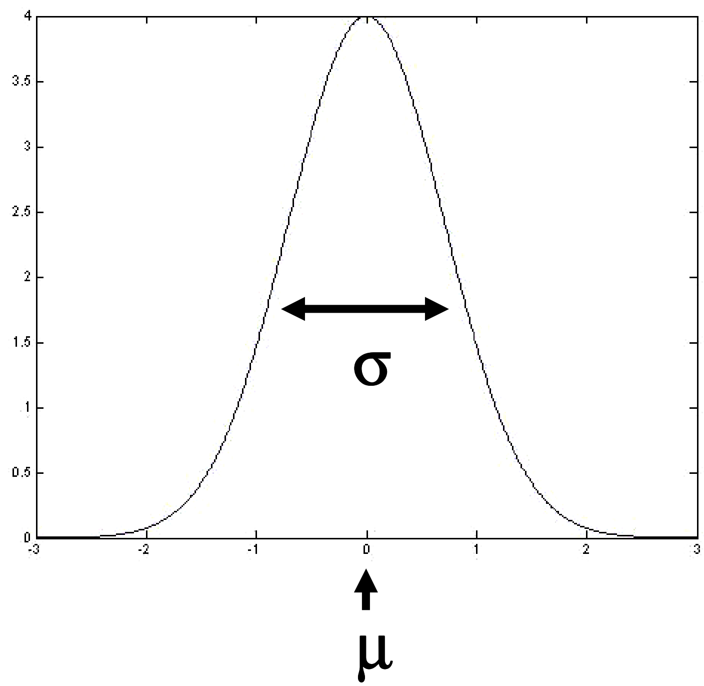
$$= 0 \quad otherwise$$



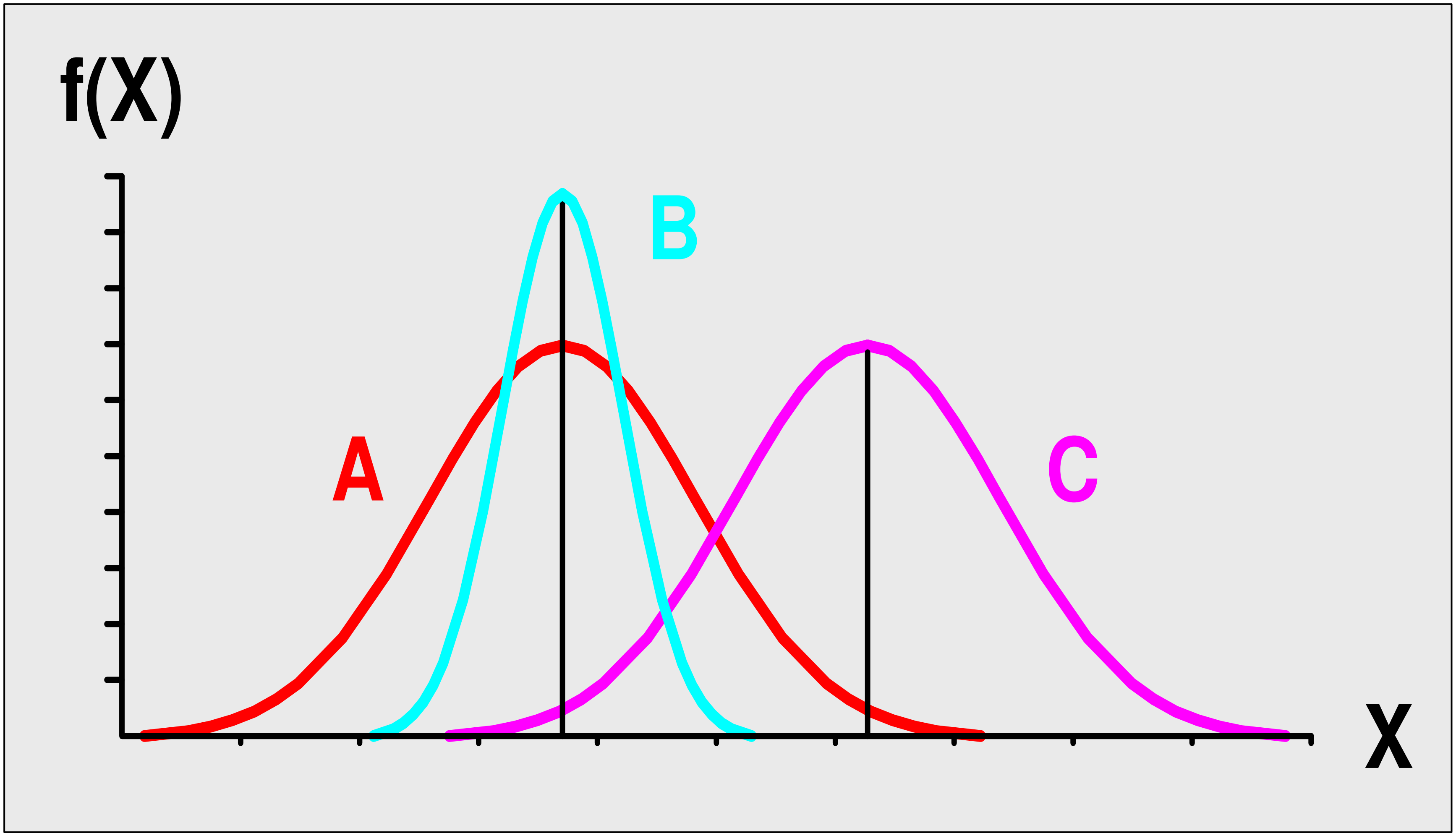
Gaussian Distributions

$$x \sim \mathrm{N}(\mu, \sigma^2)$$

$$p(x|\mu,\sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$



Gaussian Parameters (μ , σ)

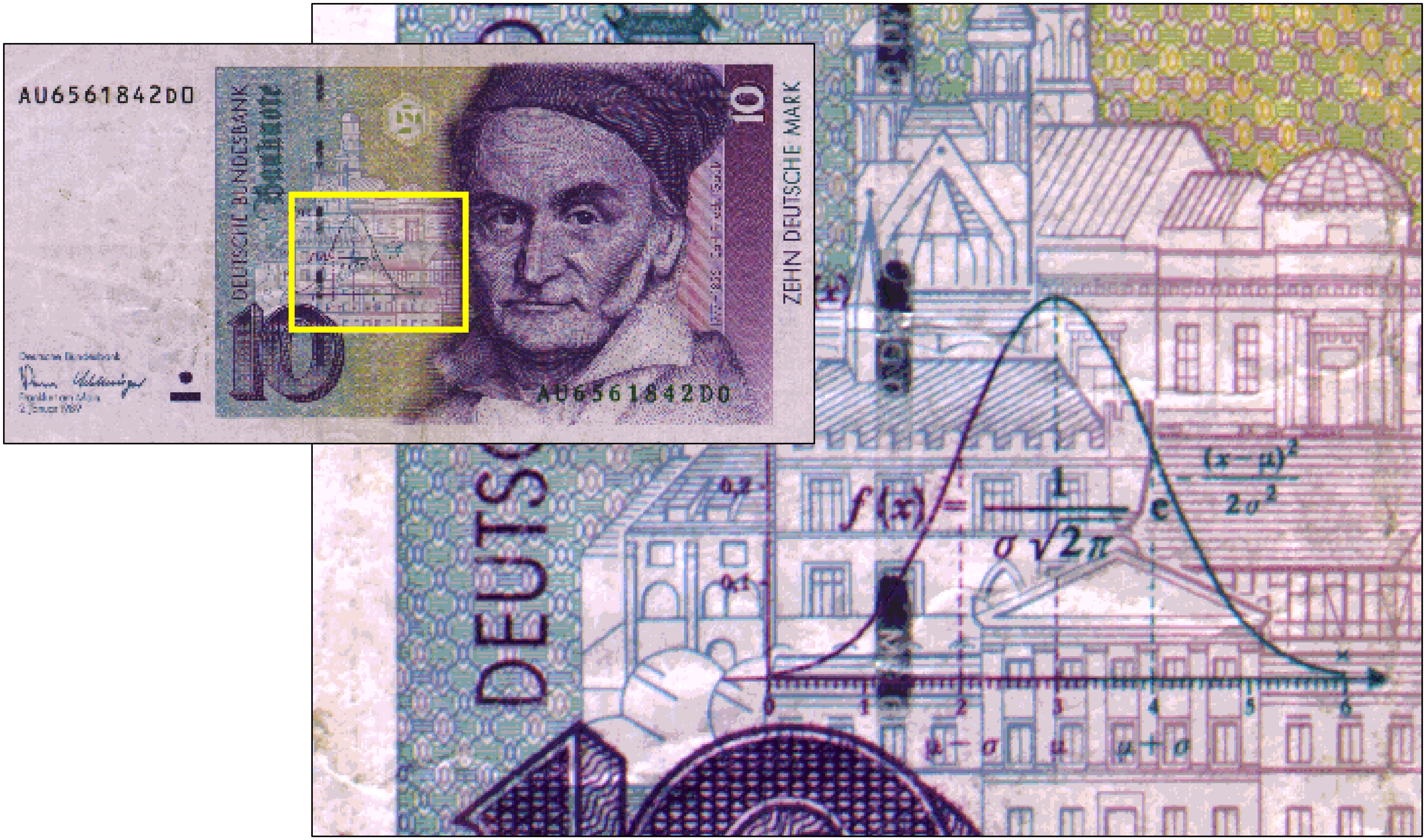


Why use Gaussians?

- Convenient analytic properties
- Central Limit Theorem
- Infinite Divisibilty
- Works well in practice
- Not for everything, but good approx
- For more theoretical reasons, see

  [Bishop 1995, Jaynes 2003]

Why use Gaussians?



Rules for Continuous PDFs

Same intuitions and rules apply

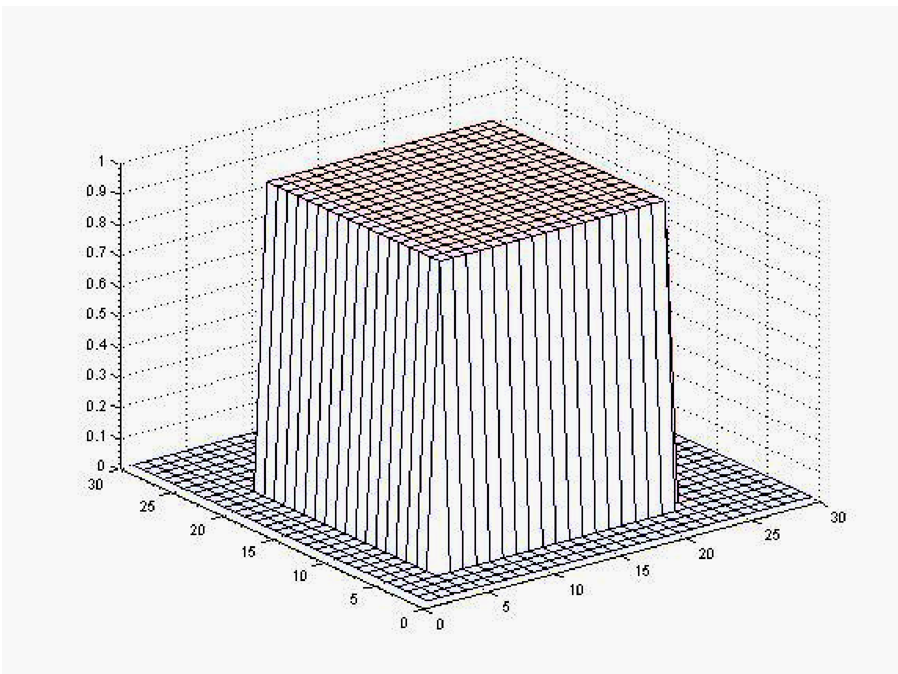
Sum rule: $\int_{-\infty}^{+\infty} p(x)dx = 1$

Product rule: $p(x,y) = p(x \mid y)\, p(y)$

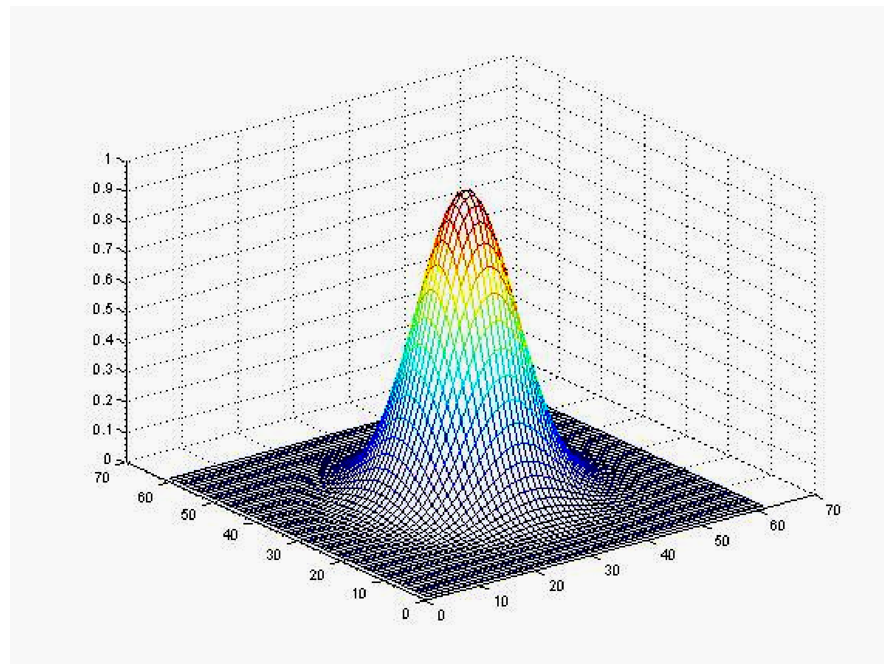
Marginalizing: $p(x) = \int p(x,y)\, dy$

… and Bayes' Rule, conditioning, *etc*.

Multivariate distributions



Uniform: $x \sim \mathrm{U}(domain)$



Normal: $x \sim \mathrm{N}(\mu, \Sigma)$

Inference

How to reason about the world from observations?

Three important sets of variables:

1. Observations (known, given, ”clamped”)
2. Unknowns (parameters, missing data, submodels)
3. Auxiliary (“nuisance”) variables
   – Any left over variables we don’t care about but *must* account for

Given the observed (known) data,
what are the probabilities of the unknowns?

Inference

Coin-flipping : Bernoulli trials

$P(C = Heads|\theta) = \theta$

$p(\theta) = \text{Uniform}(0,1)$ (Bayes & Laplace)

Suppose we flip the coin 1000 times
and get 750 heads. What is θ?

Intuitive answer : $750/1000 = 75\%$

What is θ?

$$p(\theta) = \text{Uniform}(0,1)$$

$$P(C_i = h | \theta) = \theta, \quad P(C_i = t | \theta) = 1 - \theta$$

$$P(C_{1:1000} | \theta) = \prod_{\text{i}} P(C_i = h | \theta)$$

$$p(\theta | C_{1:1000}) = \frac{P(C_{1:1000} | \theta)\, p(\theta)}{P(C_{1:1000})}$$

Bayes' Rule

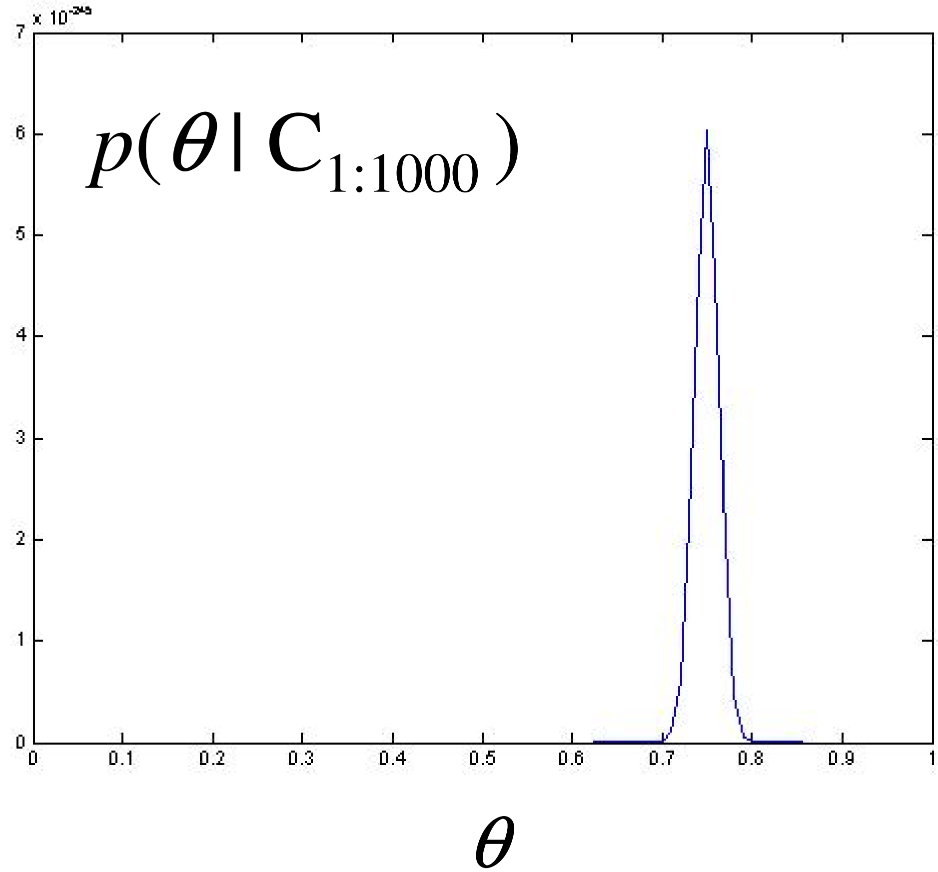
$$= \prod_{\text{i}} P(C_i \mid \theta)\, p(\theta) \;/\; P(C_{1:1000})$$

$$= \theta^H (1 - \theta)^T / P(C_{1:1000})$$

$$\propto \theta^H (1 - \theta)^T$$

H = 750
T = 250

What is θ?



The posterior *distribution* tells us everything (our revised belief about θ after seeing data)

Bayesian Prediction

What is the probability of another head?

$$P(C_{N+1} = h | C_{1:N}) = \int P(C = h,\ \theta \mid C_{1:N})\, d\theta$$

$$= \int P(C = h | \theta)\ \ p(\theta \mid C_{1:N})\, d\theta$$

$$= \int\ \theta\ p(\theta \mid C_{1:N})\, d\theta$$

$$= (H + 1)/(N + 2)$$

*** Note: we never computed an *estimate* of θ**

Parameter Estimation

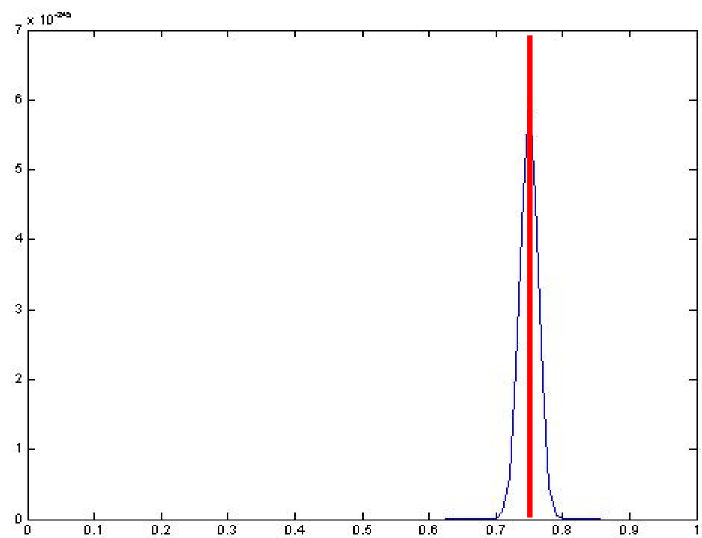
- **What if we *want* an estimate of θ?**
- **Maximum A Posteriori (MAP):**

$$\theta^* = \ arg\ max_{\theta} \ \ p(\theta \mid C_1, \ldots, C_N)$$

$$= \ H / N$$

$$= \ 750 / 1000 = \ 75\%$$

Note: with a flat prior on θ MAP and ML *mode* estimates are the same in this problem



A Problem ?

Suppose we had flipped coin just *once*

What is $P(C_2 = h \mid C_1 = h)$ **?**

ML estimate: $\theta^* = \mathrm{H} / \mathrm{N} = 1$

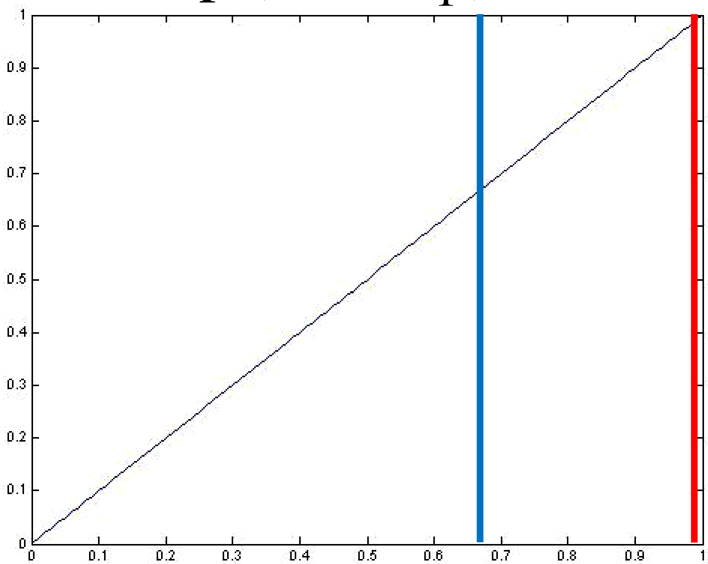
But that's absurd!

Bayesian prediction:

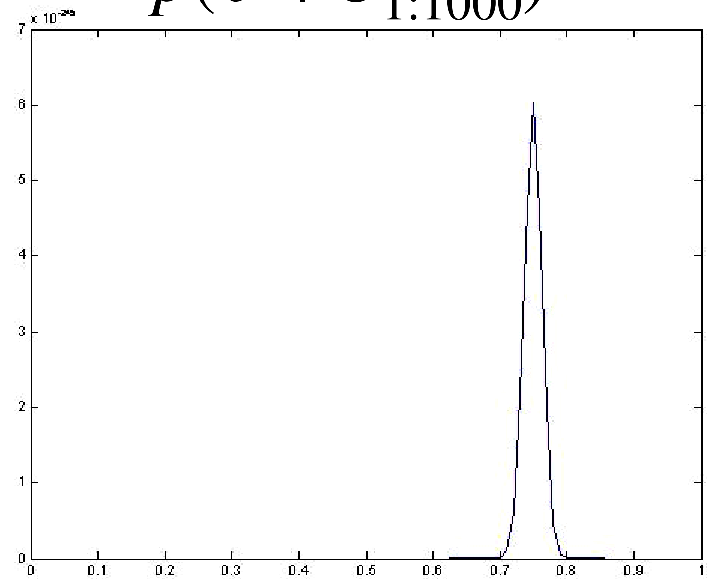
$$P(C_2 = h \mid C_1 = h) = (\mathrm{H} + 1) / (\mathrm{N} + 2) = \mathbf{2/3}$$

So what went wrong?

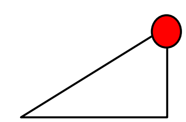
$p(\theta \mid C_1)$



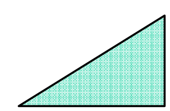
$p(\theta \mid C_{1:1000})$



ML/MAP estimate finds the posterior *peak*



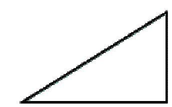
Bayes *integrates* over the posterior mass



Over-Fitting

- A model that fits the (current) data well but does not generalize (future)

- Occurs when a point-estimate is obtained from “spread-out” posterior



- Important to ask the right question: should we estimate C_{N+1} or θ ?

Principle #4:

Parameter estimation is not Bayesian.

It leads to errors, such as over-fitting.

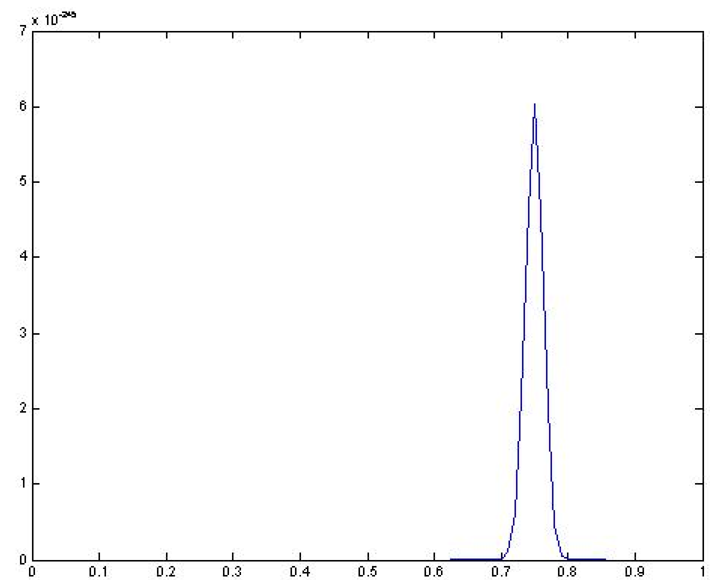
Bayesian prediction

$$p(x|D) = \int p(x,\ \theta \,|\, D)\, d\theta$$
$$= \int p(x|\theta)\, p(\theta \,|\, D)\, d\theta$$

data vector

training data

model parameters



Note "model averaging"

Advantages of Estimation

Bayesian prediction is usually difficult and/or expensive

$$p(x|D) = \int p(x, \theta \mid D)\, d\theta$$

Frequentists use "plug-in" estimates

(this ignores the uncertainty in estimates)

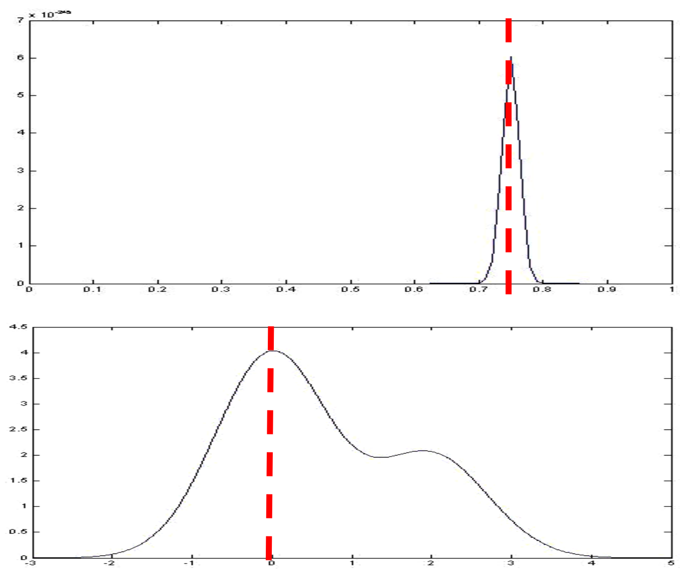
Q: When is Estimation Safe ?

A: When the posterior is “peaked”

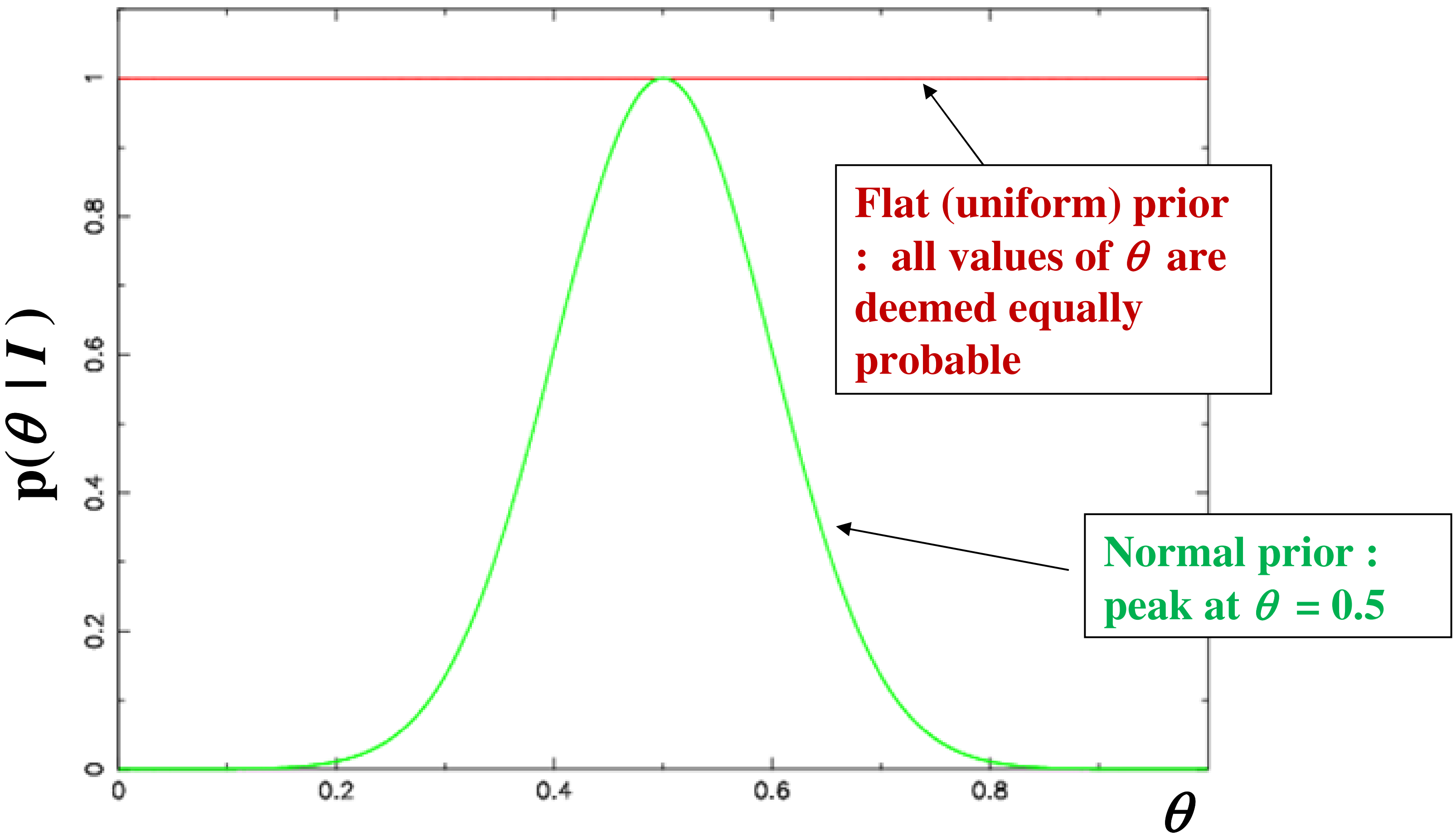
- The posterior “looks like” a spike
- Since often we have more data than parameters
- But this is not a guarantee
  (*e.g.*, fitting a line to 100 identical data points)
- In practice, use error bars (posterior variance)

Principle #4a:

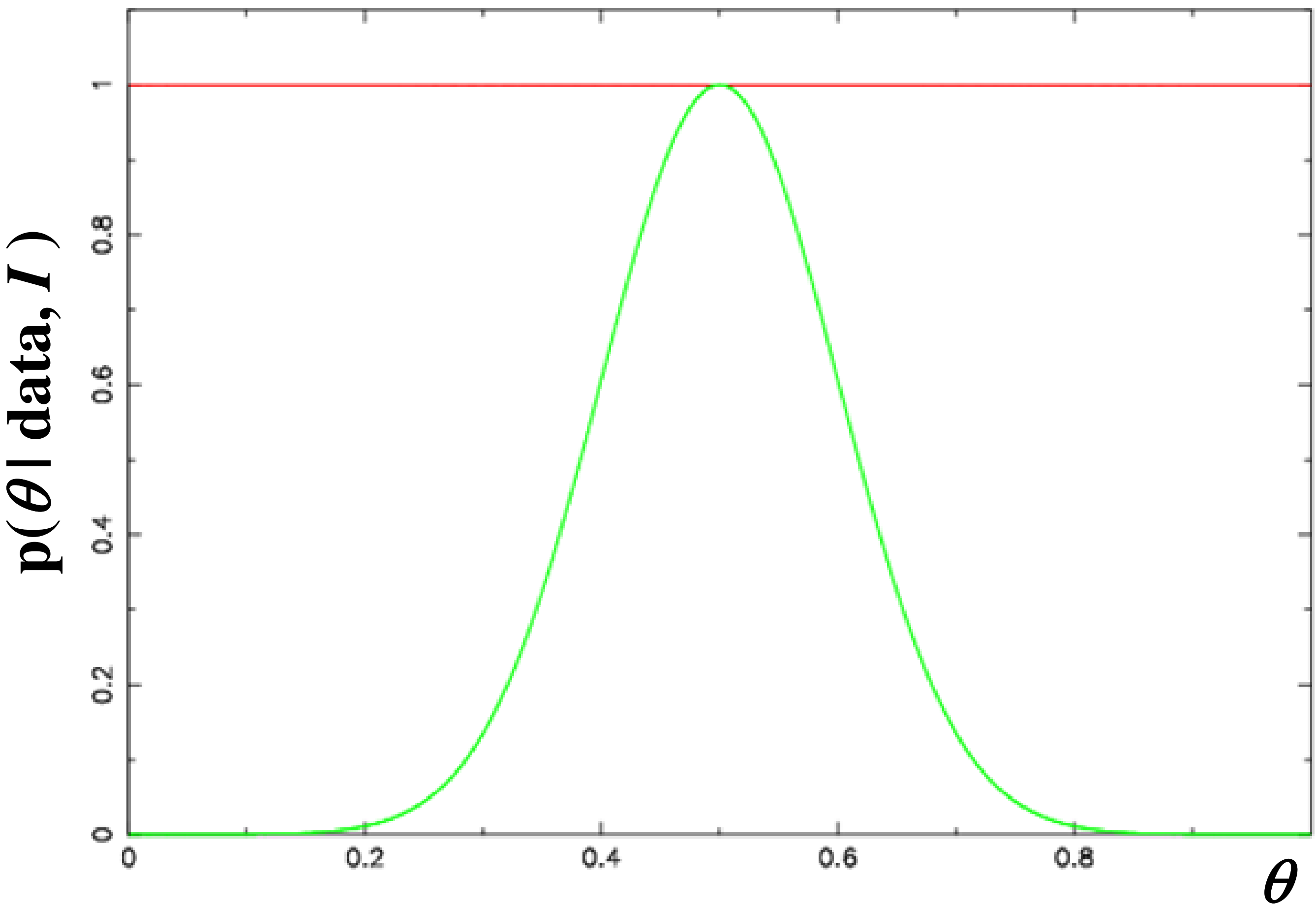
Parameter estimation is easier than prediction. It works well when the posterior is "peaked."



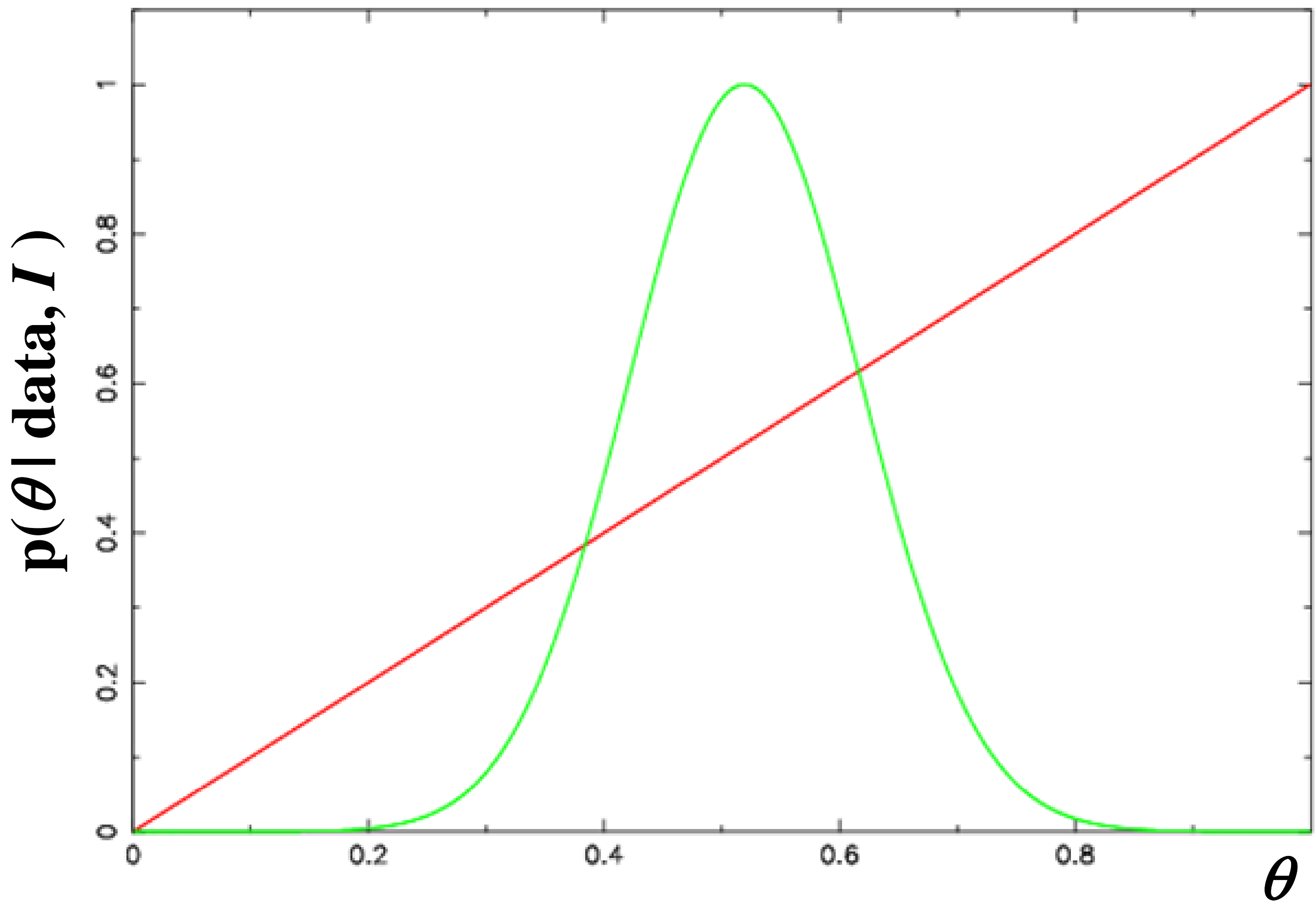
Different Priors $p(\theta)$



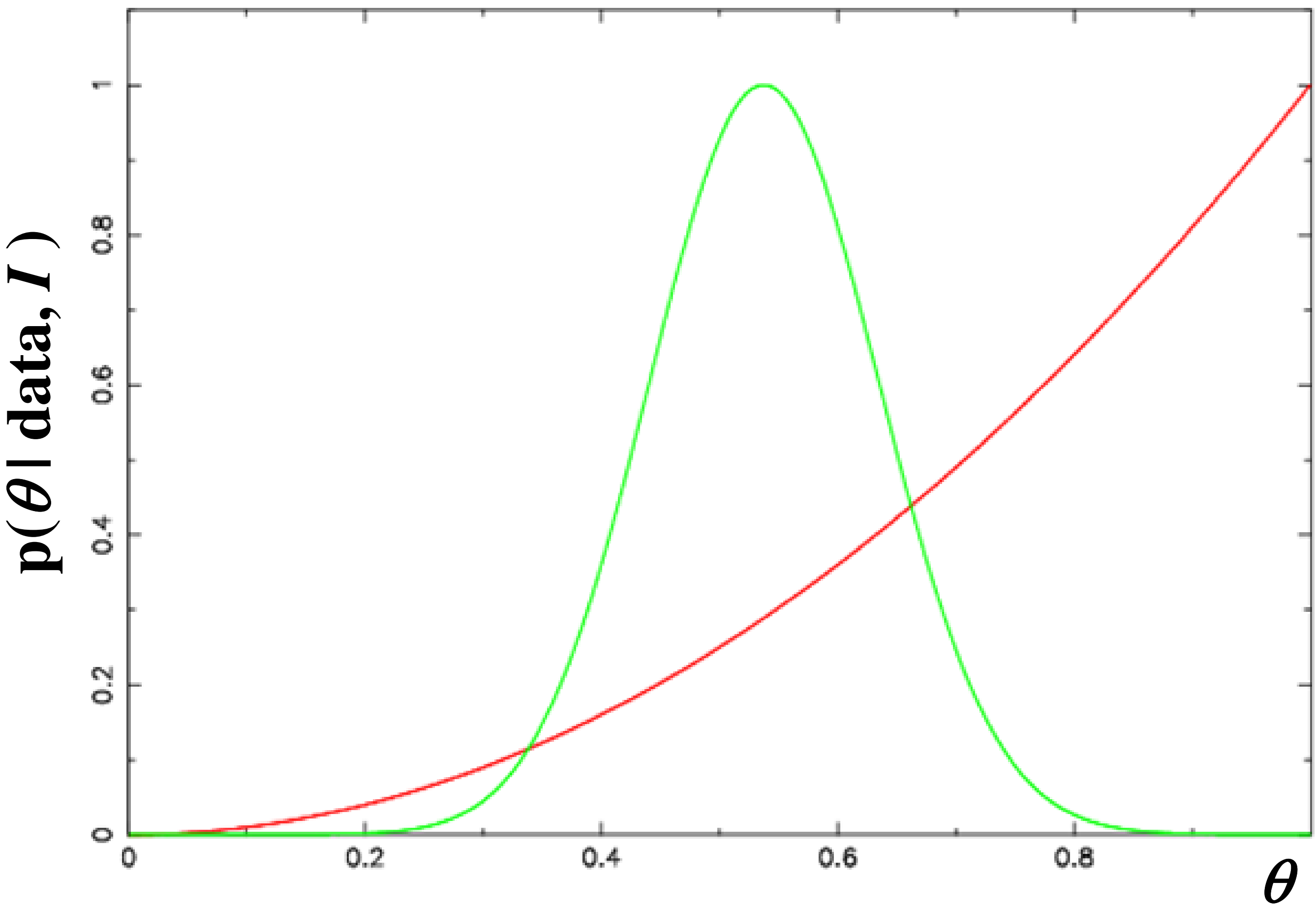
After N = 0 flips



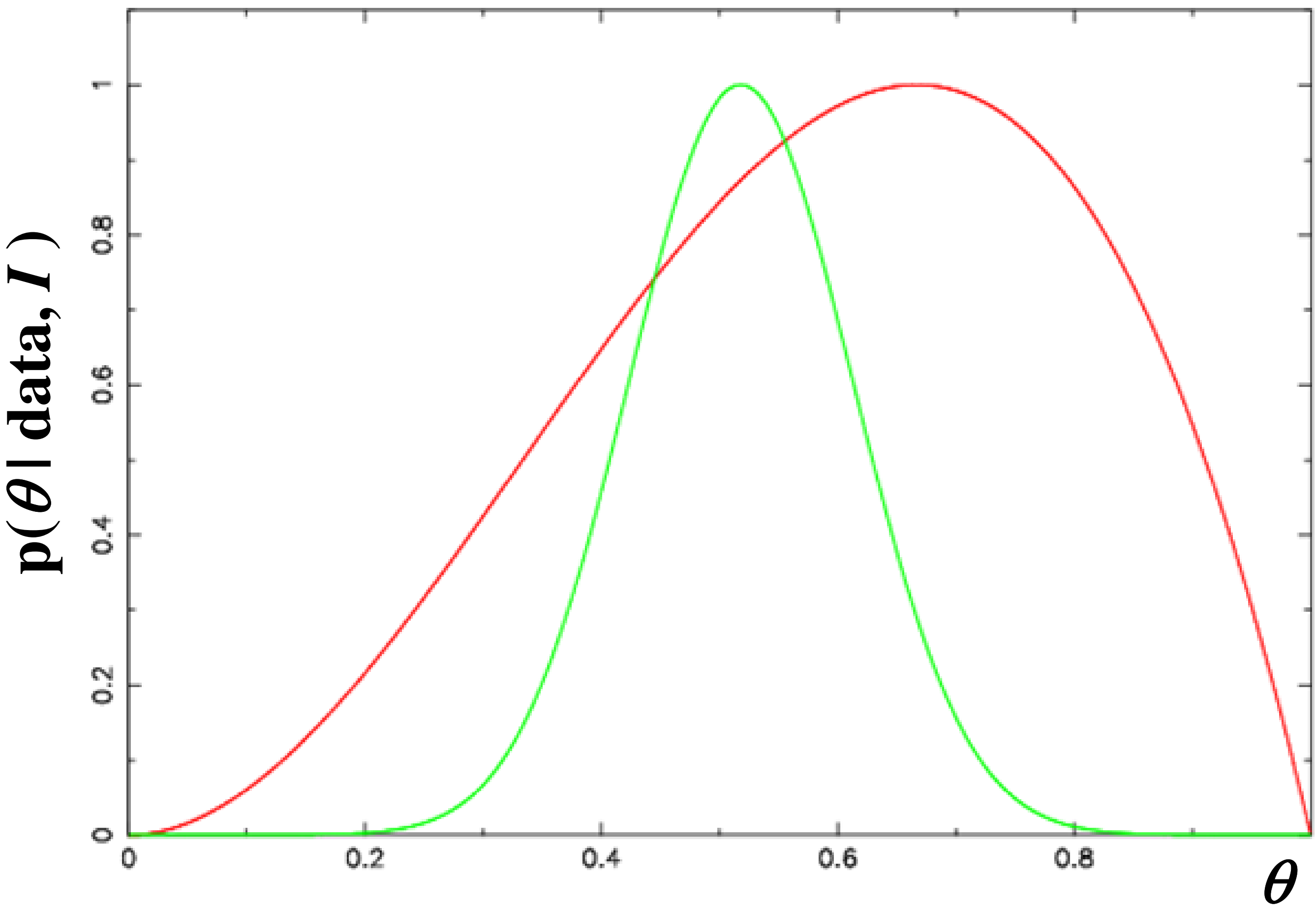
After N = 1 flips : H



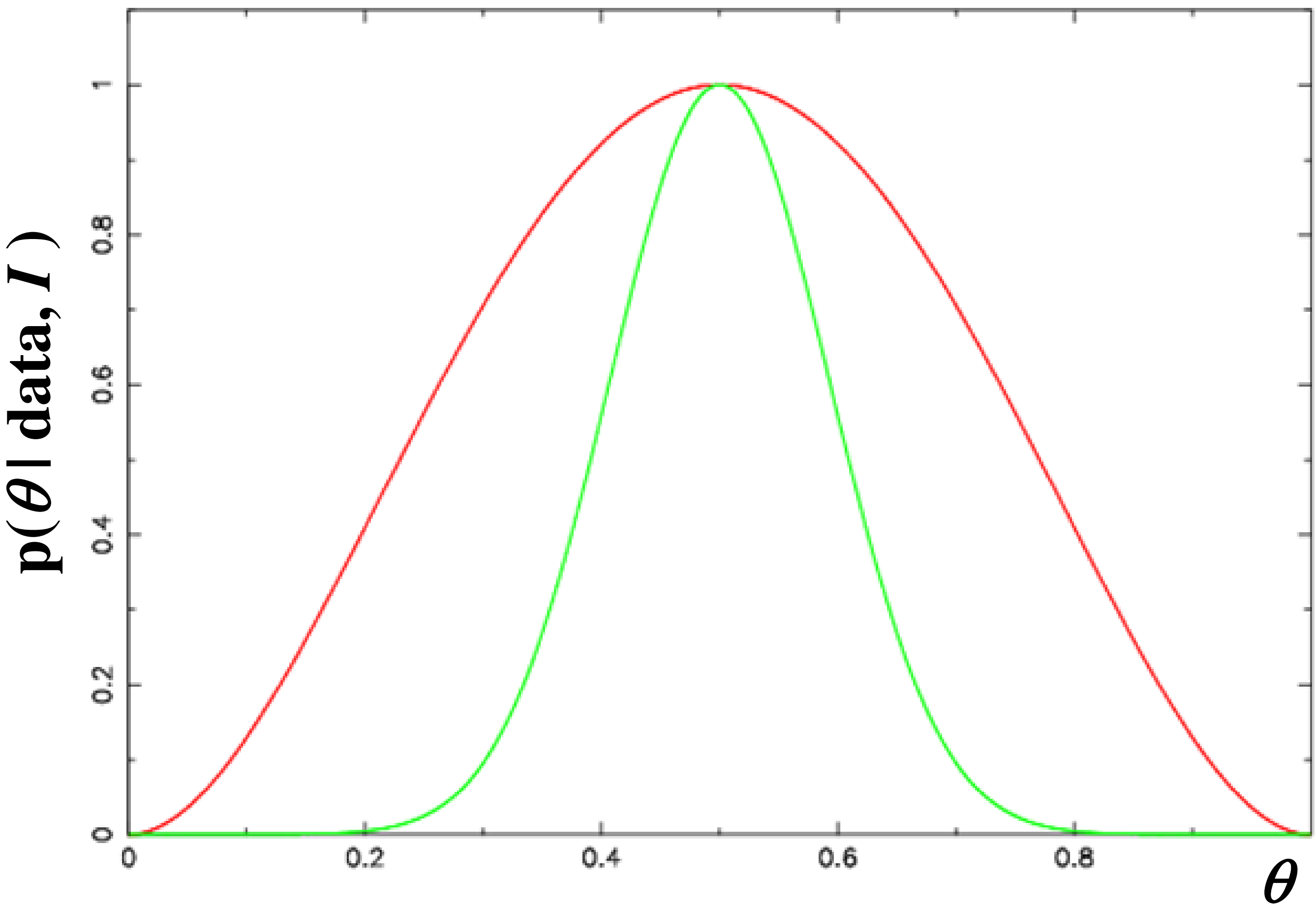
After N = 2 flips : H + H



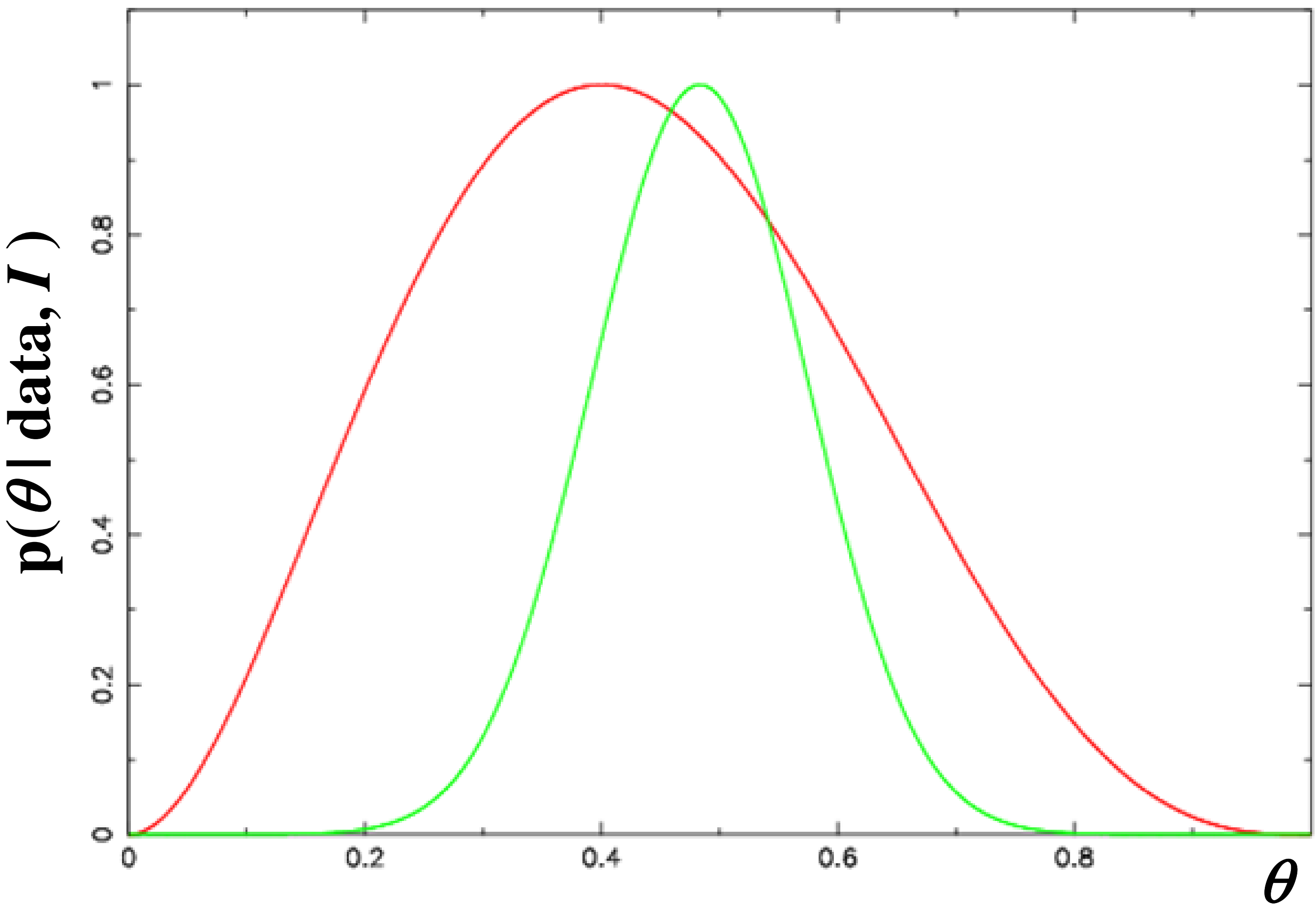
After N = 3 flips : H + H + T



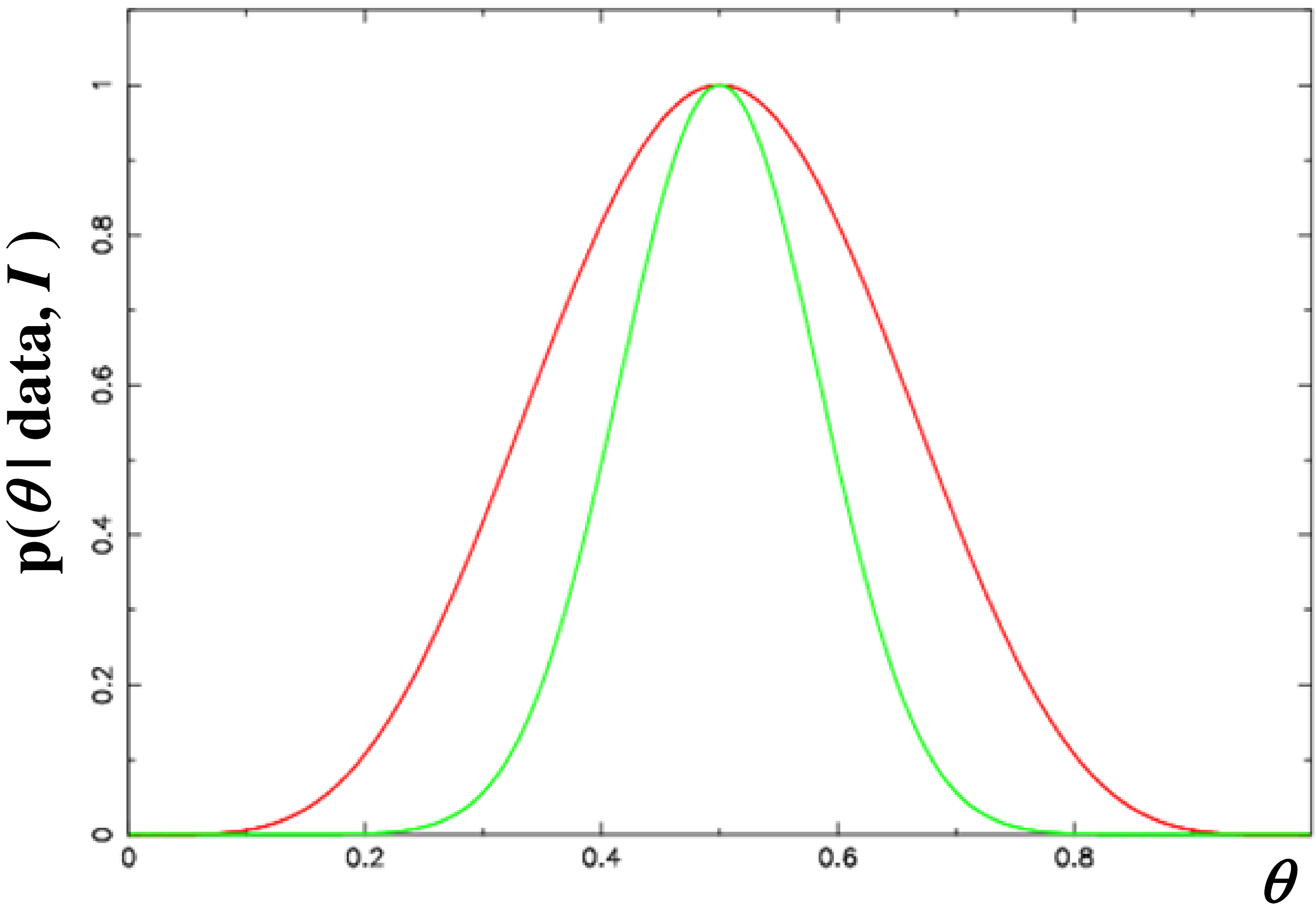
After N = 4 flips : H + H + T + T



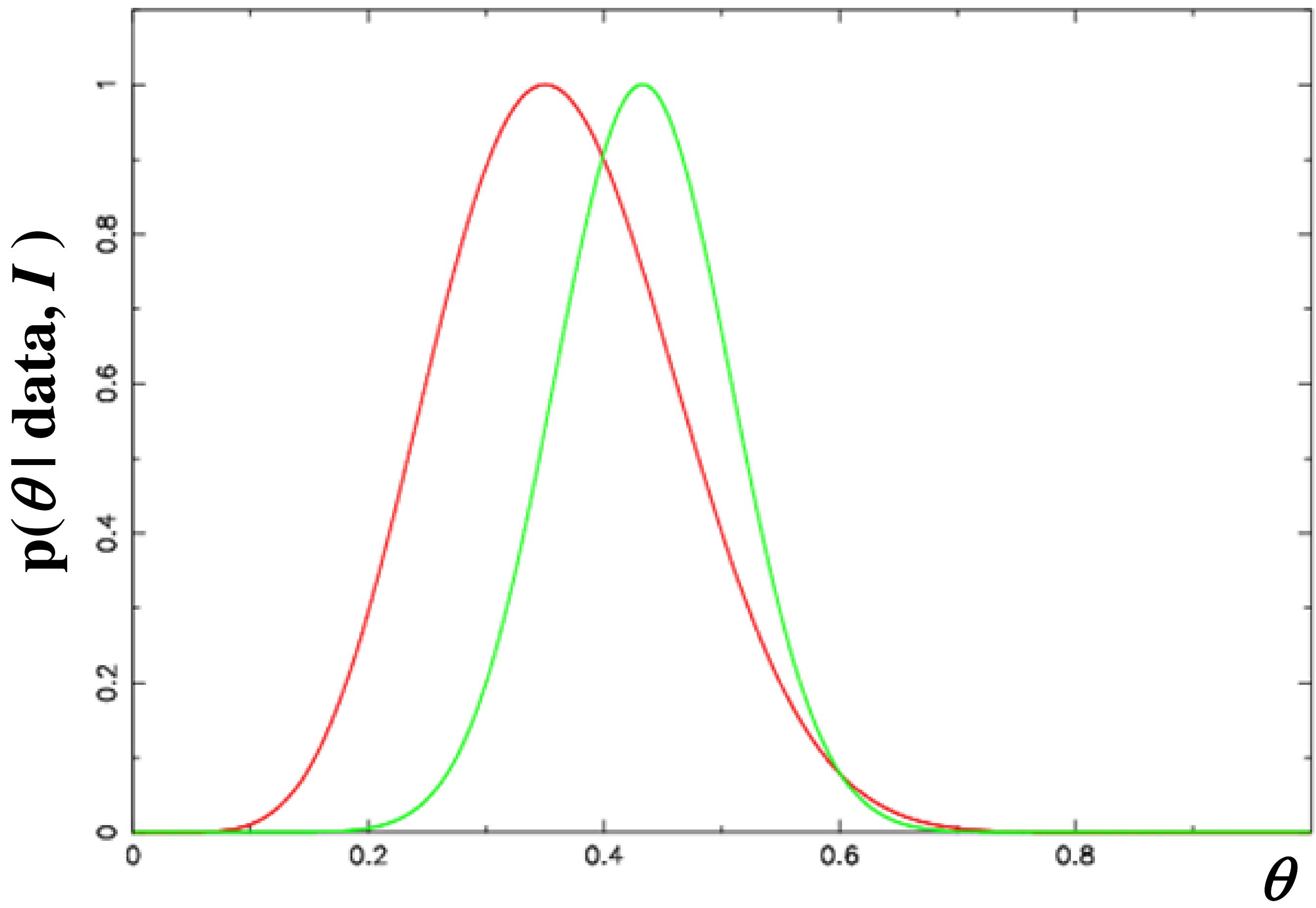
After N = 5 flips : H + H + T + T + T



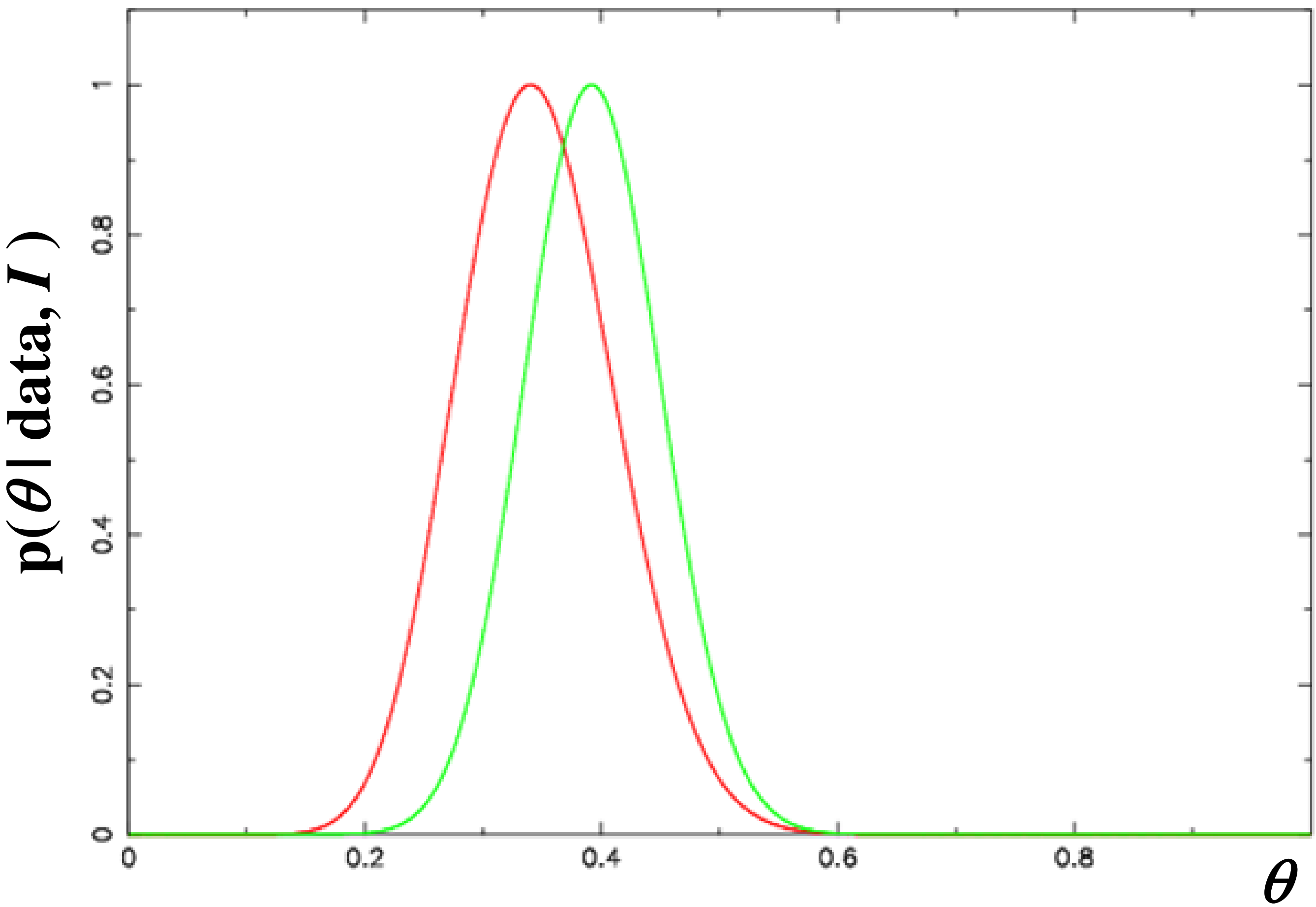
After N = 10 flips : 5 H + 5 T



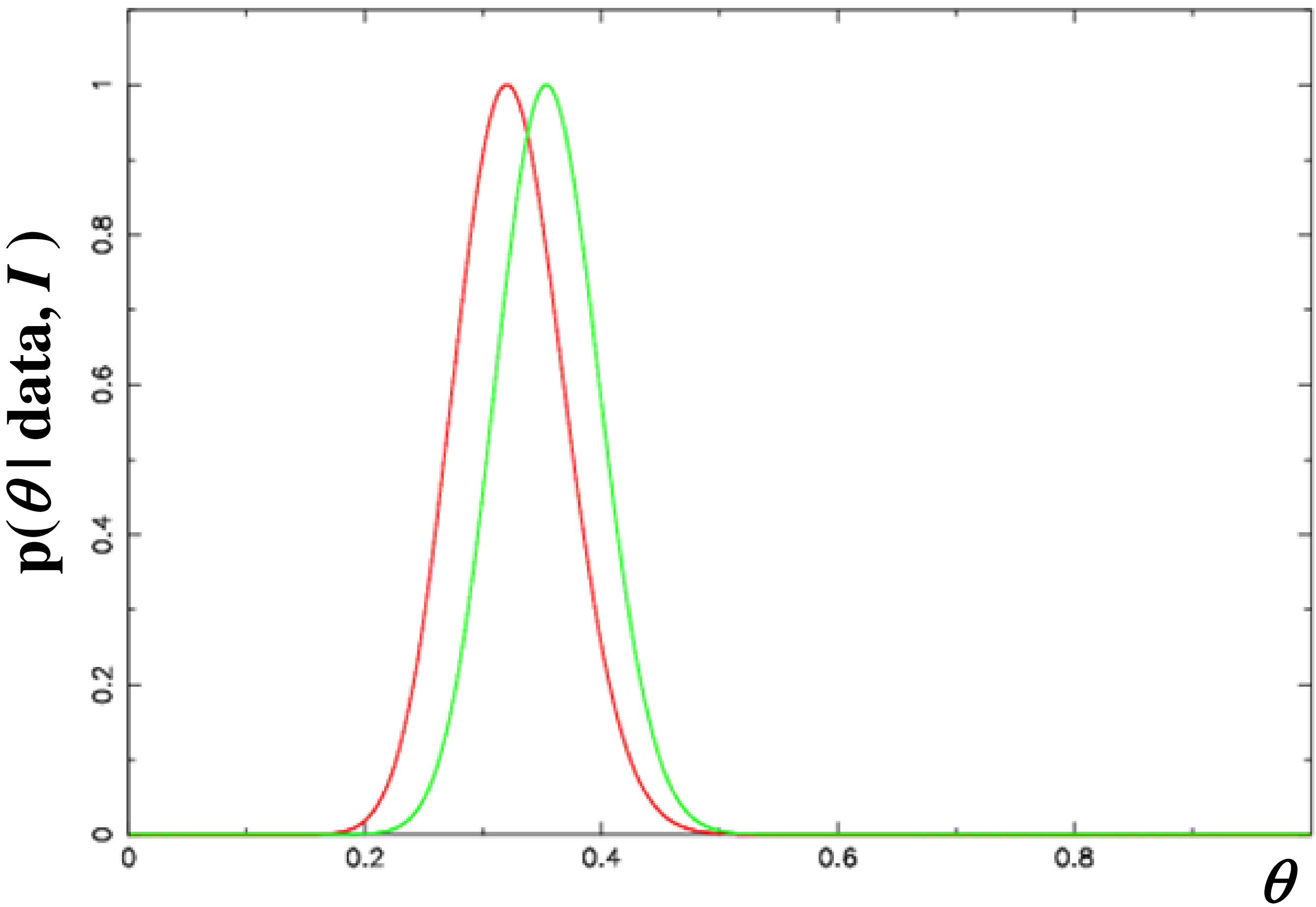
After N = 20 flips : 7 H + 13 T



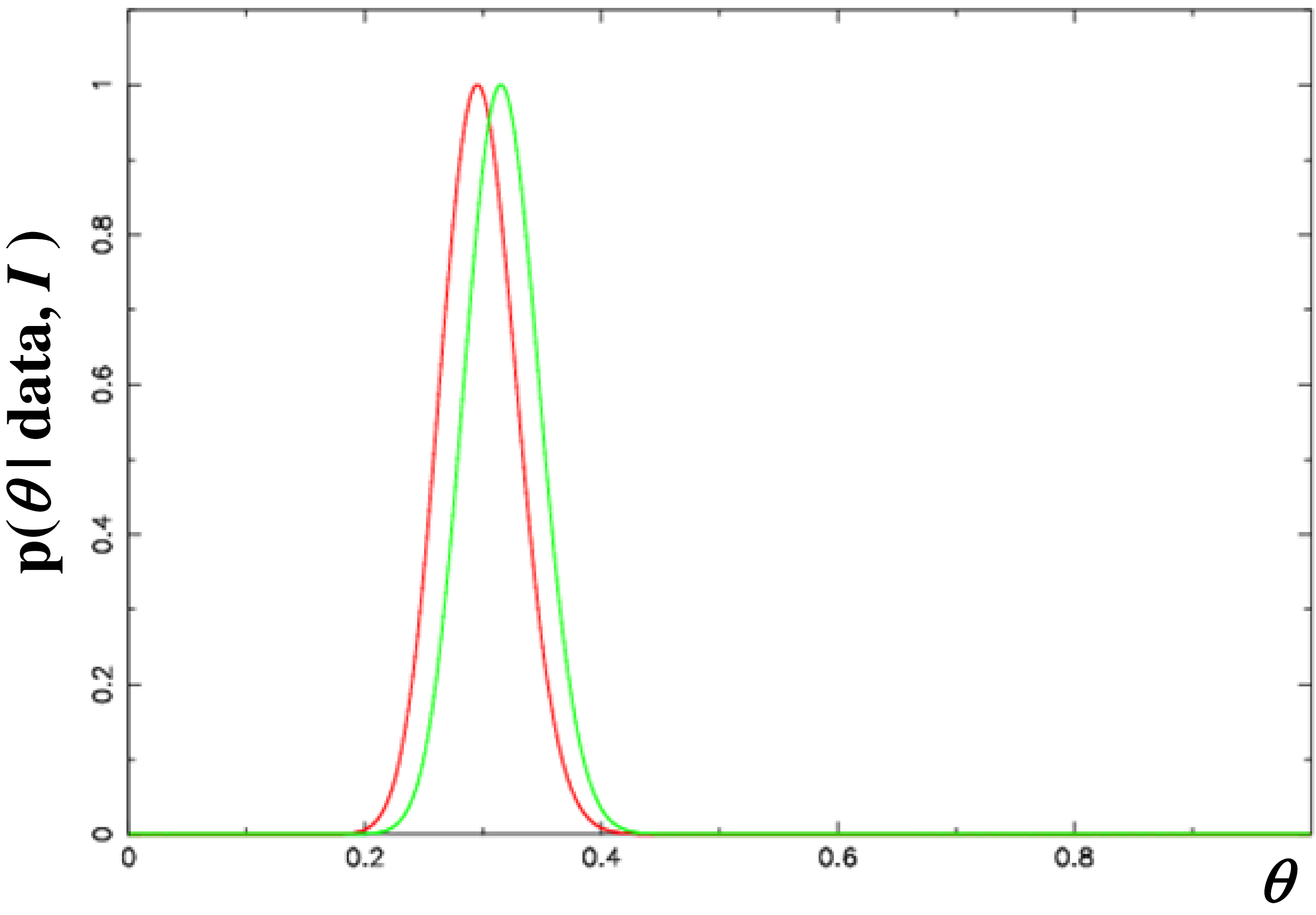
After N = 50 flips : 17 H + 33 T



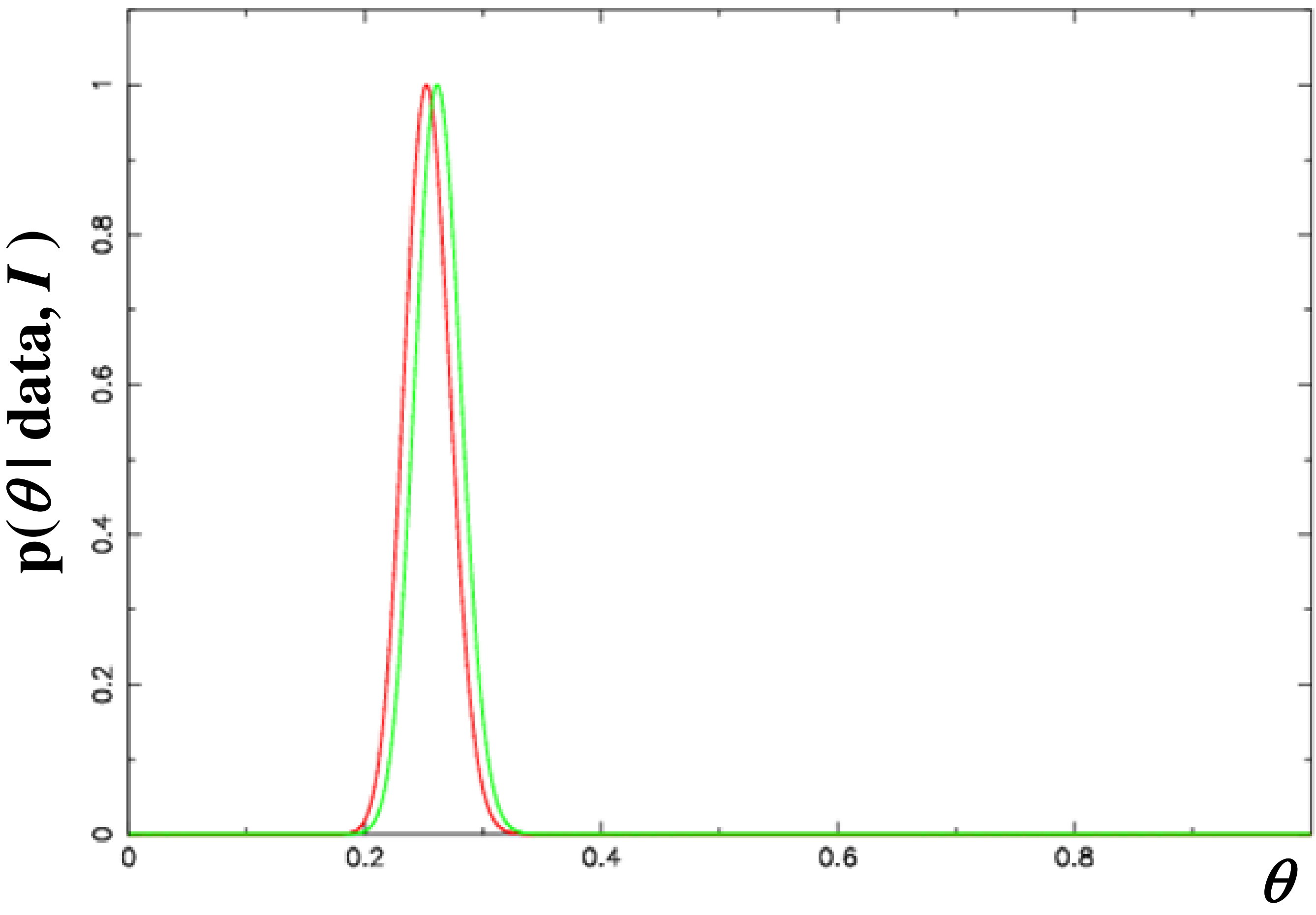
After N = 100 flips : 32 H + 68 T



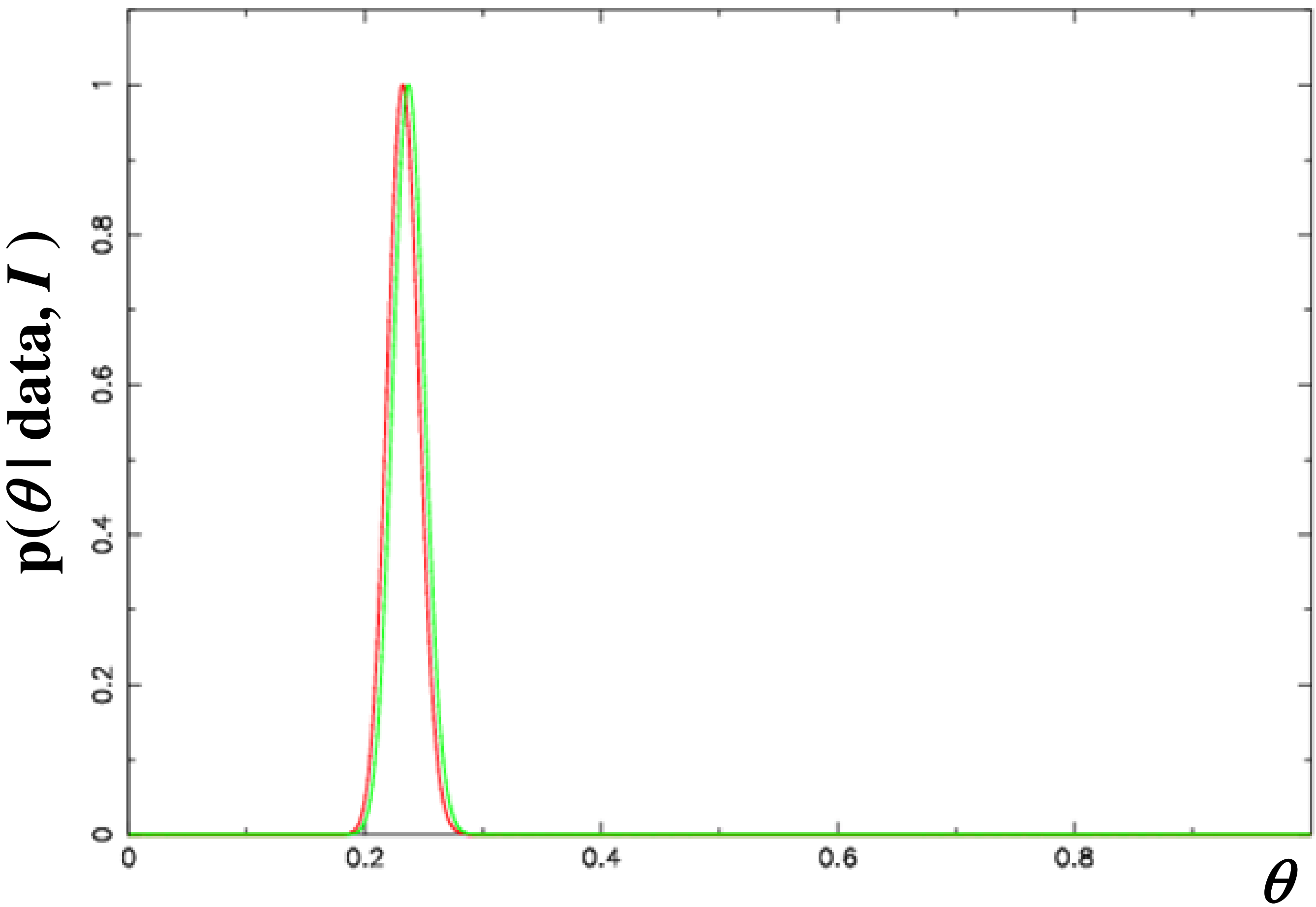
After N = 200 flips : 59 H + 141 T



After N = 500 flips : 126 H + 374 T



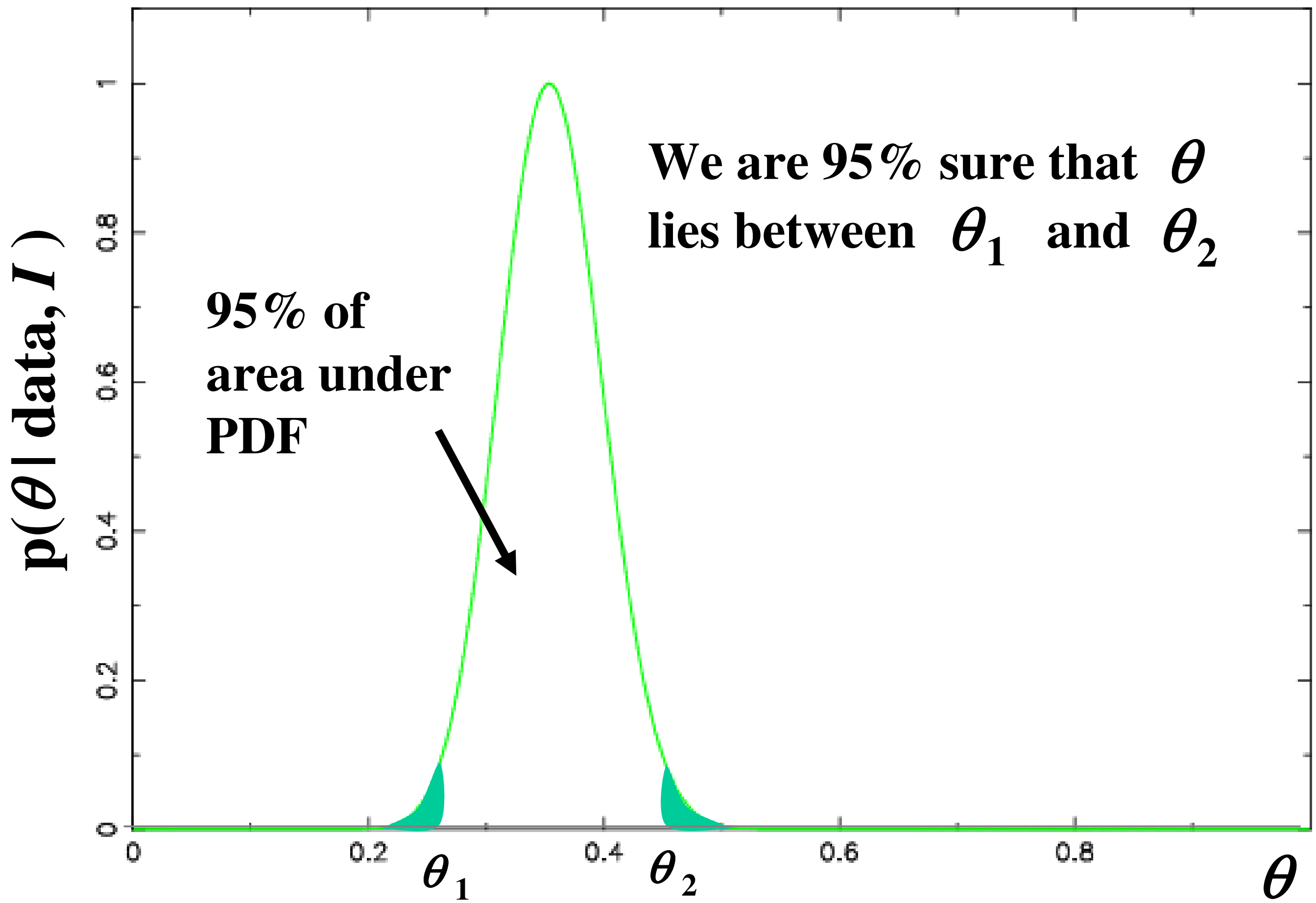
After N = 1000 flips : 232 H + 768 T



Bayesian Inference

- As data improves (and/or sample size increases), the posterior narrows and is less sensitive to choice of prior
- The posterior conveys our (evolving) degree of belief in *all* different values of θ, in the light of the observed data
- If we want to express our result as a *single number* we could use the posterior **mean**, **median**, or **mode**
- We can use the *variance* (or entropy) of the posterior to quantify the uncertainty of our belief in θ
- It is straightforward to define **credible intervals** (CI)

Bayesian Credible Intervals



Note: the credible interval is not unique, but we can define the ***shortest*** C.I.

Summary of Principles

1. **Probability theory is common sense reduced to calculation.**

2. **Given a model, we can derive any probability**

3. **Describe a model of the world, and then compute the probabilities of the unknowns with Bayes' Rule**

Problems with Bayesian methods

- **Best solution is usually intractable**
- often requires numerical computation
- But it's still far better to understand the real problem, be principled, and then approximate
- need to choose approximations carefully

Problems with Bayesian methods

2. Some complicated math needed

- Models are simple, but algorithms can be complicated
- But may still be worth it
- Bayesian toolboxes are out there

(e.g., BUGS, VIBES, Intel OpenPNL)

Problems with Bayesian methods

3. **Complex models sometimes impede creativity**

- Sometimes it's easier to tune (hack)
- Still, can hack first, be principled later
- Probabilistic models give insight that actually helps with hacking solutions

Benefits of Bayesian Approach

1. Principled modeling of noise/uncertainty
2. Unified model for learning and synthesis
3. We can learn all parameters
4. Can have more parameters than data
5. Good results from simple models
6. Especially good when data is scarce
7. Lots of new research and algorithms

Finally, some things to remember

"Probability does not exist"

– Bruno de Finetti

"All models are wrong. But some are useful"

– George E. P. Box

(son in law of Ronald Fisher)

The End