Why Bayesian?

Rigorous approach to address statistical estimation problems.

The Bayesian “philosophy” is mature and powerful.

Even if you aren’ t Bayesian, you can define an “uninformative” prior and everything reduces to maximum likelihood estimation!

It is easy to impose constraining knowledge (as priors).

It is easy to combine information from different types of sensors.

Why Bayesian?

Recursive estimators come naturally. Your posterior computed at time t-1 becomes the prior for time t. This is combined with the likelihood at time t, and renormalized to get the posterior at time t. This new posterior becomes the prior for time t+1, and so on....

Bayesian methods are crucial when you don’t have much data. With the use of a strong prior, you can make reasonable estimates from as little as one data point.

Bayes Rule

Thomas Bayes
1702-1761

Bayes rule can be derived by a simple manipulation of the rules of probability. But it has far-reaching consequences.

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

$$p(x|y) = \frac{p(y|x)p(x)}{p(y)}$$

also note...

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

$$p(x|y) = \frac{p(y|x)p(x)}{\int_x p(y|x)p(x)}$$

interpretation: multiply the prior times the likelihood, then normalize so that the result integrates to 1. This becomes the posterior distribution.

Bayes Rule

Thomas Bayes
1702-1761

Since we are normalizing anyways, our likelihood function only needs to be defined up to a constant

$$p(x|y) = \frac{p(y|x)p(x)}{\int_x p(y|x)p(x)}$$

$$\Downarrow$$

$$\boxed{p(x|y) = \frac{L(y|x)p(x)}{\int_x L(y|x)p(x)}}$$

However, often this normalization is computationally intractable. This will motivate our (later) interest in nonparametric and sampling based statistical methods.

Prior Distribution

information about the unknown parameters available from sources independent from the observed data. (these sources can, and usually do, include human intuition).

e.g.

- maximum speed of a walking person
- vehicles are likely to be seen on roads
- in this image, people are about 100 pixels high
- zebras are black and white and have stripes

Likelihood Function

This is the connection between an observation, yεY, and the unknown state, xεX.

L(y|x) = Pr{Y=y | X=x} says how likely it is that your data y was observed, given that the state of the world was x.

As depicted above, L(y|x) is a (conditional) probability distribution over the data measurement space Y.

That is: $\int_{y \in Y} L(y|x) = 1$

However, we are going to interpret y as being fixed (in fact, it is a value that we observed), and will treat x as the unknown value! As a function of x, L(y|x) does not integrate to 1. That is why it is called a likelihood function, not a likelihood distribution.

Posterior Distribution

For a Bayesian, the posterior distribution is the starting point for answering all well-posed statistical questions about the state.

e.g.

- What is the most likely location of this object?
  the mode of the posterior (MAP estimate)
- With what certainty do I know that location?
  spread of probability mass around the MAP estimate
- Are there other likely locations? If so, how many?
  analysis of multi-modality of the posterior

Important point: output of Bayesian approach is not a single point estimator, but a whole probability distribution over state values.

Some (overly) Simple Examples

We hope to gain some intuition about characteristics of Bayesian estimation through some simple examples:

- MLE vs Bayes: analysis of coin-tosses
- point observation model
- ray (bearings only) observation model
- combining bearings only + distance observations

Maximum Likelihood Estimation

Basic approach:

- form likelihood function L(y | x) as product of individual likelihoods L(yi | x)of each observation (assuming observations are iid!)
- treat it as a function of parameters L(x)
- take derivative of L(x) or log L(x) wrt x and set the equation(s) equal to 0
- solve for x
- should really also verify that this solution $\hat{x}$ is a maximum and not a minimum.

MLE Example: Coin Flip

example motivated by Bishop, PRML, chapter 2

I flip a coin N times and get m heads.
How do I estimate x = probability of heads?

$$p(y_i|x) = x^{y_i}(1-x)^{1-y_i}$$

Bernoulli Distribution

form likelihood function from N sample observations yi

$$p(\mathbf{y}|x) = \prod^{N} p(y_i|x) = \prod^{N} x^{y_i}(1-x)^{1-y_i}$$

take log of that. This is what we want to maximize, as a function of x

$$\ln p(\mathbf{y}|x) = \sum_{n=1}^{N} \ln p(y_i|x) = \sum_{n=1}^{N} \{y_i \ln x + (1-y_i)\ln(1-x)\}$$

take deriv wrt x and set to 0, then solve for x [will do on board]

$$\hat{x} = \frac{\sum y_i}{N} = \frac{m}{N}$$

(proportion of observations that were heads – result is completely intuitive!)

MLE Example: Coin Flip

So, what, if anything, is wrong with that?

Well, first of all, I had to do it in batch mode, rather than one observation at a time. We' ll save this complaint until later.

Let' s say I flip a coin 3 times, and every time I get heads. Should I then say my probability of getting heads is 100 percent?

No, my belief about the way coins work leads me to think that is overly-optimistic.

Solution: encode my prior knowledge that flipped coins tend to come up tails sometimes...

Bayesian Coin Flip Analysis

I flip a coin N times and get m heads.
How do I estimate x = probability of heads?

$$p(y_i|x) = x^{y_i}(1-x)^{1-y_i}$$ **Bernoulli Distribution**

form likelihood function from N sample observations yi

$$p(\mathbf{y}|x) = \prod^{N} p(y_i|x) = \prod^{N} x^{y_i}(1-x)^{1-y_i}$$

So far, this is an exact copy of what we did for MLE. The difference will be encoding my prior knowledge about coins typically being fair. I need a prior distribution on x. Here is where a little “art” comes in. Notice that our likelihood is of the form

x^something * (1-x)^somethingelse

We would like a prior distribution that also has that form, so that the posterior has that form too! Prior distributions that have this nice property are called “conjugate” priors.

Bayesian Coin Flip Analysis

Beta distribution is a conjugate prior for estimating the x parameter in a Bernoulli distribution.

$$p(x) = \beta(x|a,b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} x^{a-1}\,(1-x)^{b-1}$$

pictures
of B(x|a,b)
for various
values of a,b

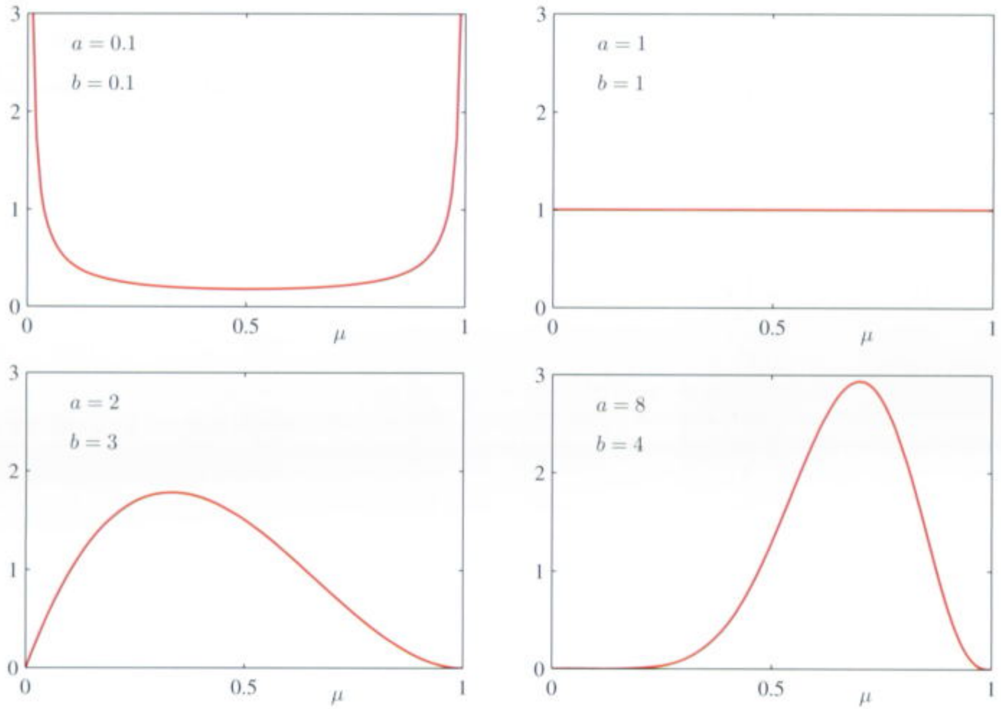


Figure 2.2 Plots of the beta distribution $\mathrm{Beta}(\mu|a,b)$ given by (2.13) as a function of μ for various values of the hyperparameters a and b.

aside...

How do we know what values of a and b to pick????

pictures
of B(x|a,b)
for various
values of a,b

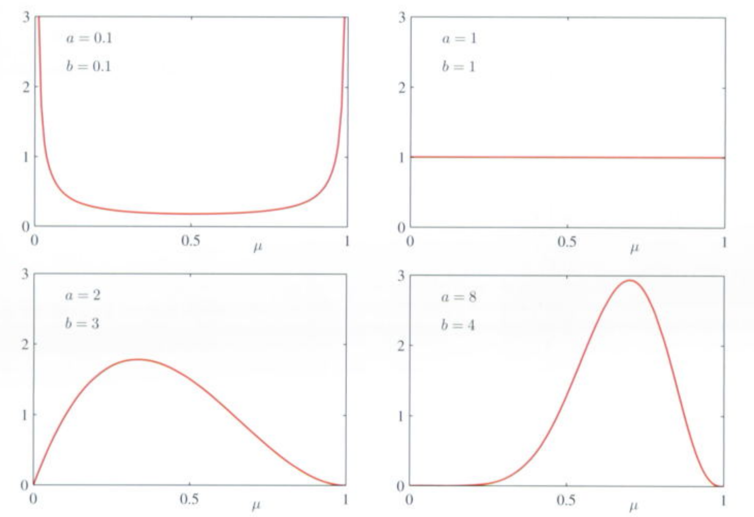


Figure 2.2 Plots of the beta distribution $\mathrm{Beta}(\mu|a,b)$ given by (2.13) as a function of μ for various values of the hyperparameters a and b.

This is where true Bayesian analysis starts to get scarily self-referential. You treat them as unknowns to be estimated. And so... they have their own prior distributions, which have their own parameters (hyperparameters), which need to be selected, so maybe there are even more priors and hyperparameters... and so you sit on the brink and contemplate infinity [or "Bayesian frenzy", as David Capel calls it].

Or, we could just say a=2 and b=2.

Bayesian Coin Flip Analysis

Multiply the prior (Beta distribution) times the Likelihood function (product of Bernoullis), to get the posterior.

$$p(x|y) \propto p(x)L(y|x)$$

$$= \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} x^{a-1} (1-x)^{b-1} \prod^{N} x^{y_i} (1-x)^{1-y_i}$$

$$\propto x^{a-1} (1-x)^{b-1} \prod^{N} x^{y_i} (1-x)^{1-y_i}$$

$$= x^{a-1} (1-x)^{b-1} x^{m} (1-x)^{l}$$

m = number observed heads
l = number of observed tails
a,b = parameters of Beta prior

$$= x^{m+a-1} (1-x)^{l+b-1}$$

Rather than compute an integral to normalize, we can note that this is in the form of a Beta distribution, so could just write it down (recall, it is the ratio of some Gamma functions).

Bayesian Coin Flip Analysis

We have computed the posterior, and it is a Beta function

$$p(x|a,b,m,l) \propto x^{m+a-1}\,(1-x)^{l+b-1}$$

We now have a whole function (distribution) for x. What if we just want a single number? We could compute its expected (mean) value, for example.

$$\hat{x} \equiv \int_{x=0}^{1} x\,p(x|a,b,m,l)\,\mathrm{dx}$$

We won’t derive it here, but it is the mean of a Beta distribution, and in our case has value

$$\hat{x} = \frac{m+a}{m+a+l+b}$$

m = number observed heads
l = number of observed tails
a = number of “fictitious” heads
b = number of “fictitious” tails

note how this result allows us to understand the a,b parameters as a “virtual” number of heads and tails included to the dataset in addition to the observed data.

Bayesian Coin Flip Analysis

For example, using a=2, b=2 (prior belief that the coin is fair), if we are allowed only one observation, and it is heads, then instead of inferring that the probability of heads is 100% (as MLE would tell us), we instead see...

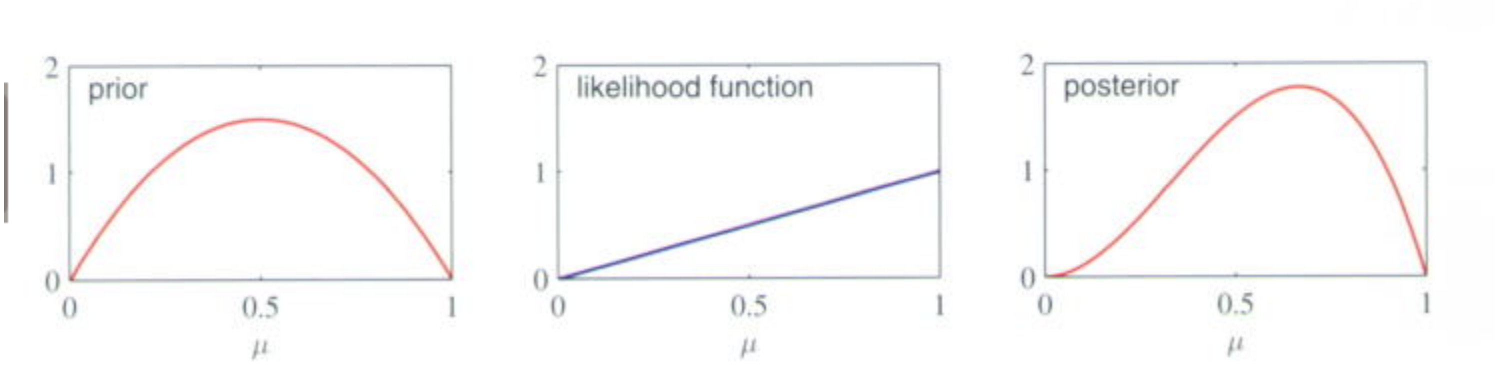


Figure 2.3 Illustration of one step of sequential Bayesian inference. The prior is given by a beta distribution with parameters $a = 2$, $b = 2$, and the likelihood function, given by (2.9) with $N = m = 1$, corresponds to a single observation of $x = 1$, so that the posterior is given by a beta distribution with parameters $a = 3$, $b = 2$.

The distribution shifts towards heads being more likely, but it stays well-behaved. And assuming we later observe some tails, it will shift back towards “fair”, and so on.

Example 2: Gaussian Point Observation

Assume state x is a 1D location. It has some prior distribution that we will take as a normal distribution. We also will get an observation y, which is noisy, and we will characterize the noise by a zero-mean normal distribution. What is the posterior distribution on x, given the prior and the observation?

Our model

$$\begin{aligned}
p(x) &= N(\mu, \sigma^2) && (1)\\
y &= x + \varepsilon && (2)\\
p(\varepsilon) &= N(0, s^2) && (3)\\
p(y|x) &= N(x, s^2) && (4)\\
p(x|y) &\propto p(x)p(y|x) && (5)
\end{aligned}$$

Example 2: Gaussian Point Observation

Recall the form of a 1D Normal distribution...

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

$$\propto \exp\left\{-\frac{1}{2}\left[\frac{1}{\sigma^2}x^2 - \frac{2\mu}{\sigma^2}x + \frac{\mu^2}{\sigma^2}\right]\right\}$$

useful fact for later on... if we ever get an exponential posterior distribution of the form

$$\exp\left\{-\frac{1}{2}\left[ax^2 - 2bx + c\right]\right\}$$

we can identify it as a Normal distribution, with parameters

$$\sigma^2 = 1/a \qquad \mu = b/a$$

this is called
“completing the square”

Example 2: Gaussian Point Observation

Back to our example, combining our prior and likelihood yields

$$\begin{aligned} p(x) &= N(\mu,\sigma^2) &= c_1 \exp\left\{-\tfrac{1}{2}\tfrac{(x-\mu)^2}{\sigma^2}\right\} \\ p(y|x) &= N(x,s^2) &= c_2 \exp\left\{-\tfrac{1}{2}\tfrac{(y-x)^2}{s^2}\right\} \end{aligned}$$

$$\begin{aligned} p(x|y) &\propto p(x)p(y|x) \\ &\propto c_3 \exp\left\{-\frac{1}{2}\left[\frac{(x-\mu)^2}{\sigma^2}+\frac{(y-x)^2}{s^2}\right]\right\} \\ &\propto \exp\left\{-\frac{1}{2}\left[\underbrace{\left(\frac{s^2+\sigma^2}{s^2\sigma^2}\right)}_{\mathbf{a}}x^2-2\underbrace{\left(\frac{s^2\mu+\sigma^2 y}{s^2\sigma^2}\right)}_{\mathbf{b}}x+\underbrace{\left(\frac{s^2\mu^2+\sigma^2 y^2}{s^2\sigma^2}\right)}_{\mathbf{c}}\right]\right\} \end{aligned}$$

Example 2: Gaussian Point Observation

Based on our earlier slide, we can identify this as a Normal distribution, so we now know our posterior distribution is also Normal:

$$p(x|y) = N(\mu_{new}, \sigma^2_{new})$$

And completing the square, we see that

$$\sigma^2_{new} = 1/a = \frac{s^2\sigma^2}{s^2+\sigma^2}$$

$$\mu_{new} = b/a = \frac{s^2\mu+\sigma^2 y}{s^2+\sigma^2}$$

Example 2: Gaussian Point Observation

Instead of using the whole posterior distribution, we may want just a single point estimate. Since a Normal distribution is unimodal, it is quite natural to choose the mean (which is also the mode).

Rewriting the mean slightly, we see that our estimator is

$$\hat{x} = \frac{\frac{1}{s^2}y + \frac{1}{\sigma^2}\mu}{\left(\frac{1}{s^2} + \frac{1}{\sigma^2}\right)}$$

This is a weighted mean of the prior location and observed location. Each location is weighted by the inverse of its variance (so low variance yields a high weight, and vice versa).

Verify to yourself that this estimator makes a lot of sense, by plugging in some different values for s and sigma.

Bearings-Only Example

Some sensors (like cameras) don’t explicitly measure locations of objects in the world, but can only measure “directions” towards objects in the world.

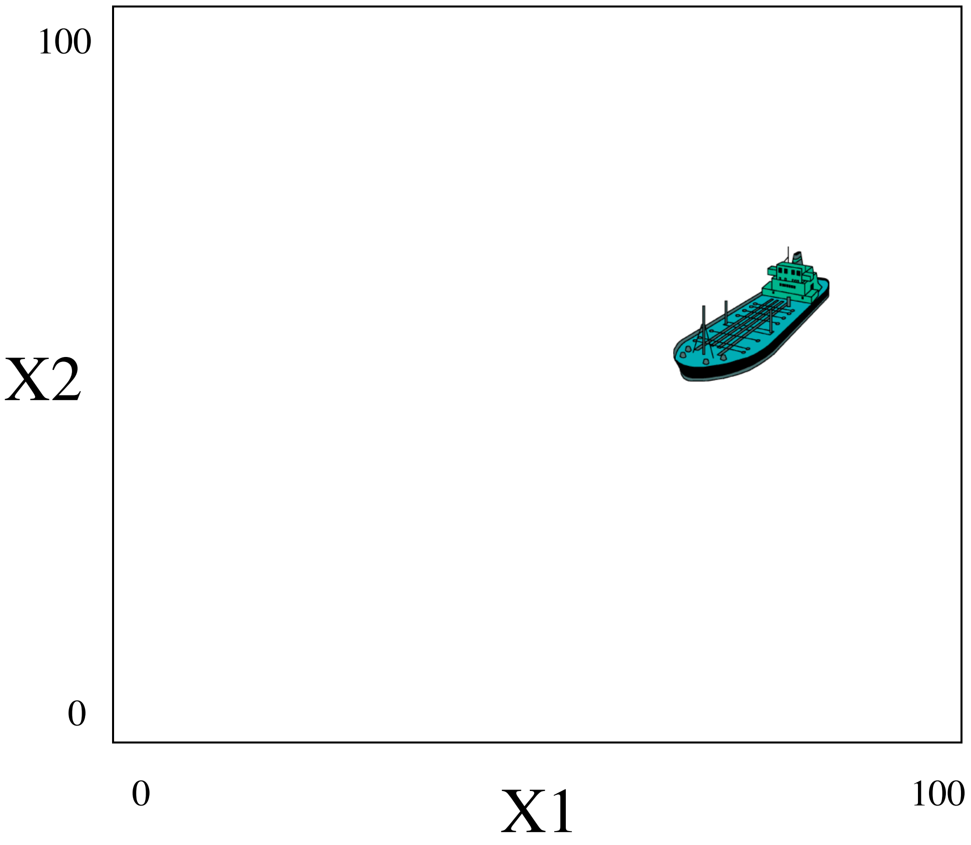
For cameras, these directional observations are called viewing rays.

For other kinds of sensors, they are called “bearings-only” measurements.

Bearings-Only Example

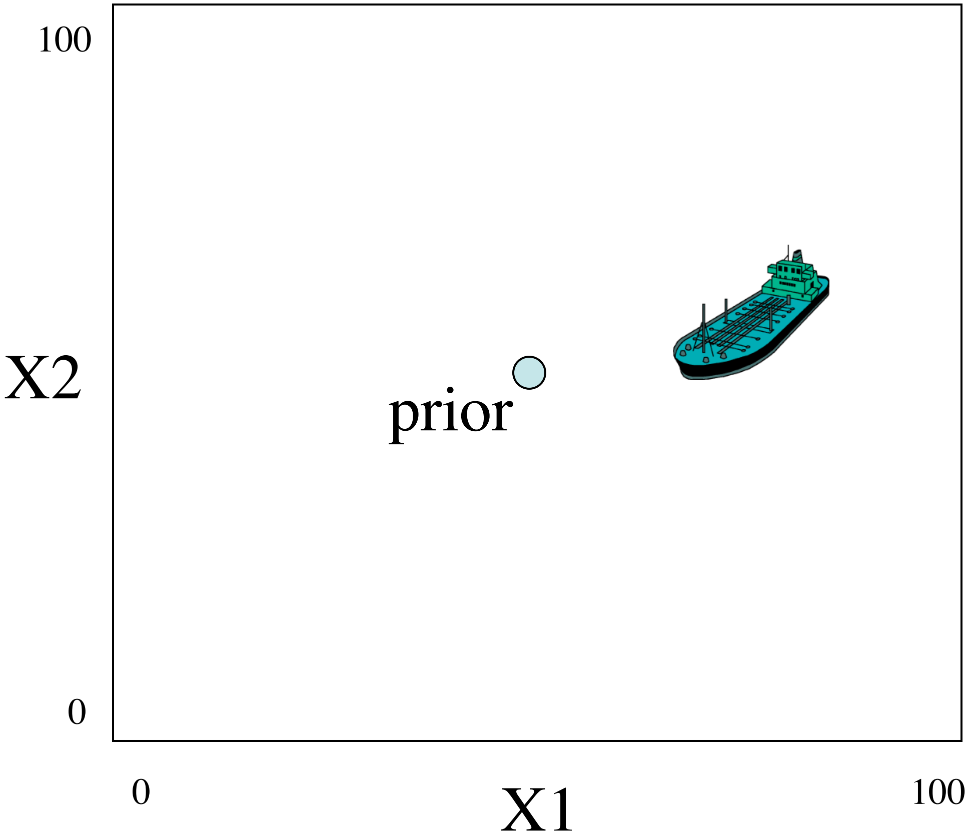
Let’s assume our target state x is a 2D location (x1, x2), and that our target is within some bounded region of interest, say

$0 < x1 < 100$ and $0 < x2 < 100$



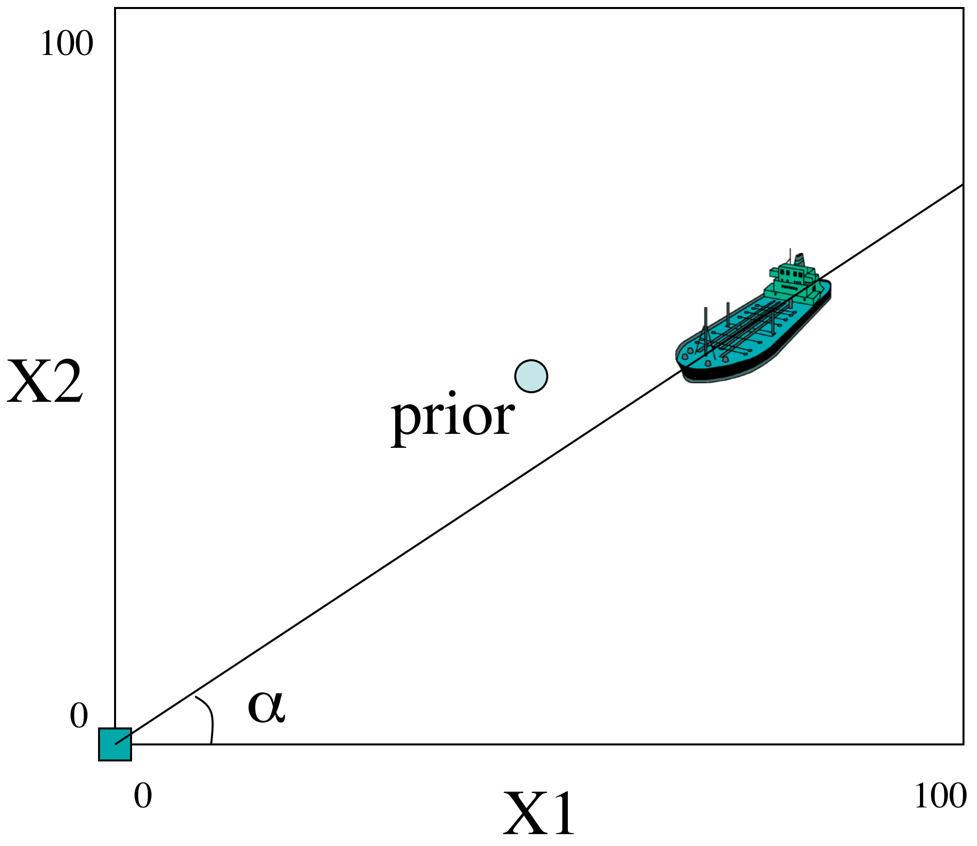
Bearings-Only Example

Initially we don't know the target is, so we guess it is at (50,50). We are uncertain, so model that uncertainty as a Gaussian with high variance (truncated and normalized so all mass lies within the region of interest)



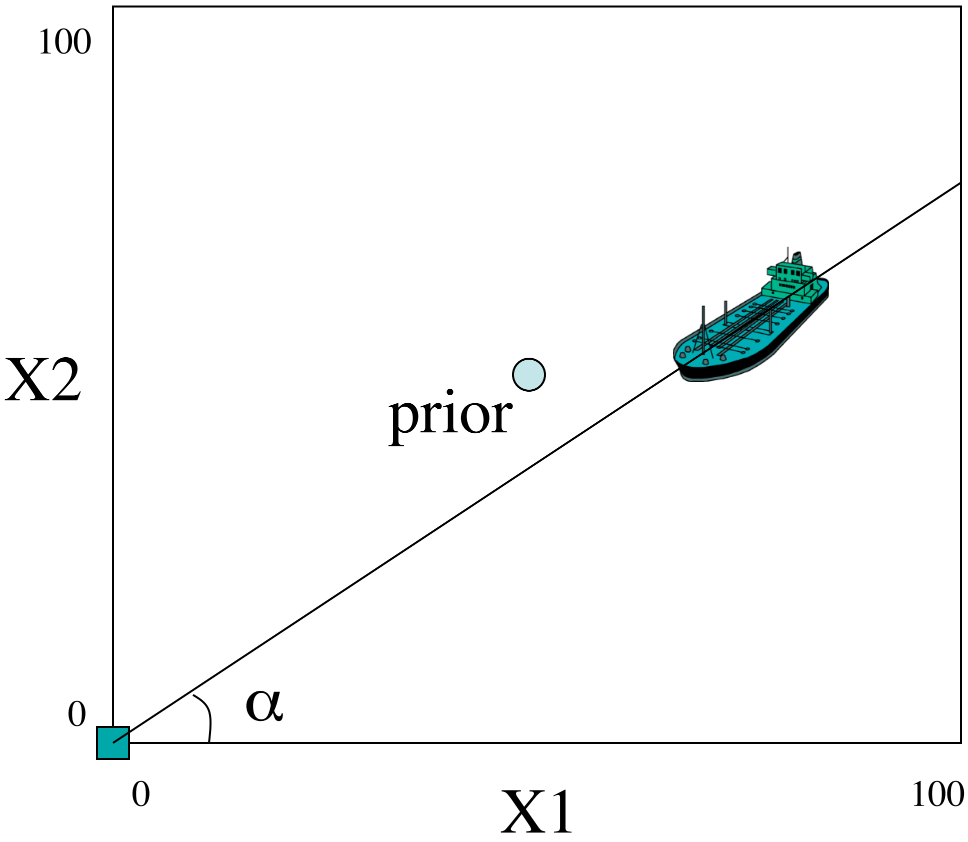
Bearings-Only Example

A bearings-only sensor located at (0,0) takes a noisy reading of the angle (alpha) towards the target. We will model the difference between the measured angle and actual angle as a zero-mean, 1D Gaussian.



Bearings-Only Example

Unlike previous examples, our distributions here are messy to deal with in closed form, due to truncation within a finite region of interest (as well as angular data, which can’t really be Gaussian) .

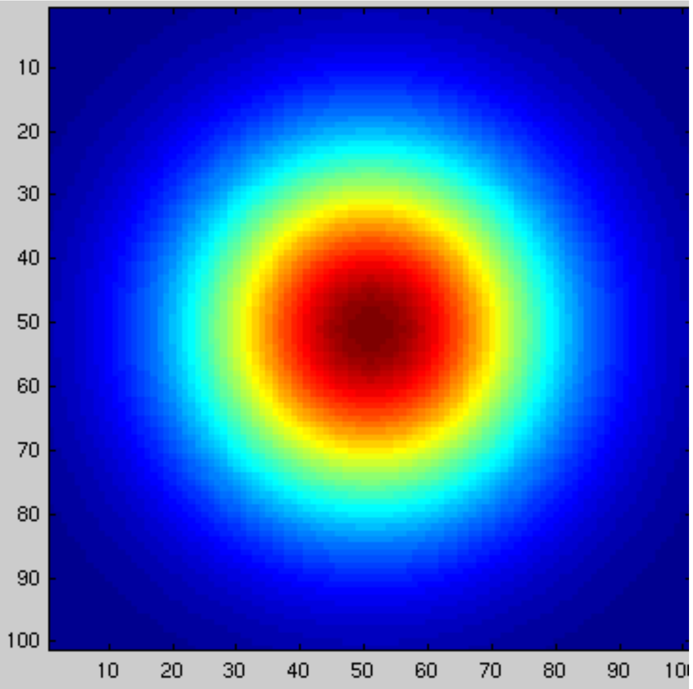


Bearings-Only Example

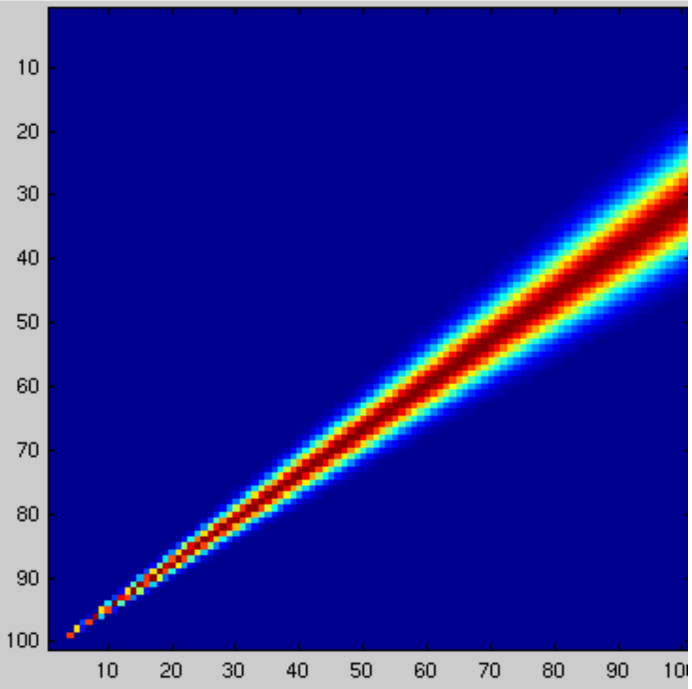
To deal with our intractable functions, we will “gridify” the region of interest into a discrete grid of spatial cells, and compute the probability mass within each cell.

This is a first, simple example of using a non-parametric representation to describe probability density functions!

Bearings-Only Example



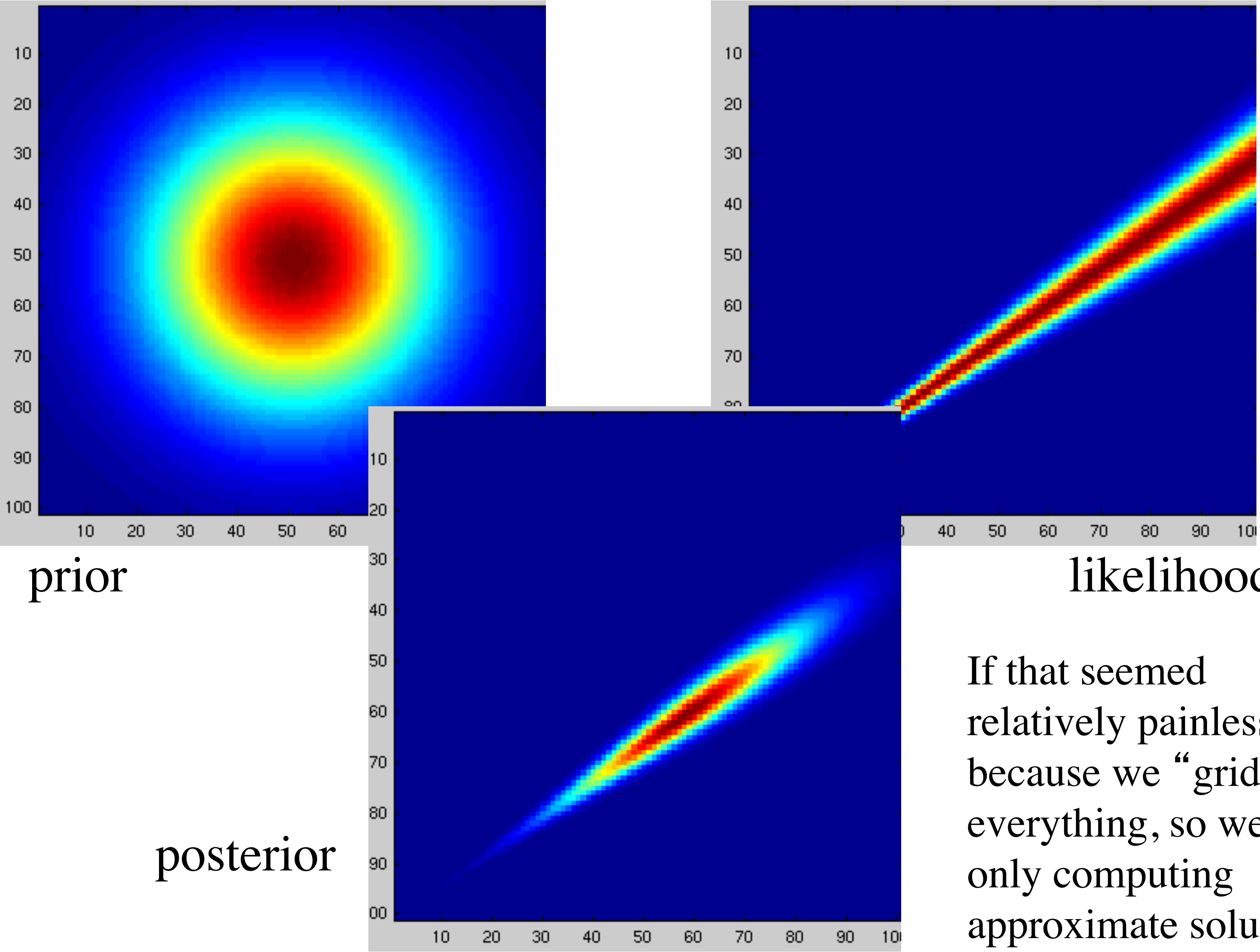
prior



likelihood

to combine using Bayes rule: point-wise multiply the prior times the likelihood, then renormalized the result so that to total mass sums up to 1.

Bearings-Only Example



prior

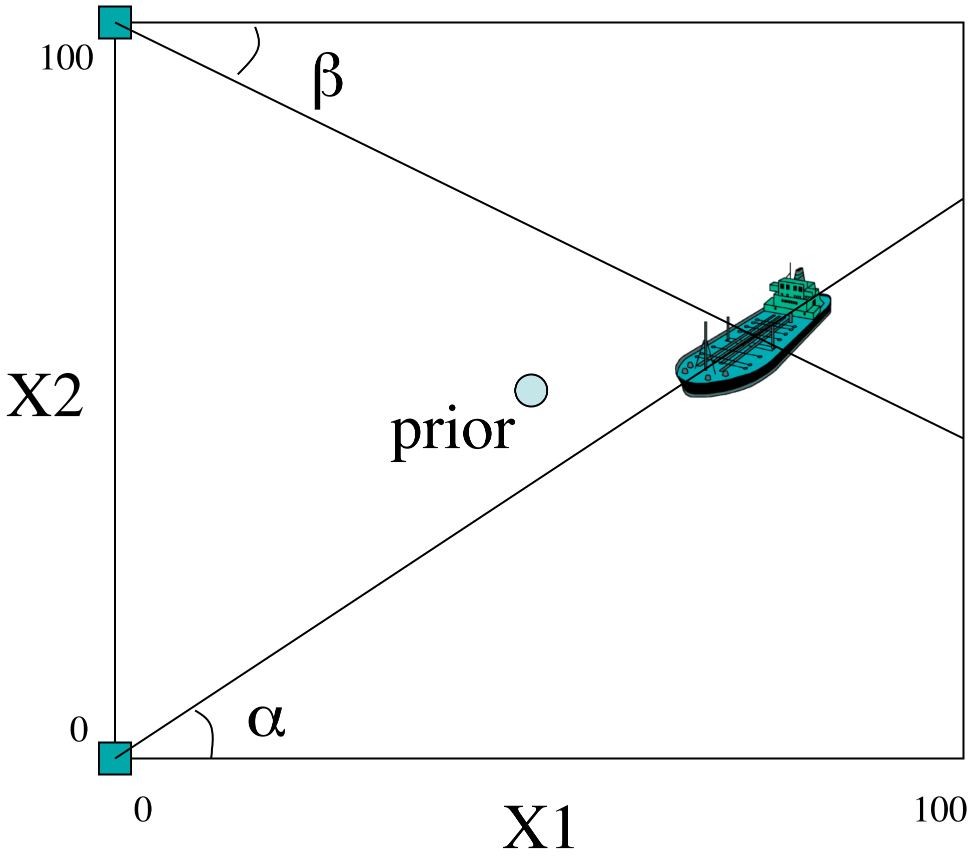
likelihood

posterior

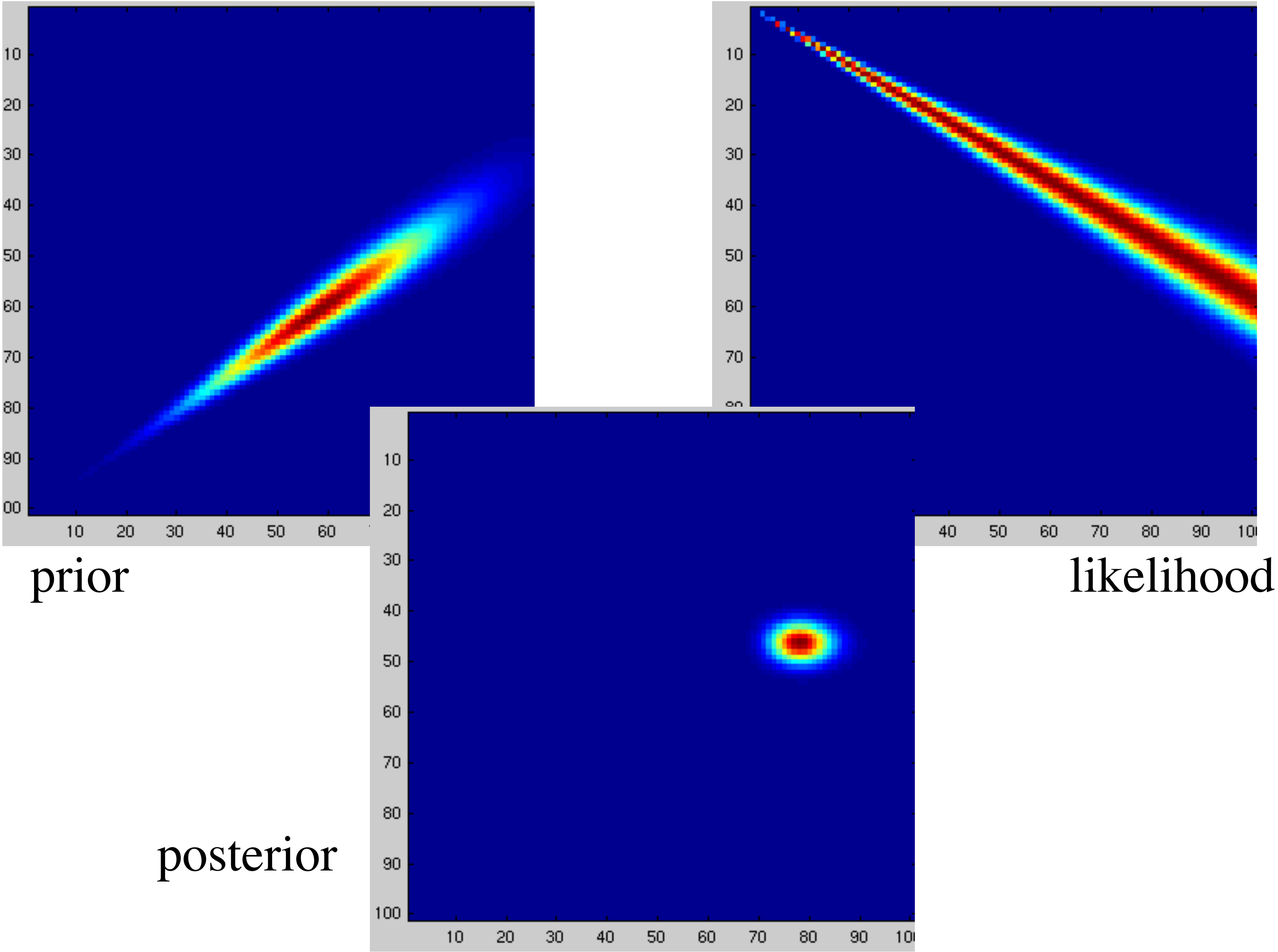
If that seemed relatively painless, it is because we “gridded” everything, so we are only computing approximate solutions.

Bearings-Only Example

Say a second sensor at (0,100) also takes a noisy bearings measurement of the target. We this have another likelihood function to represent this second observation.



Bearings-Only Example



prior

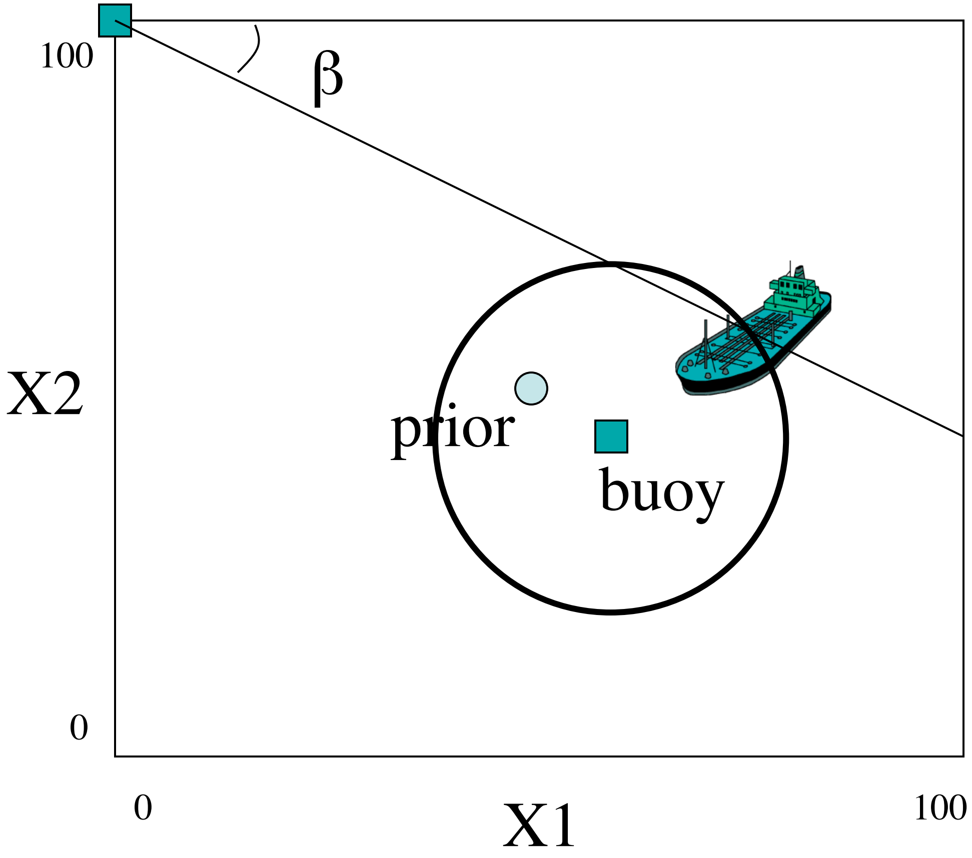
likelihood

posterior

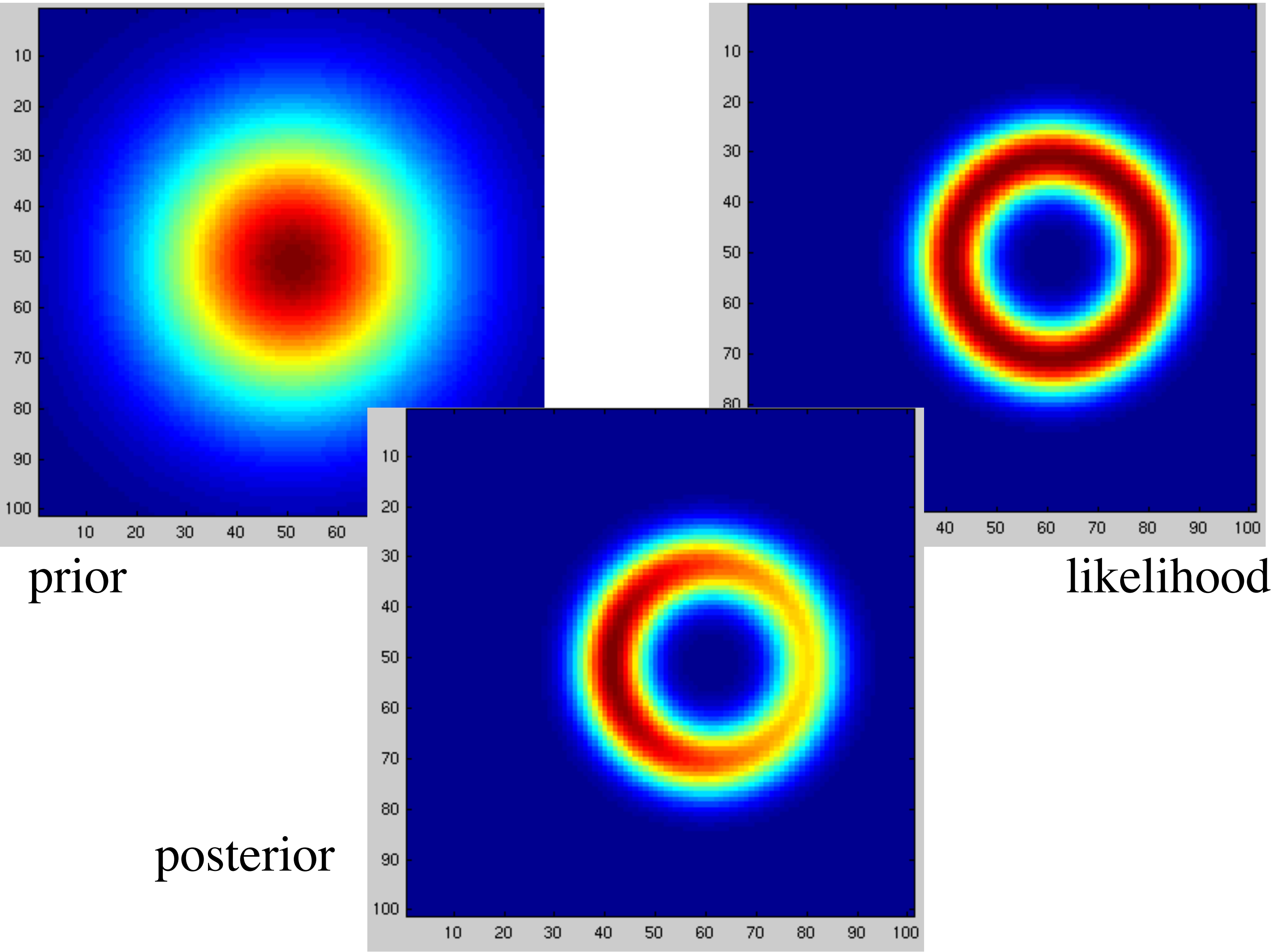
Range+Bearings Example

Bayesian methods are good for combining information from different kinds of sensors (sensor fusion). Let's say our ship wants to be found, and is broadcasting a radio signal, picked up by a transmitter on a buoy. That gives us a known distance to the ship.

A second, bearings-only reading, is also taken...



Range+Bearings Example

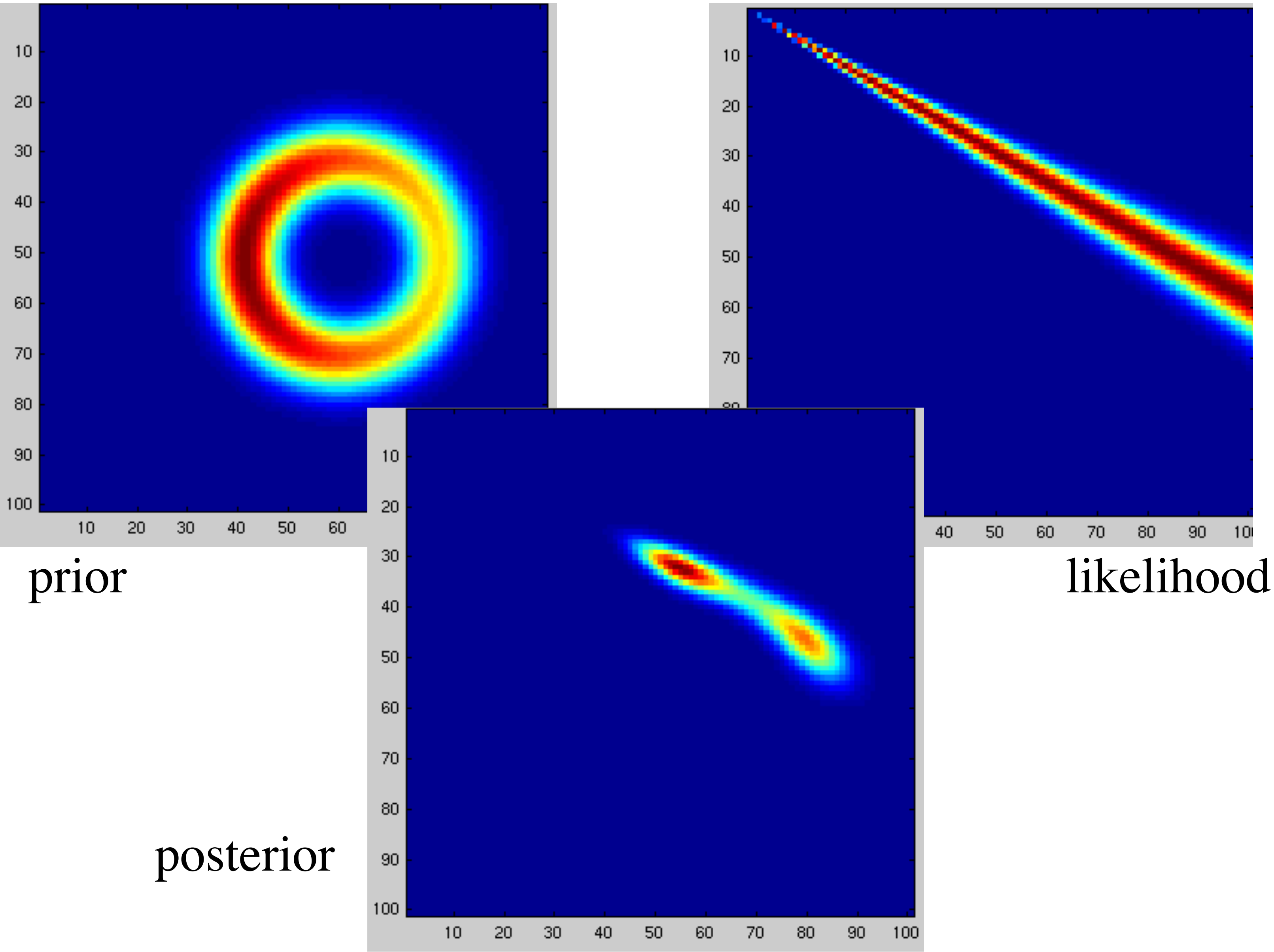


prior

likelihood

posterior

Range+Bearings Example



prior

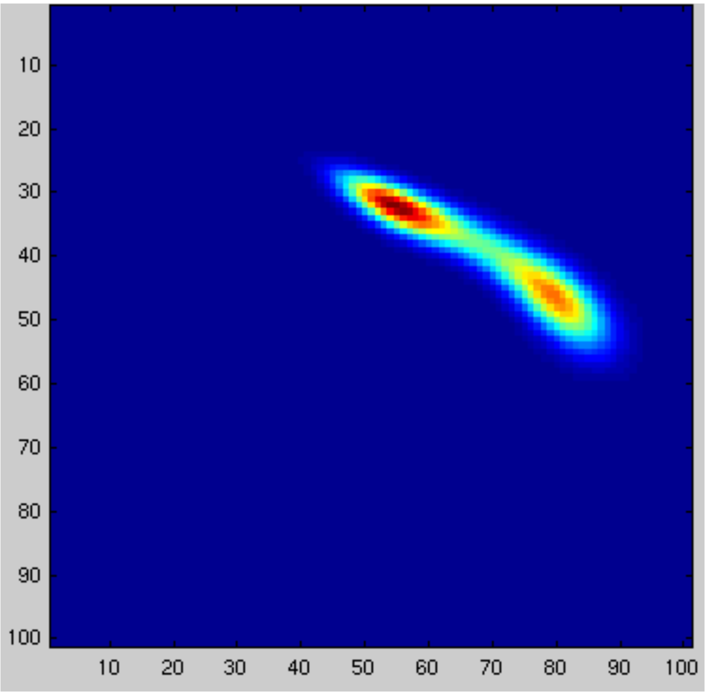
likelihood

posterior

Range+Bearings Example

This example points out that the posterior distribution may be multimodal. Presumably a second bearings reading from a sensor at the lower left of the region would disambiguate the location. But it is important that the multimodality be preserved, in order for that to happen! If we only kept the highest peak (in this example), we would get the wrong answer.

posterior


10
20
30
40
50
60
70
80
90
100
10
20
30
40
50
60
70
80
90
100