

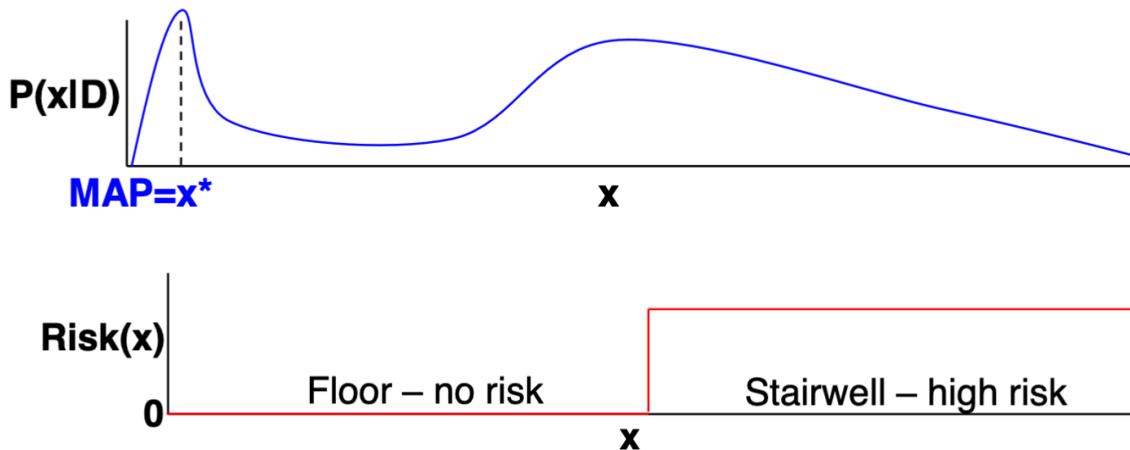
Introduction to Bayesian Data Analysis

MIE223
Winter 2025

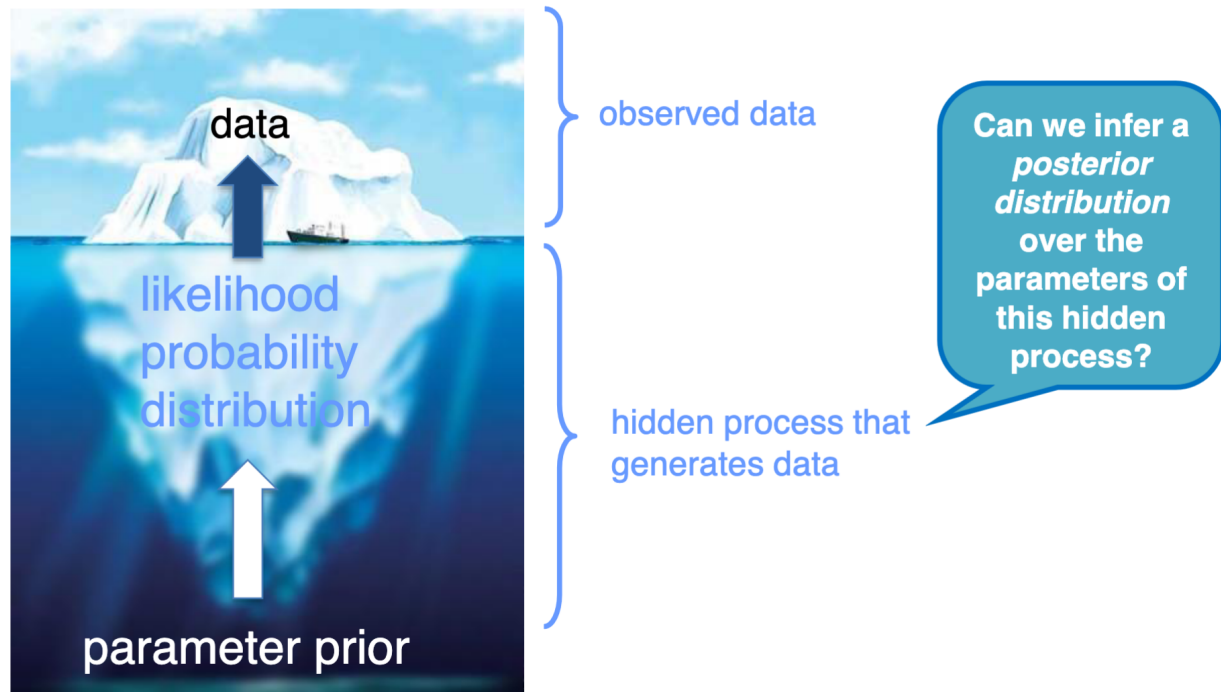
1 Bayesian Data Analysis

1.1 Bayesian Motivation: Bayesian Decision Theory

- Robot has belief $P(x|D)$ over position
 - D consists of noisy range finder readings
- Associate $Risk(x)$ w position x (e.g., stairs!)
 - $MAP\ Risk = Risk(x^*) = 0$ (Max Posterior)
 - $E(p(x|D))(Risk(x))$
 - Full Bayesian Risk = $\int Risk(x)p(x|D) dx > 0$
- Which risk estimate would you use?



1.2 Bayesian Generative View of Data



1.3 A coin tossing example

- Suppose we know nothing about coins except that each tossing event produces a head with some unknown probability p and a tail with probability $1-p$.
 - Our model of a coin has one parameter, p .
- If we observe 1 heads in 1 toss, what is p ?
- If we observe 5 heads in 10 tosses, what is p ?
- If we observe 53 heads in 100 tosses, what is p ?
 - How confident are we in each case?
- The frequentist answer (also called maximum likelihood): Pick the value of p that makes the observation of 53 heads and 47 tails most probable.
 - This value is $p=0.53$

1.4 A coin tossing example: the math

probability of a particular sequence containing 53 heads and 47 tails. \Rightarrow

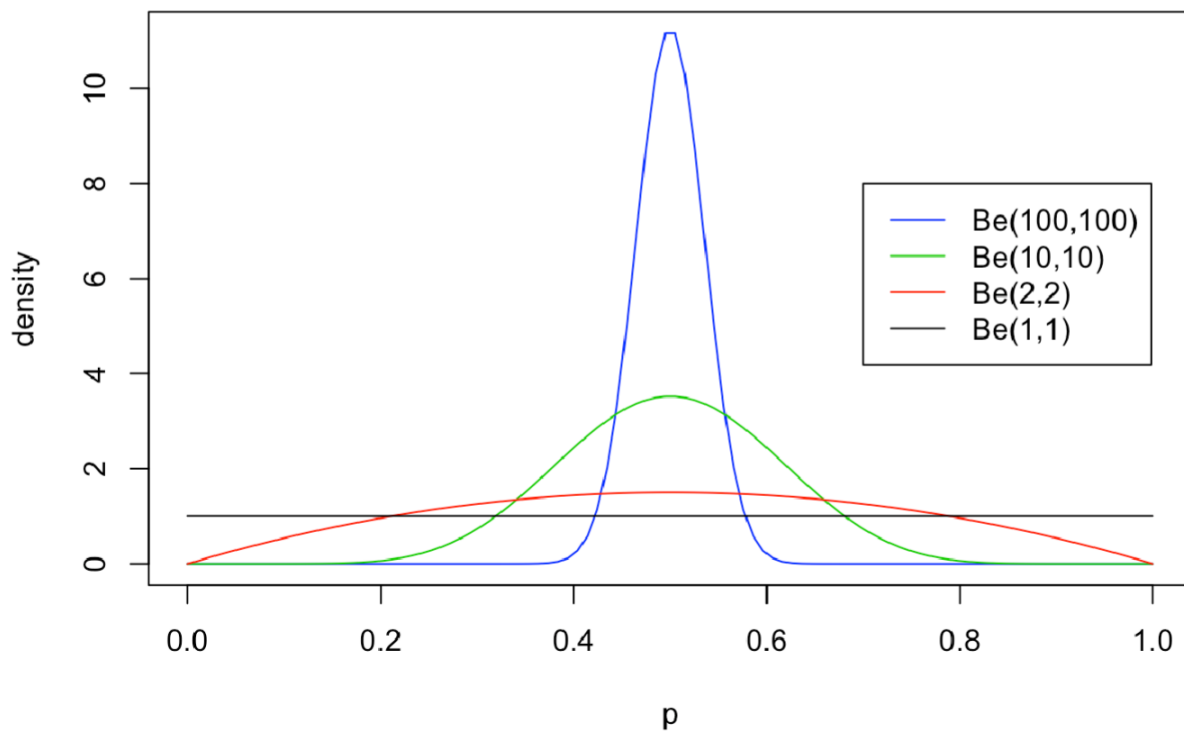
$$P(D) = p^{53}(1-p)^{47}$$
$$\frac{dP(D)}{dp} = 53p^{52}(1-p)^{47} - 47p^{53}(1-p)^{46}$$
$$= \left(\frac{53}{p} - \frac{47}{1-p} \right) [p^{53}(1-p)^{47}]$$
$$= 0 \text{ if } p = .53$$

1.5 Some problems with picking the parameters that are most likely to generate the data

- What if we only tossed the coin once and we got 1 head?
 - Is $p=1$ a sensible answer?
 - Surely $p=0.5$ is a much better answer.
- Is it reasonable to give a single answer?
 - If we don't have much data, we are unsure about p .
 - Our computations of probabilities will work much better if we take this uncertainty into account.

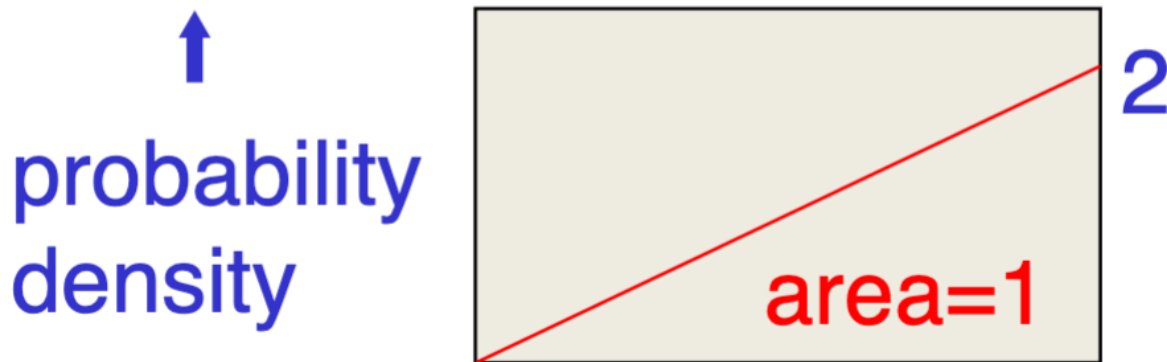
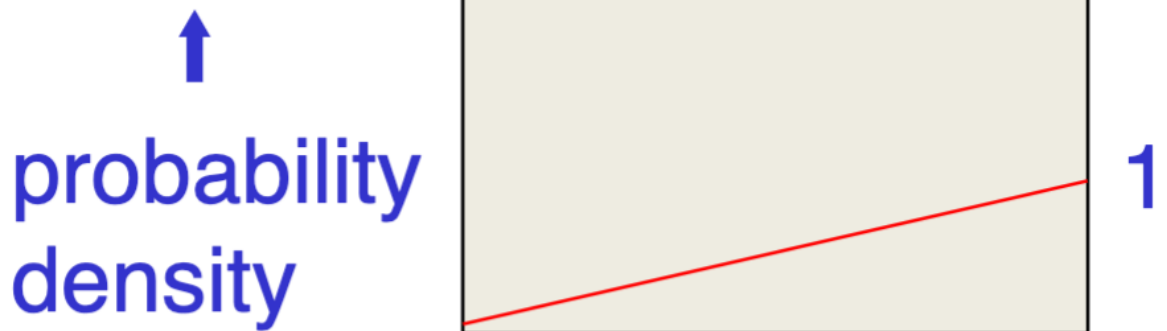
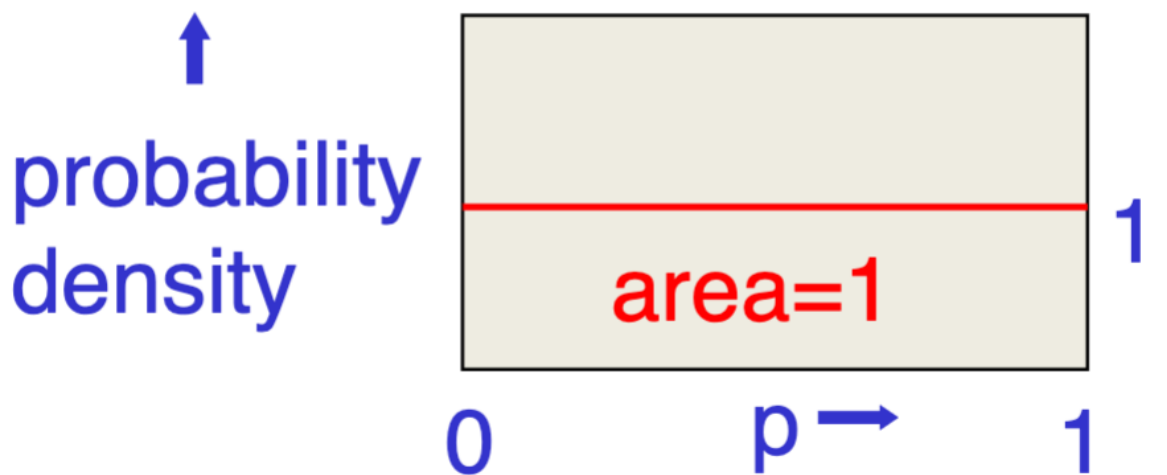
1.6 The Bayesian Perspective

- Rather than taking the most likely estimate of p (i.e. maximum likelihood)...
- Let's build a Bayesian posterior distribution over our belief in p
- The more data we see, the more "peaked" our belief in the correct value of p



1.7 Using a distribution over parameter values

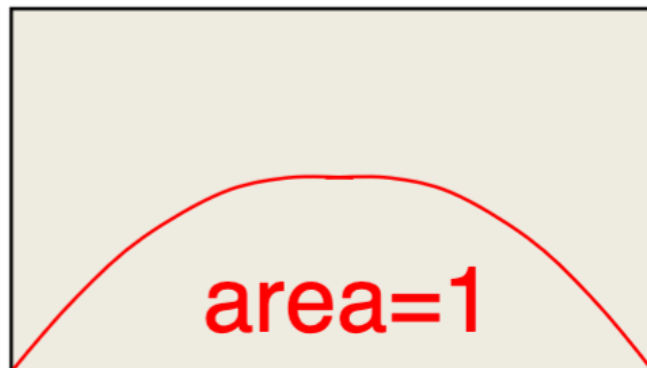
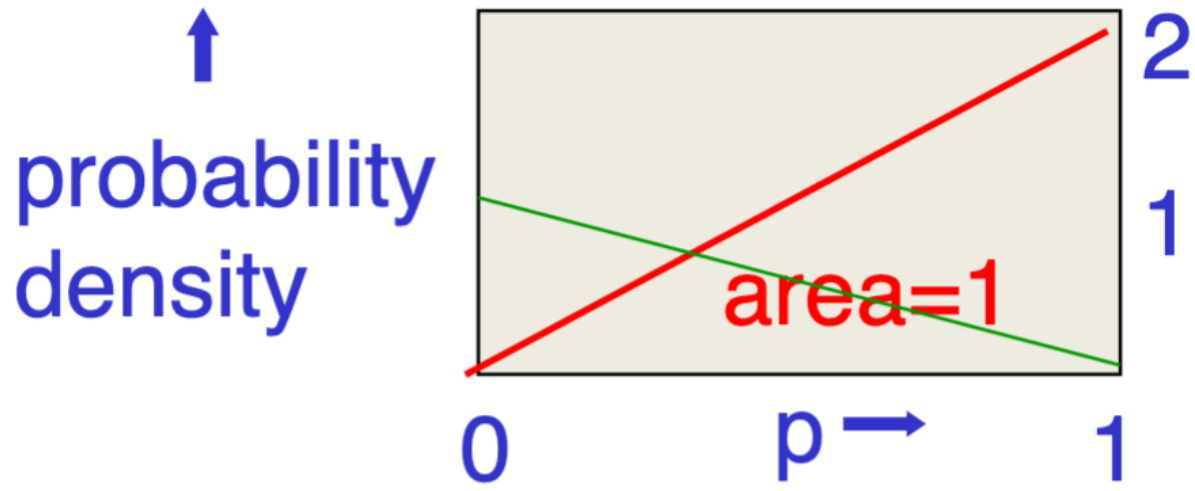
- Start with a prior distribution over p . In this case we used a uniform distribution.
- Multiply the prior probability of each parameter value by the probability of observing a head given that value (likelihood).
- Then scale up all of the probability densities so that their integral comes to 1. This gives the posterior distribution.



1.8 Lets do it again: Suppose we get a tail

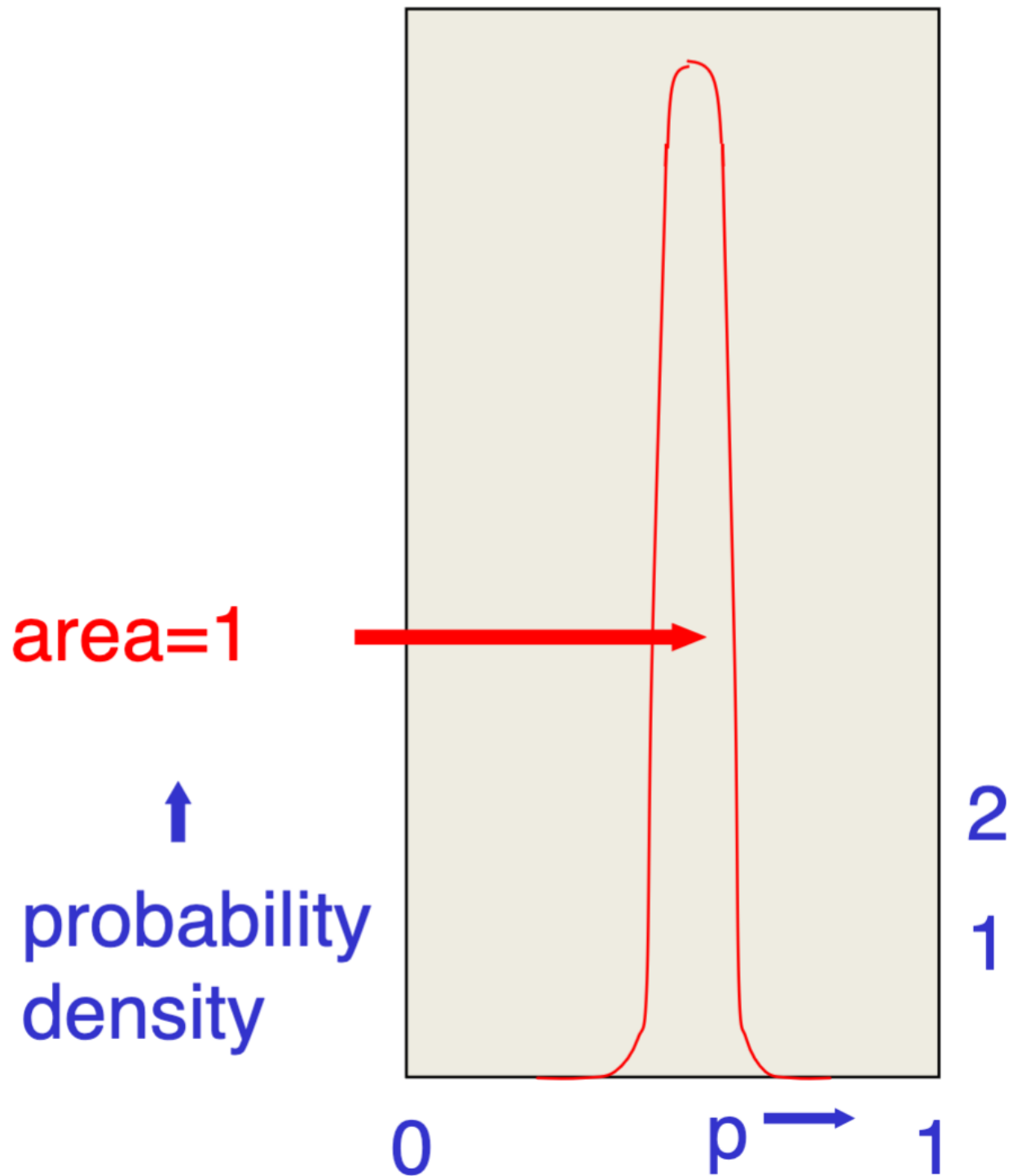
- Start with a prior distribution over p .
- Multiply the prior probability of each parameter value by the probability of observing a tail given that value.

- Then renormalize to get the posterior distribution. Look how sensible it is!



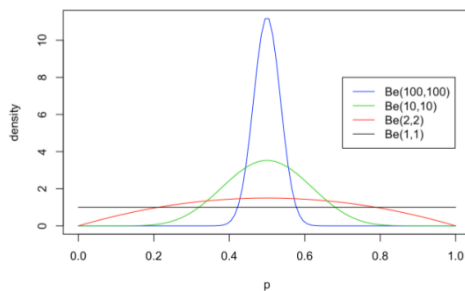
1.9 Lets do it another 98 times

- After 53 heads and 47 tails we get a very sensible posterior distribution that has its peak at 0.53 (assuming a uniform prior)

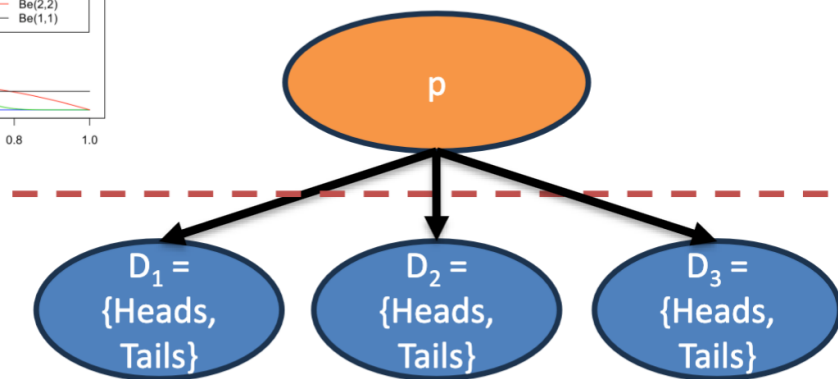


1.10 Generative Graphical Model for Biased Coin

- We have one latent parameter p
 - With a prior distribution (Beta)
- Generates data via a likelihood
 - With a Bernoulli distribution
- Infer posterior distribution
 - Also Beta since this is a conjugate prior-posterior pair



Latent Data Generating Process



likelihood
distribution

Data Observations

1.11 The Bayesian framework

The Bayesian framework assumes that we always have a prior distribution for everything.

- The prior may be very vague.
- When we see some data, we combine our prior distribution with a likelihood term to get a posterior distribution.
- The likelihood term takes into account how probable the observed data is given the parameters of the model.
 - It favors parameter settings that make the data likely.
 - It fights the prior
 - With enough data the likelihood terms always wins.

1.12 Bayes Theorem

joint probability

$$p(D)p(W|D) = p(D, W) = p(W)p(D|W)$$
conditional probability

prior probability of weight vector W

$$p(W|D) = \frac{p(W) \quad \text{Likelihood probability of observed data given W}}{p(D)}$$
posterior probability of weight vector W given training data D

$$\int_W p(W)p(D|W)$$

1.13 Aside: Bayesian Graphical Model Notation

- Recursive Bayesian update:

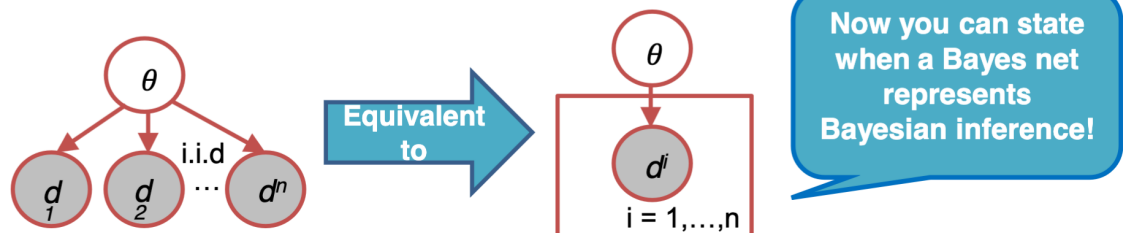
$$P(\vec{\theta}|D^n) \propto P(d^n|\vec{\theta})P(\vec{\theta}|D^{n-1}) \text{ where } D^n = \{d^1, \dots, d^n\}$$

- Recursive Bayesian update unrolled (data observed):

$$P(\vec{\theta}|D^n) \propto P(\vec{\theta}, D^n) = P(\vec{\theta}) \prod_{i=1}^n P(d^i|\vec{\theta})$$

- Inference in graphical models below is Bayesian**

- **Plate notation** invented to represent i.i.d. template:



1.14 A More Complex Generative Process

A More Complex Generative Process

- **Data are just the observables of a generative process**
 - We observe election **votes** of a population
 - We observe Gallup poll **survey** of a population's voting disposition
 - But what process generated both?
- **Example: 2016 US Presidential Election**
 - Gallup Poll Surveys predicted a landslide for Hillary Clinton
 - We know how the election turned out
 - Why was there a discrepancy?
 - Sample bias in surveys
 - Human behavioural biases in revealing their true preferences
- **We care about the (latent) variables that generate the data**
 - We can use understanding of this generative process for prediction
 - Aside: critically connected to modern perspective of Generative AI

