Introduction to Bayesian Data Analysis

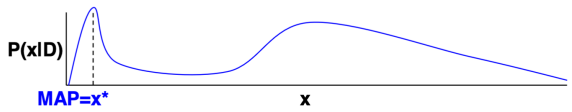
MIE223

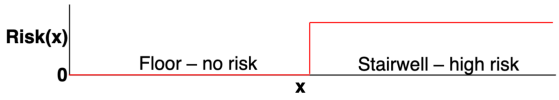
Winter 2025

1 Bayesian Data Analysis

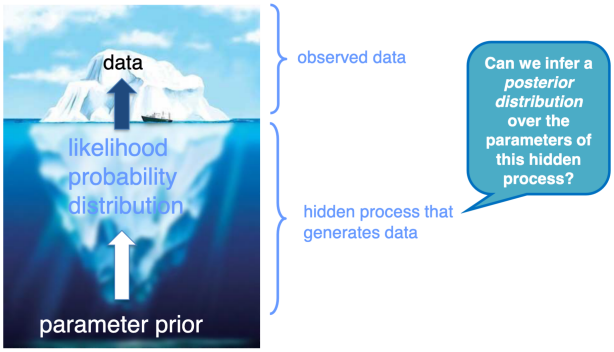
1.1 Bayesian Motivation: Bayesian Decision Theory

- 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 = x Risk(x)p(x—D) ¿ 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: maximum likelihood (frequentist estimation)

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

$$P(D) = p^{53}(1-p)^{47}$$

$$\frac{dP(D)}{dp} = 53p^{52}(1-p)^{47} - 47p^{53}(1-p)^{46}$$

$$= \left(\frac{53}{p} - \frac{47}{1-p}\right)\left[p^{53}(1-p)^{47}\right]$$

$$= 0 \;\; if \;\; p = .53$$

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

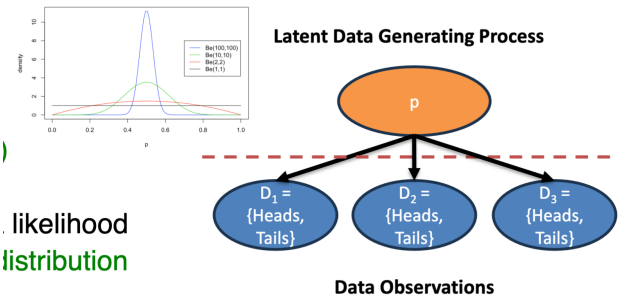
- What if we only tossed the coin once and we got 1 head?
  - Is p=1 a sensible answer?
  - Surely p=0.5 is a much better answer.
- Is it reasonable to give a single answer?
  - If we don't have much data, we are unsure about p.
  - Our computations of probabilities will work much better if we take this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

1.6 The Bayesian Perspective

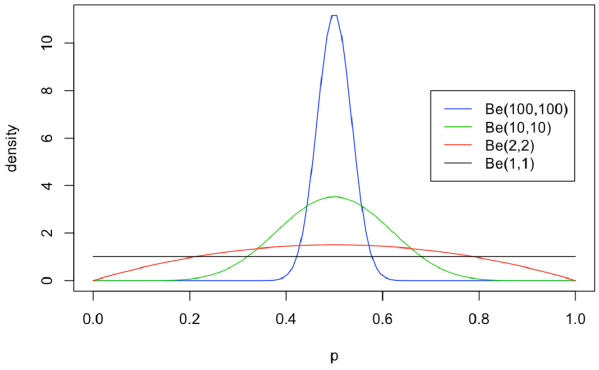
- Rather than taking the most likely estimate of p (i.e. maximum likelihood)...
- Let's build a Bayesian posterior distribution over our belief in p
- The more data we see, the more "peaked" our belief in the correct value of p

1.7 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



How likely is HTH given p? p(1-p)p.



1.8 Using a distribution over parameter values

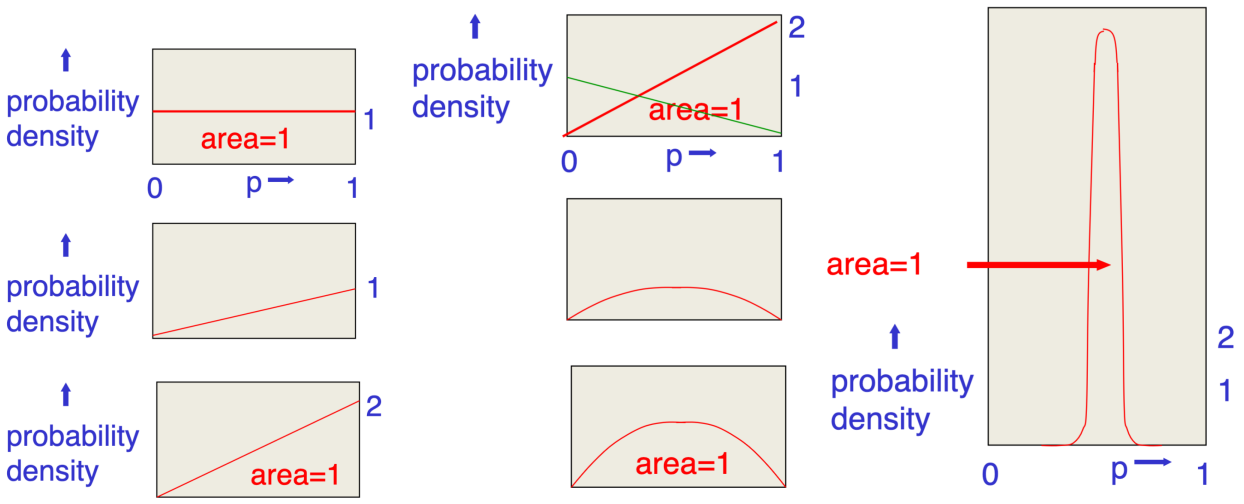
- Start with a prior distribution over p. In this case we used a uniform distribution.
- Multiply the prior probability of each parameter value by the probability of observing a head given that value (likelihood).
- Then scale up all of the probability densities so that their integral comes to 1. This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

1.9 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!

1.10 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)

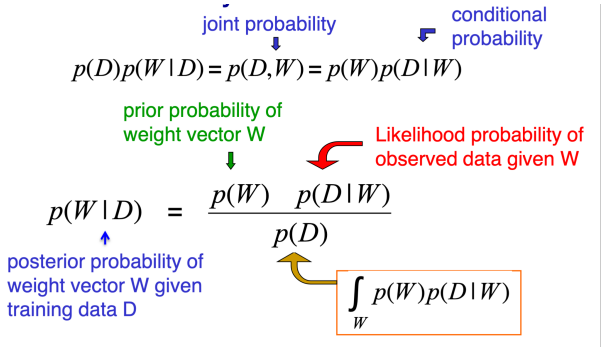


1.11 The Bayesian framework

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



1.13 Aside: Bayesian Graphical Model Notation

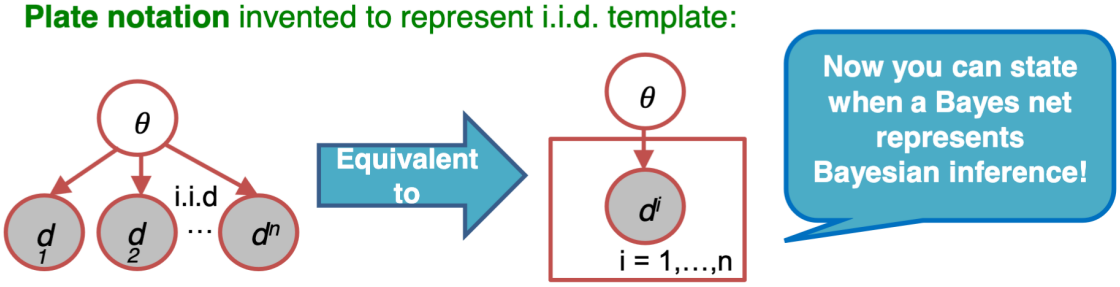
- Recursive Bayesian update:

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

- 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:



bernoulli processes are iid.

1.14 Sanner on the Chalkboard

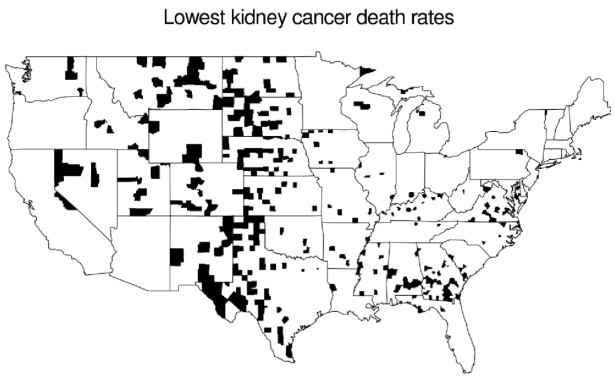
$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \tag{1}$$

It's proportional to P(B—A)P(A) because P(B) is a constant.
posterior is proportional to likelihood times prior.
proportiional to:

$$\prod_{i=1}^{n} P(x_i|\theta)P(\theta) \tag{2}$$

1.15 Why Bayesian Data Analysis? Sparse Data

- Choropleth at right shows countries with lowest 10% death rates for kidney cancer
- Should we live in the Midwest to avoid kidney cancer?
- What went wrong?



Source: Bayesian Data Analysis (Gelman et al, 2013)

1.16 Why Bayesian Data Analysis? Small Data

- Why Bayesian? Small Data. (Gelman, 2005)
  - Sample sizes are never large.
  - If N is too small to get a sufficiently-precise estimate, you need to get more data (or make more assumptions).
  - But once N is "large enough", you can start subdividing the data to learn more (for example, in a public opinion poll, once you have a good estimate for the entire country, you can estimate among men and women, northerners and southerners, different age groups, etc.).
  - N is never enough because if it were "enough" you'd already be on to the next problem for which you need more data.

1.17 A More Complex Generative Process