Computer Science

CSC535: Probabilistic Graphical Models

Bayesian Probability and Inference

Prof. Jason Pacheco

Administrative Items

- HW1 grades / solutions posted
- For homeworks, please submit PDF *as separate file*
  - D2L makes it difficult to grade otherwise
- To generate uniform random integers in the range [1,6]

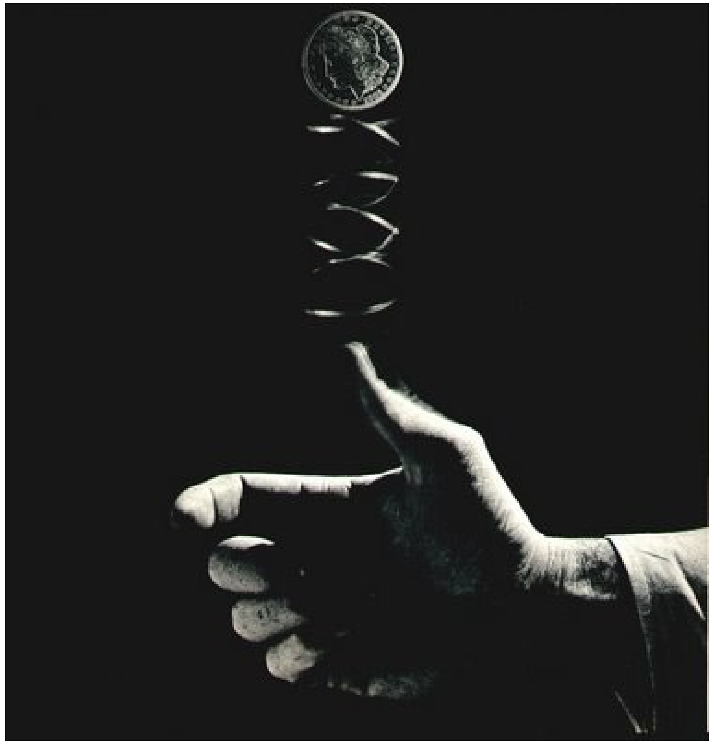
```
numpy.random.randint(1,7)
```

Second argument is (exclusive)

This is the case for most Numpy functions that take intervals

What is Probability?

What does it mean that the probability of heads is ½ ?



Two schools of thought…

Frequentist Perspective
Proportion of successes (heads) in repeated trials (coin tosses)

Bayesian Perspective
Belief of outcomes based on assumptions about nature and the physics of coin flips

Neither is better/worse, but we can compare interpretations…

Frequentist & Bayesian Modeling

We will use the following notation throughout:

θ - Unknown (e.g. coin bias) y - Data

Frequentist

(Conditional Model)

$$p(y;\theta)$$

- θ is a non-random unknown parameter
- $p(y;\theta)$ is the *sampling / data generating distribution*

Bayesian

(Generative Model)

Prior Belief → $p(\theta)p(y \mid \theta)$ ← **Likelihood**

- θ is a random variable (latent)
- Requires specifying $p(\theta)$ the prior belief

Bayes' Rule

Posterior represents all uncertainty after observing data…

prior probability

likelihood function for the parameters

$$p(\theta \mid y) = \frac{p(\theta)p(y \mid \theta)}{p(y)}$$

posterior probability

marginal likelihood
or: evidence
or: partition function
or: normalizer

Bayes' Rule : Marginal Likelihood

$$p(\theta \mid y) = \frac{p(\theta)p(y \mid \theta)}{p(y)} \propto \underbrace{p(\theta)p(y \mid \theta)}_{\text{Often know this (the model)}}$$

Often hard to calculate

Marginal likelihood integrates (marginalizes) over unknown θ :

$$p(y) = \int p(\theta)p(y \mid \theta)\, d\theta$$

Marginal likelihood is less problematic in discrete models (not always)

This integral often lacks a closed form and cannot be computed…

Aside : Proportionality

Recall PMF / PDF must sum / integrate to 1,

PMF

$$\sum_x p(x) = 1$$

PDF

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

May only know distribution constant that does not depend on RV x,

$$\int \widetilde{p}(x)\,dx = \mathcal{Z} \qquad \text{so} \qquad p(x) \propto \widetilde{p}(x)$$

Properly normalized distribution by dividing our <u>normalization constant:</u>

$$\int p(x)\,dx = \int \frac{1}{\mathcal{Z}}\widetilde{p}(x)\,dx = \frac{1}{\int \widetilde{p}(x)\,dx}\int \widetilde{p}(x)\,dx = 1$$

Aside : Proportionality

Example Let X be a Bernoulli RV (coinflip) with probabilities *proportional to:*

$$\widetilde{p}(X=0) = 0.5 \qquad \widetilde{p}(X=1) = 1.5$$

Greater than 1, but It is an *unnormalized* probability

Compute normalization constant,

$$\mathcal{Z} = \widetilde{p}(X=0) + \widetilde{p}(X=1) = 2.0$$

Normalize probability distribution,

$$p(X) = \frac{1}{\mathcal{Z}}\widetilde{p}(X) = \begin{pmatrix} 1/4 \\ 3/4 \end{pmatrix}$$

Sums to 1

Bayesian Inference Example

About 29% of American adults have high blood pressure (BP). Home test has 30% false positive rate and no false negative error.

A recent home test states that you have high BP. Should you start medication?

An Assessment of the Accuracy of Home Blood Pressure Monitors When Used in Device Owners

Jennifer S. Ringrose,[1] Gina Polley,[1] Donna McLean,[2–4] Ann Thompson,[1,5] Fraulein Morales,[1] and Raj Padwal[1,4,6]

Bayesian Inference Example

About 29% of American adults have high blood pressure (BP). Home test has 30% false positive rate and no false negative error.

- Latent quantity of interest is hypertension: $\theta \in \{true, false\}$
- Measurement of hypertension: $y \in \{true, false\}$
- Prior: $p(\theta = true) = 0.29$
- Likelihood: $p(y = true \mid \theta = false) = 0.30$

$$p(y = true \mid \theta = true) = 1.00$$

Bayesian Inference Example

About 29% of American adults have high blood pressure (BP). Home test has 30% false positive rate and no false negative error.

Suppose we get a positive measurement, then posterior is:

$$p(\theta = true \mid y = true) = \frac{p(\theta = true)p(y = true \mid \theta = true)}{p(y = true)}$$

$$= \frac{0.29 * 1.00}{0.29 * 1.00 + 0.71 * 0.30} \approx 0.58$$

What conclusions can be drawn from this calculation?

Bayesian Updating

Suppose we plan to take another test…

Question What is our belief about blood pressure status *before* the second test?

(a) Posterior: $p(\theta = true \mid y_1 = true)$ ←

(b) Likelihood: $p(y_1 = true \mid \theta = true)$

(c) Marginal Likelihood: $p(y_1 = true)$

Bayesian Updating

Suppose we plan to take another test…

Question What is the probability that we get *true* on the second test if we have high blood pressure?

(a) Posterior: $p(\theta = true \mid y_1 = true, y_2 = true)$

(b) Likelihood: $p(y_2 = true \mid \theta = true)$ ←

(c) Marginal Likelihood: $p(y_2 = true)$

Why not: $p(y_2 = true \mid \theta = true, y_1 = true)$

Bayesian Updating

Suppose we plan to take another test…

Question What is the probability that we get *true* on the second test if we have high blood pressure?

(a) Posterior: $p(\theta = true \mid y_1 = true, y_2 = true)$

(b) Likelihood: $p(y_2 = true \mid \theta = true)$

(c) Marginal Likelihood: $p(y_2 = true)$

Because $y_1 \perp y_2 \mid \theta$ *so* $p(y_2 \mid \theta, y_1) = p(y_2 \mid \theta)$

Bayesian Updating

Suppose we receive another positive test $y_2 = true$…

Posterior belief given *both* tests is then,

$$p(\theta = true \mid y_1 = true, y_2 = true) =$$

$$= \frac{p(\theta = true \mid y_1 = true)p(y_2 = true \mid \theta)}{p(y_2 = true \mid y_1 = true)}$$

Probability of getting two positive tests regardless of BP status

$$\propto \underbrace{p(\theta = true \mid y_1 = true)}_{\text{Inference from first test}}\underbrace{p(y_2 = true \mid \theta = true)}_{\text{Likelihood of positive test}}$$

Bayesian Updating

Consider two *conditionally independent* observations X_1 and X_2, their joint distribution is:

Probability chain rule

$$p(\theta, X_1, X_2) = p(\theta)p(X_1 \mid \theta)p(X_2 \mid \theta) = p(\theta \mid X_1)p(X_1)p(X_2 \mid \theta)$$

So, conditioned on X_1:

Update prior belief after seeing X_1

$$p(\theta, X_2 \mid X_1) = p(\theta \mid X_1)p(X_2 \mid \theta)$$

This is proportional to the **full posterior** by Bayes' rule:

$$p(\theta \mid X_1, X_2) \propto p(\theta \mid X_1)p(X_2 \mid \theta)$$

Normalizer is $p(X_2 \mid X_1)$

Step 1: Do inference after seeing X1

Step 2: Update posterior by multiplying likelihood of X2

Bayesian Updating

Given conditionally independent $X_1, \ldots, X_N$ posterior belief is,

$$p(\theta \mid X_1, \ldots, X_N)$$

Receive N+1th observation X_{N+1} and update posterior,

$$p(\theta \mid X_1, \ldots, X_{N+1}) \propto p(\theta \mid X_1, \ldots, X_N) p(X_{N+1} \mid \theta)$$

Belief after seeing N+1th observation

Belief before seeing N+1th observation

Belief about N+1th observation

Updates are more complicated if observations are dependent…

Frequentist vs. Bayesian Inference

We have data $X_1, \ldots, X_N$ and want to infer unknown parameter θ

Frequentist Inference

The data *uniquely determines* θ, *e.g.* by the likelihood:

Not a distribution on parameter $$p(X_1, \ldots, X_N; \theta)$$ **How well it explains the data**

Bayesian Inference

The data *updates our belief* about θ, which is random:

$$p(\theta \mid X_1, \ldots, X_N) \propto p(\theta \mid X_1, \ldots, X_{N-1}) p(X_N \mid \theta)$$

Our belief changes with more data

Minimum Mean Squared Error (MMSE)

Posterior mean minimizes squared error,

$$\hat{\theta}^{\mathrm{MMSE}} = \arg\min \mathbb{E}[(\hat{\theta} - \theta)^2 \mid y] = E[\theta \mid y]$$

- Minimizes error <u>conditioned on observed data</u>

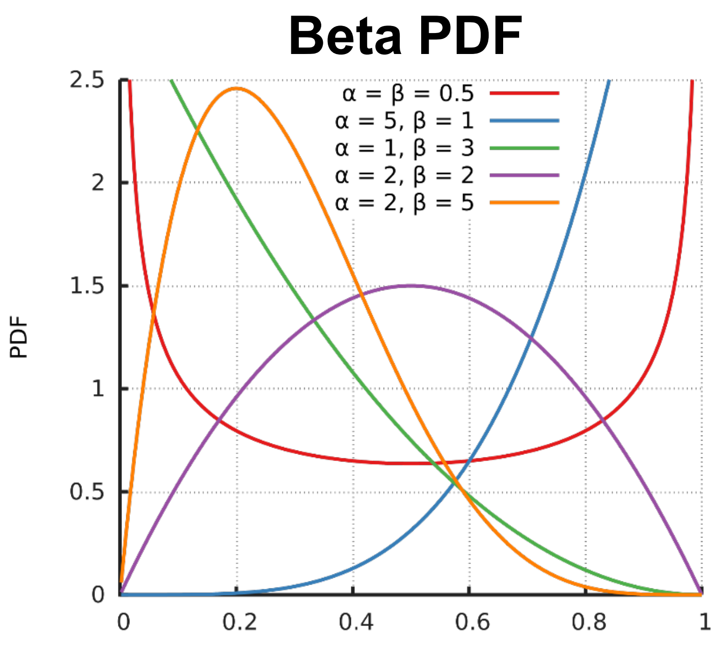
- MMSE is an **unbiased estimator**

- MMSE is **asymptotically unbiased** and **asymptotically normal**,

$$\sqrt{N}(\hat{\theta}^{\mathrm{MMSE}} - \theta) \to \mathcal{N}(0, \sigma^2)$$

Example: Beta-Bernoulli MMSE

Let $Y_1, \ldots, Y_N \sim \text{Bernoulli}(\pi)$ and $\pi \sim \text{Beta}(\alpha, \beta)$.



- Beta is a distribution on probabilities $\pi \in [0, 1]$
- *Shape parameters* α and β with mean,

$$\mathbf{E}[\theta] = \frac{\alpha}{\alpha+\beta}$$

- Beta-Bernoulli has Beta posterior distribution,

$$p(\pi \mid X_1^N) = \text{Beta}(\alpha + \textbf{number of heads}, \; \beta + \textbf{number of tails})$$

MMSE given by posterior mean,

Prior belief (pseudo-heads)

$$\hat{\pi}^{\text{MMSE}} = \frac{\alpha + \textbf{number of heads}}{\alpha + \beta + N}$$

Q What happens to MMSE when we have limited data?

Q What happens to MMSE when we have a lot of data?

Bayes Estimators

Minimizes expected loss function,

$$\hat{\theta} = \arg\min_{\hat{\theta}} \mathbf{E}\left[L(\theta, \hat{\theta}) \mid y\right]$$

Expected loss referred to as *Bayes risk.*

MMSE minimizes squared-error loss $L(\theta, \hat{\theta}) = (\theta - \hat{\theta})^2$

Minimum absolute error (MAE) is posterior *median,*

$$\arg\min \mathbb{E}[|\hat{\theta} - \theta| \mid y] = \text{median}(\theta \mid y)$$

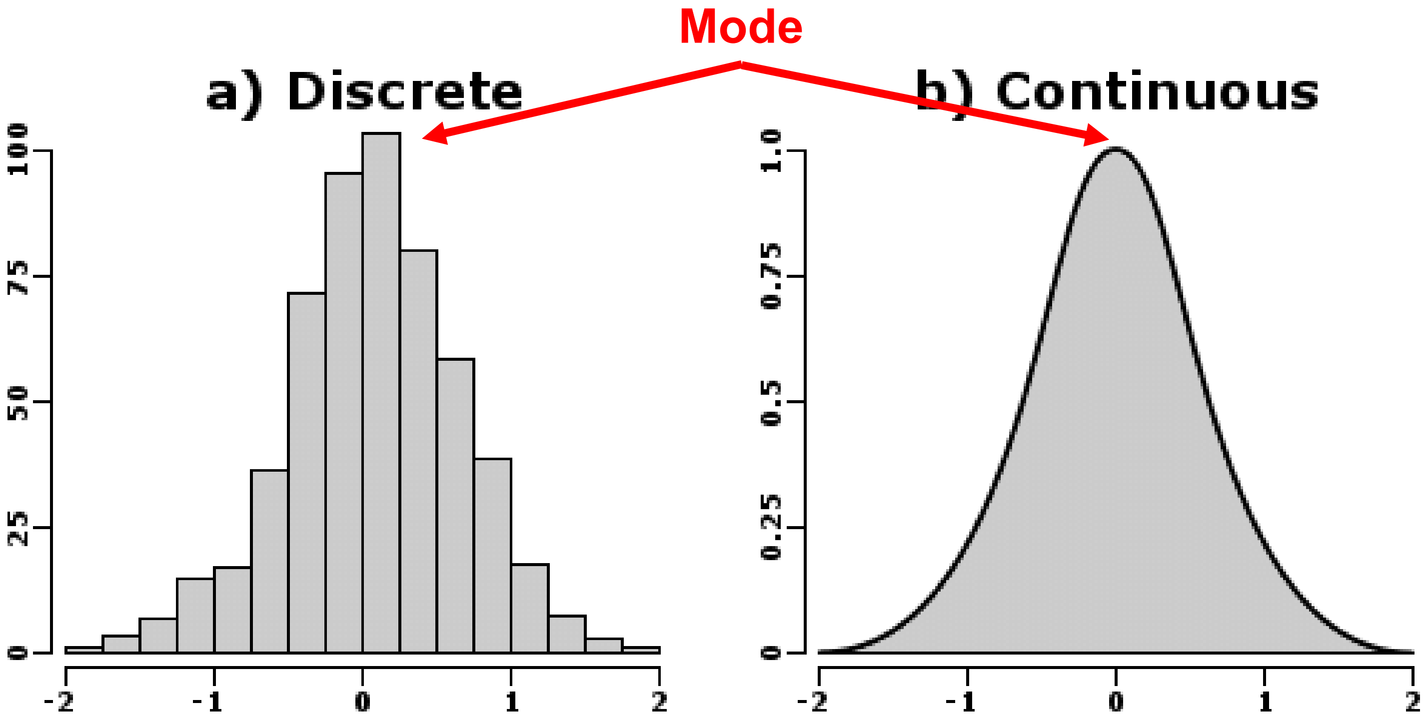
Note: Same answer for linear function: $L(\theta, \hat{\theta}) = c|\hat{\theta} - \theta|$

Maximum a Posteriori (MAP)

Very common to produce maximum probability estimates,

$$\hat{\theta}^{\mathrm{MAP}} = \arg max \; p(\theta \mid y)$$

MAP is the ***mode*** *(highest probability outcome) of the posterior*



Maximum a Posteriori (MAP)

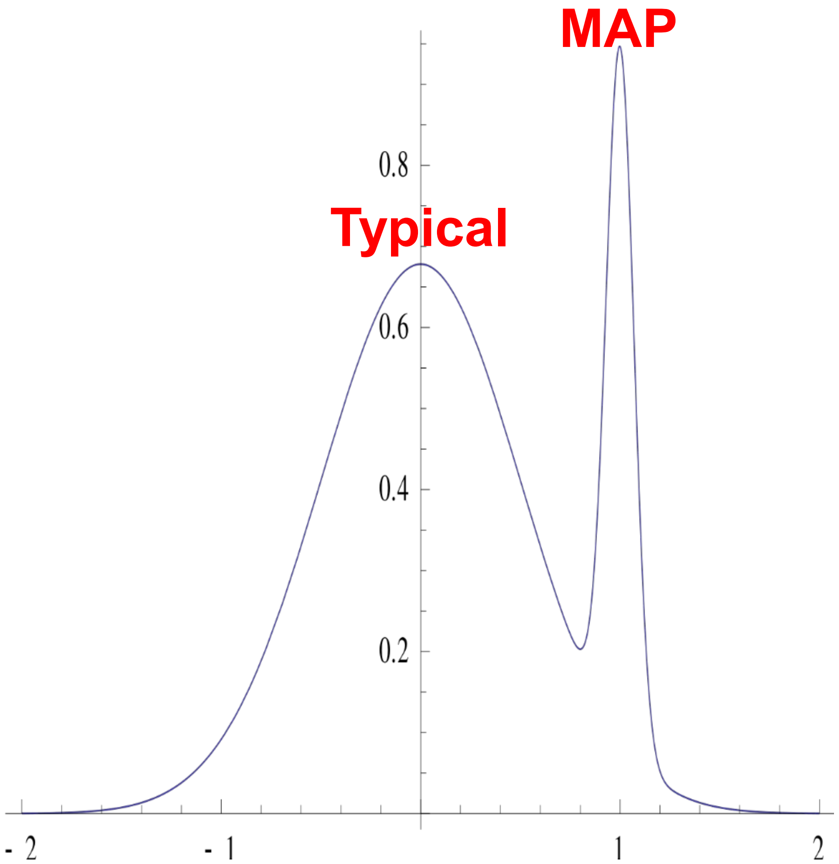
MAP (mode) may not be representative of typical outcomes

Also, not a Bayes estimator (unless discrete),

$$\lim_{c \to 0} L(\theta, \hat{\theta}) = \begin{cases} 0, & \text{if } |\hat{\theta} - \theta| < c \\ 1, & \text{otherwise} \end{cases}$$

Degenerate loss function

Despite its issues, MAP is frequently used in “Bayesian” inference and estimation



Posterior PDF

Example: Beta-Bernoulli MAP

Let $X_1, \ldots, X_N \sim \mathrm{Bernoulli}(\pi)$ and $\pi \sim \mathrm{Beta}(\alpha, \beta)$ then posterior is,

$$p(\pi \mid X_1^N) = \mathrm{Beta}(\alpha + \underbrace{\textbf{number of heads}}_{\mathbf{N_H}},\ \beta + \textbf{number of tails})$$

Highest probability (mode) of Beta given by,

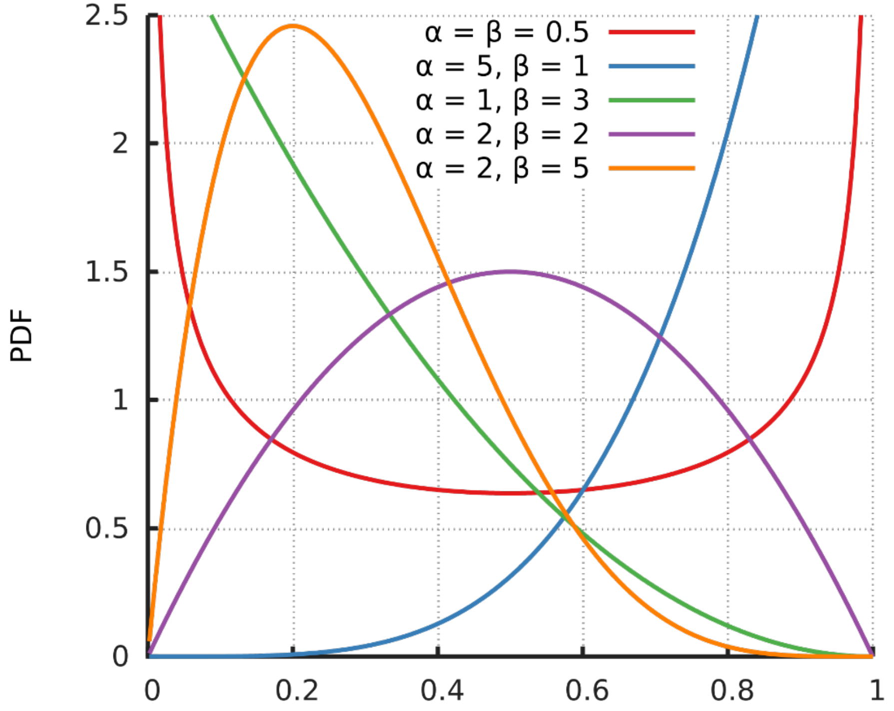
Take derivative, set to zero, solve.

$$\hat{\pi}^{\mathrm{MAP}} = \frac{\alpha + N_H - 1}{\alpha + \beta + N - 2}$$

Beta distribution is not always convex!

- MAP is any value for $\alpha = \beta = 1$
- Two modes (bimodal) for $\alpha, \beta < 1$

Beta Posterior PDF



Maximum a Posteriori (MAP)

Equivalent to maximizing joint probability,

Constant

$$\arg\max_{\theta} p(\theta \mid y) = \arg\max_{\theta} \frac{p(\theta, y)}{p(y)} = \arg\max_{\theta} p(\theta, y)$$

For iid $y_1, \ldots, y_N$ solve in log-domain (like *maximum likelihood est.*),

$$\hat{\theta}^{\mathrm{MAP}} = \arg\max_{\theta} \log p(\theta, y_1, \ldots, y_N) = \underbrace{\sum_i \log p(y_i \mid \theta)}_{\substack{\textbf{Log-Likelihood} \\ \textbf{(how well it fits data)}}} + \underbrace{\log p(\theta)}_{\substack{\textbf{Log-Prior} \\ \textbf{(how well it} \\ \textbf{agrees with prior)}}}$$

Intuition *MAP is like MLE but with a "penalty" term (log-prior)*

Bayes' Rule : Reminder

prior probability

likelihood function for the parameters

$$p(\theta \mid D) = \frac{p(\theta)p(D \mid \theta)}{p(D)}$$

posterior probability

normalizer, often is not of interest

Bayes' Rule : Reminder

prior probability

likelihood function
for the parameters

$$p(\theta \mid D) \propto p(\theta)p(D \mid \theta)$$

posterior probability

Posterior is **proportional**
to the joint

Bayes' Rule : Reminder

prior probability

likelihood function for the parameters

$$q(\theta \mid D) \propto p(\theta)\omega(D \mid \theta)$$

posterior probability

In general, distributions are different functions

Bayes' Rule : Reminder

Prior and **likelihood** chosen for model

$$q(\theta \mid D) \propto p(\theta)\omega(D \mid \theta)$$

Posterior determined by algebra

In general, distributions are different functions

Conjugate Pairs

For some special models the posterior takes a simple form

$$p(\theta \mid D) \propto p(\theta)\omega(D \mid \theta)$$

Prior and posterior are the same distribution (with different parameters)

We have already seen one example, the Beta-Bernoulli conjugate pair:

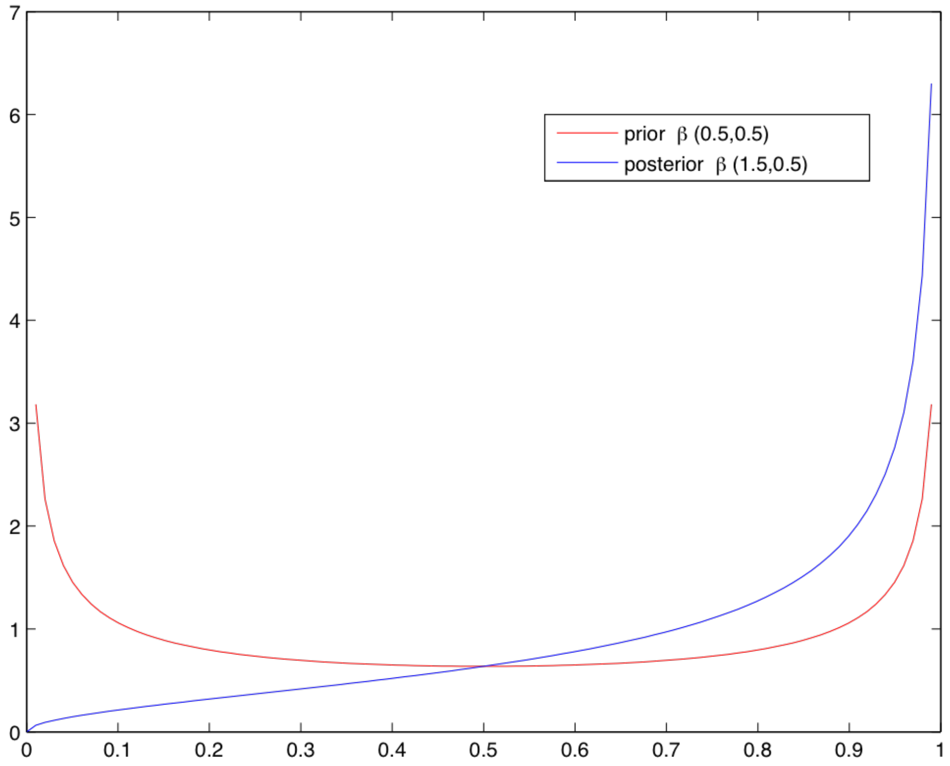
$$\mathrm{Beta}(\theta \mid \alpha + \textbf{num.-heads}, \beta + \textbf{num.-tails}) \propto \mathrm{Beta}(\theta \mid \alpha, \beta) \prod_i \mathrm{Bernoulli}(x_i \mid \theta)$$

Same PDF

Example: Beta-Bernoulli

After a single coinflip of heads (x=1) the posterior is…

$$\mathrm{Beta}(\theta \mid \alpha + x, \beta + 1 - x)$$



The prior (red) is a fair coin,

$$\mathrm{Beta}(\theta \mid \alpha = 0.5, \beta = 0.5)$$

After observing one head, the posterior (blue) concentrates on heads,

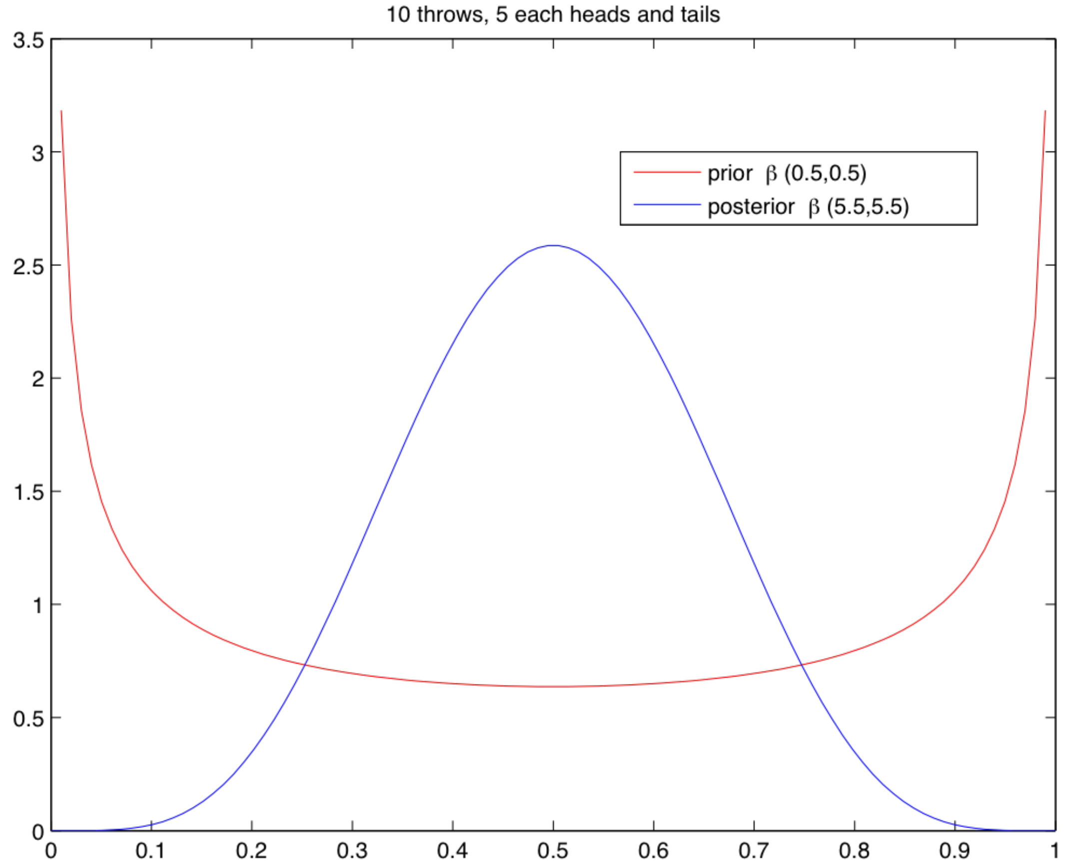
$$\mathrm{Beta}(\theta \mid 1.5, 0.5)$$

What do you expect if we flip N=10 times with 5 heads and 5 tails?

Example: Beta-Bernoulli

After a N=10 flips (5 heads, 5 tails) we have…

$$\mathrm{Beta}(\theta \mid \alpha + 5, \beta + 5) = \mathrm{Beta}(\theta \mid 5.5, 5.5)$$



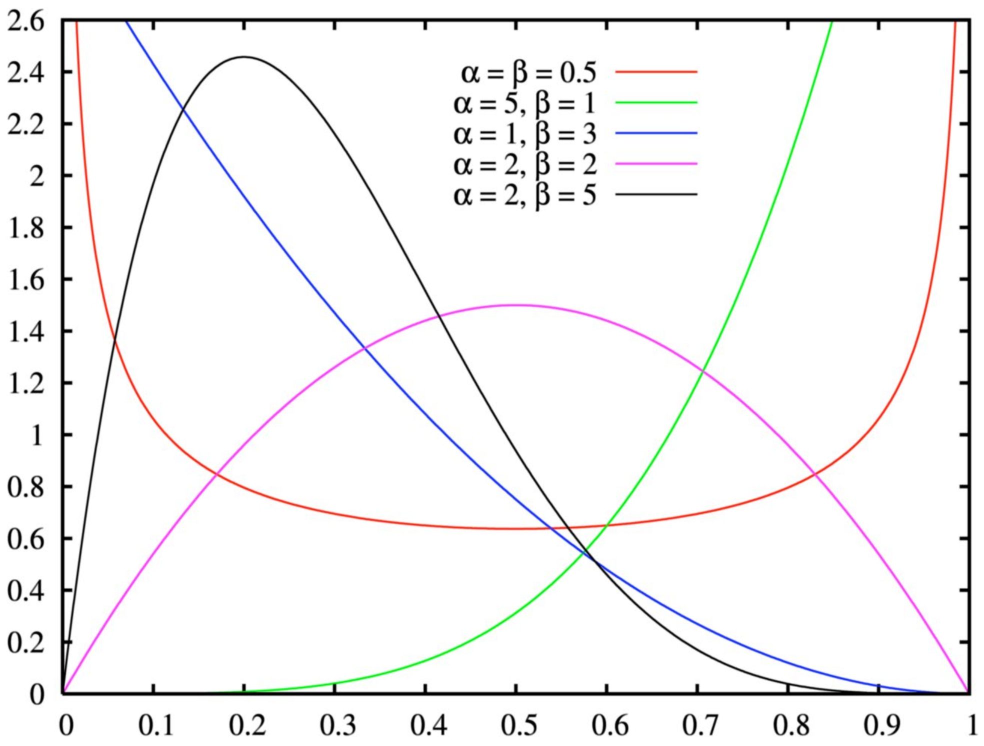
Posterior concentrates on fair coin $\theta = 0.5$

Example: Beta-Bernoulli

Bernoulli *A.k.a. the* ***coinflip*** *distribution on* <u>*binary*</u> *RVs* $X \in \{0, 1\}$

$$\text{Bernoulli}(X \mid \theta) = \theta^X (1 - \theta)^{(1-X)}$$

Beta *distribution on* $\theta \in (0, 1)$ *with* $\alpha, \beta > 0$ *has PDF,*



$$\text{Beta}(\theta \mid \alpha, \beta) \propto \theta^{\alpha - 1} (1 - \theta)^{\beta - 1}$$

For N coinflips $x_1, \ldots, x_N$ the posterior is,

$$\text{Beta}(\theta \mid \alpha + \sum_i x_i, \beta + N - \sum_i x_i)$$

Example: Beta-Bernoulli

$$\mathrm{Beta}(\theta \mid \alpha, \beta) \prod_{i=1}^{N} \mathrm{Bernoulli}(x_i \mid \theta) \propto$$

$$\propto \theta^{\alpha-1}(1-\theta)^{\beta-1} \prod_i \theta^{x_i}(1-\theta)^{1-x_i}$$

$$= \theta^{\alpha-1}(1-\theta)^{\beta-1}\theta^{\sum_i x_i}(1-\theta)^{\sum_i (1-x_i)}$$

$$= \theta^{\alpha-1}(1-\theta)^{\beta-1}\theta^{\sum_i x_i}(1-\theta)^{\left(N-\sum_i x_i\right)}$$

$$= \theta^{\alpha-1+\sum_i x_i}(1-\theta)^{\beta-1+N-\sum_i x_i}$$

$$\propto \mathrm{Beta}(\theta \mid \alpha + \sum_i x_i, \beta + N - \sum_i x_i)$$

Other Conjugate Pairs

Likelihood	Model Parameters	Conjugate Prior
Normal	Mean	Normal
Normal	Mean / Variance	Normal-Inv-Gamma
Multivariate Normal	Mean / Variance	Normal-Inv-Wishart
Multinomial	Probability vector	Dirichlet
Gamma	Rate	Gamma
Poisson	Rate	Gamma
Exponential	Rate	Gamma

Wikipedia has a nice list of standard conjugate forms…

https://en.wikipedia.org/wiki/Conjugate_prior

Priors in AI / ML / Data Science

- Priors are often used as *regularizers* (promote smoothing)
  - Reduces overfitting as random noise is not smooth
  - Often regularizers can be of simple form, even conjugate
- Priors often house sophisticated domain knowledge
  - Possibly from earlier encounters with data
  - Possibly problem constraints (e.g. θ must be nonnegative)
  - World knowledge is complex, so good priors are often complex and **not conjugate**

Choosing a Prior

- Conjugate priors can keep posteriors in closed form
  - This can speed up our codes (a lot!)
- The conjugate priors for standard distributions are fairly expressive
  - Often they can serve the purpose
- They are cool (better than doing nothing or the wrong thing)
- But they require that the likelihood is of a standard form
  - This is often a lot to hope for!
- Simply expressed functions may not be able to encode what you know
  - Constraints, non-local relationships

Prediction

Can make predictions of unobserved $\tilde{y}$ before seeing any data,

$$p(\widetilde{y}) = \sum_k p(\theta = k) p(\widetilde{y} \mid \theta = k)$$

Similar calculation to marginal likelihood

This is the ***prior predictive*** *distribution*

For continuous parameters sum turns into integral,

$$p(\tilde{y}) = \int p(\theta) p(\tilde{y} \mid \theta) \, d\theta$$

This is a prediction based on ***no observed data***

Prediction

When we observe y we can predict future observations $\tilde{y}$,

$$p(\widetilde{y}) = \sum_k \underbrace{p(\theta = k \mid y)}_{\textbf{This is now the posterior}} p(\widetilde{\ } | \theta = k)$$

This is the ***posterior predictive*** *distribution*

Again, for continuous parameters sum turns into integral,

$$p(\tilde{y} \mid y) = \int p(\theta \mid y) p(\tilde{y} \mid \theta) \, d\theta$$

Prediction Example

About 29% of American adults have high blood pressure (BP). Home test has 30% false positive rate and no false negative error.

What is the likelihood of *another* positive measurement?

$$p(\tilde{y} = true \mid y = true) = \sum_{\theta \in \{true, false\}} p(\theta \mid y = true) p(\tilde{y} = true \mid \theta)$$

$$= 0.42 * 0.30 + 0.58 * 1.00 \approx 0.71$$

What conclusions can be drawn from this calculation?

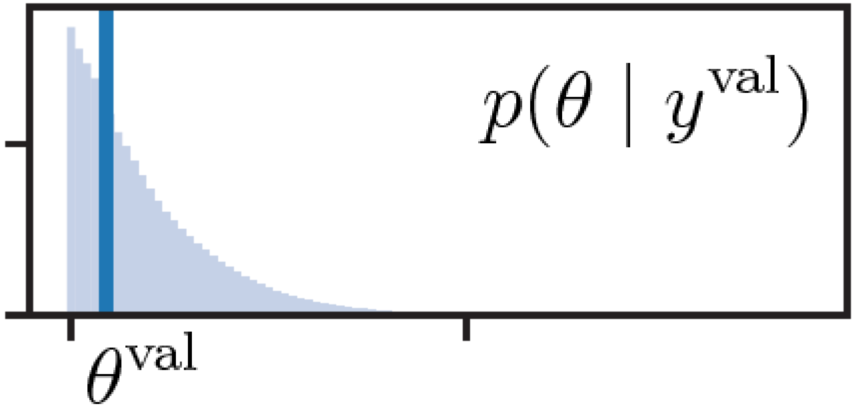
Model Validation

How do we know if the model $p(\theta, y)$ *is* ***good***?

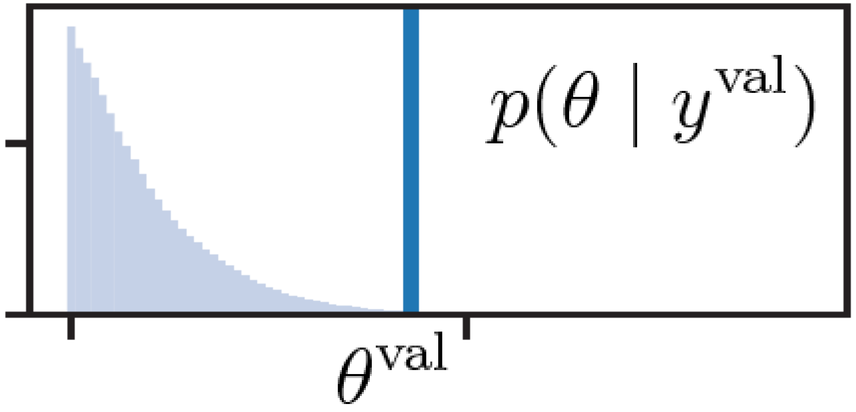
Supervised Learning

Validation set $\{(\theta^{\text{val}}, y^{\text{val}})\}$ consists of known θ^{val}. Are true values typically preferred under the posterior?

Good (maybe lucky)



Not Good (maybe unlucky)



Repeat trials over validation set for more certainty

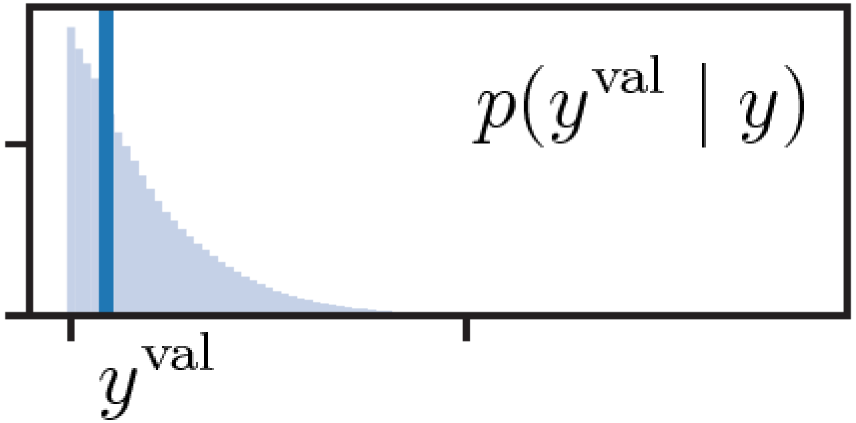
Model Validation

How do we know if the model $p(\theta, y)$ *is* ***good***?

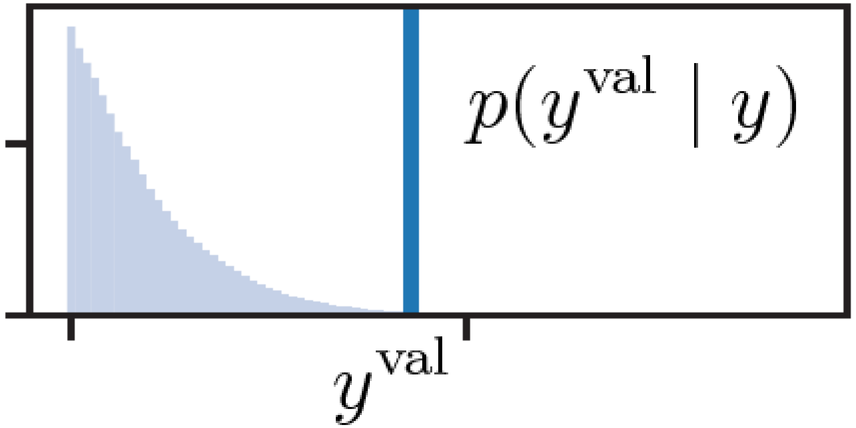
Unsupervised Learning

Validation set $\{y^{\text{val}}\}$ only contains observable data. Check validation data against posterior-predictive distribution.

Good (maybe lucky)



Not Good (maybe unlucky)



Repeat trials over validation set for more certainty

Likelihood and Odds Ratios

Which parameter value θ_1 or θ_2 is more likely to have generated the observed data y?

The **posterior odds ratio** is:

$$\frac{p(\theta_1 \mid y)}{p(\theta_2 \mid y)} = \underbrace{\frac{p(\theta_1)}{p(\theta_2)}}_{\text{Prior Odds Ratio}} \underbrace{\frac{p(y \mid \theta_1)}{p(y \mid \theta_2)}}_{\text{Likelihood Ratio}} \cancel{\frac{p(y)}{p(y)}}$$

Observe: the marginal likelihood $p(y)$ cancels!

Posterior Summarization

Ideally we would report the <u>full posterior distribution</u> as the result of inference…but this is not always possible

Summary of Posterior Location:

Point estimates: mean (MMSE), mode, median (min. absolute error)

Summary of Posterior Uncertainty:

Credible intervals / regions, posterior entropy, variance

Bayesian analysis should report uncertainty when possible

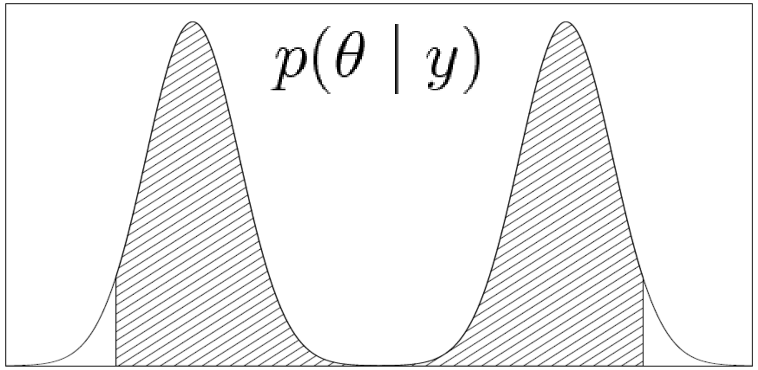
Credible Interval

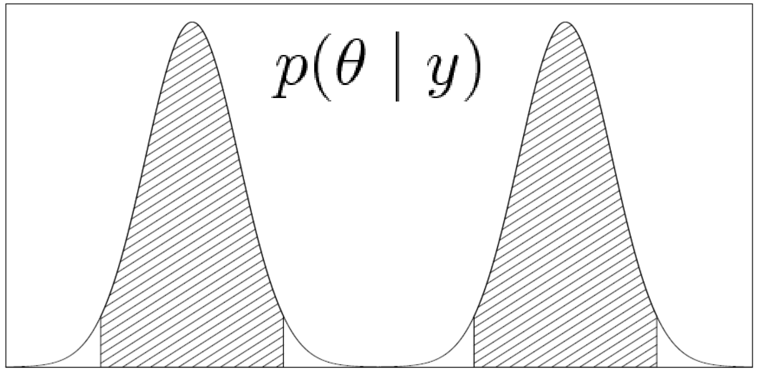
Def. For parameter $0 < \alpha < 1$ the $100(1-\alpha)\%$ credible interval $(L(y), U(y))$ satisfies,

$$p(L(y) < \theta < U(y) \mid y) = \int_{L(y)}^{U(y)} p(\theta \mid y) = 1 - \alpha$$

Interval containing fixed percentage of posterior probability density.

Note: This is <u>not unique</u> -- consider the 95% intervals below:





[Source: Gelman et al., “Bayesian Data Analysis”]

Frequentist Inference

Example: Suppose we observe the outcome of N coin flips.
$y = \{y_1, \ldots, y_N\}$. What is the probability of heads θ (coin bias)?

- Coin bias θ is not random (e.g. there is some *true* value)
- Uncertainty reported as confidence interval (typically 95%)

Correct Interpretation: On repeated trials of N coin flips θ will fall inside the confidence interval 95% of the time (in the limit)

- Inferences are valid for multiple trials, **never on single trials**

Wrong Interpretation: For *this trial* there is a 95% chance θ falls in the confidence interval

Bayesian Inference

Posterior distribution is complete representation of uncertainty

Prior Belief

Likelihood

$$p(\theta \mid y) = \frac{p(\theta)p(y \mid \theta)}{p(y)}$$

Marginal Likelihood
(more on this later)

- Must specify a <u>prior belief</u> $p(\theta)$ about coin bias
- Coin bias θ is a <u>random quantity</u>
- Interval $p(l(y) < \theta < u(y) \mid y) = 0.95$ can be reported in lieu of full posterior, and takes intuitive interpretation for a <u>single trial</u>

Interval Interpretation: For <u>this experiment</u> there is a 95% chance that θ lies in the interval

Summary

- Bayesian statistics interprets probability differently than classical stats
  - Frequentist: Probability → Long run odds in repeated trials
  - Bayesian: Probability → Belief of outcome that captures all uncertainty

- Bayesian models treat unknown parameter as random, with a prior

- Bayesian inference via the *posterior distribution* using Bayes' rule

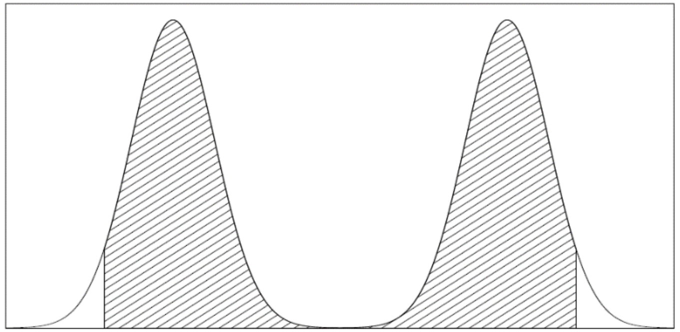
$$p(\theta \mid y) = \frac{p(\theta)p(y \mid \theta)}{p(y)}$$

- Bayesian estimators minimize expected risk (e.g. MMSE)

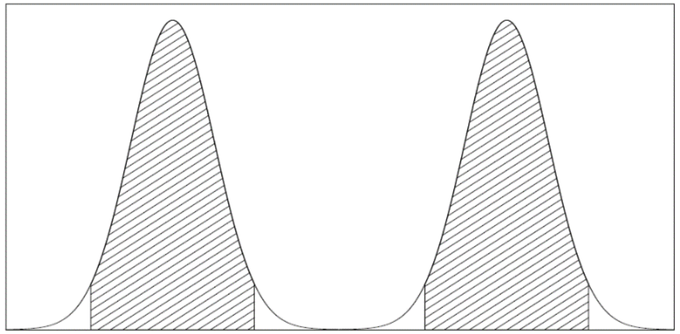
- Maximum a posteriori (MAP) estimate maximizes posterior probability

Summary

- Conjugate prior-posterior pairs ensure closed-form posterior inference
- Posterior uncertainty can be characterized by credible intervals



Not necessarily unique



- Selecting models can be done via posterior odds ratio

$$\frac{p(\theta_1 \mid y)}{p(\theta_2 \mid y)} = \frac{p(\theta_1)}{p(\theta_2)} \frac{p(y \mid \theta_1)}{p(y \mid \theta_2)} \cancel{\frac{p(y)}{p(y)}}$$

- Parameter can be marginalized out via prior/posterior predictive dist’n