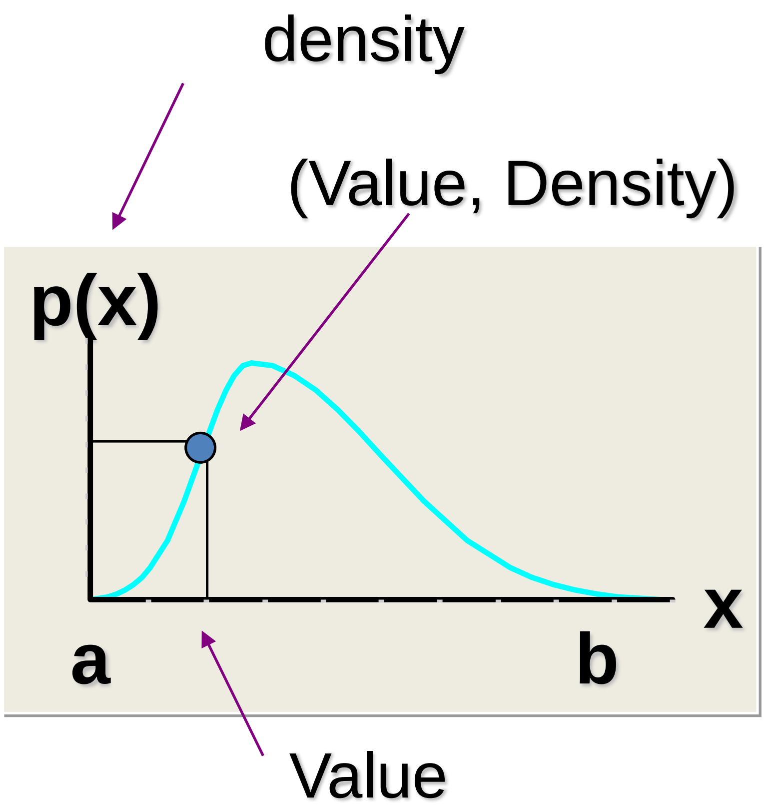
Continuous Probability Distributions

Suresh Manandhar
suresh@cs.york.ac.uk

Continuous Probability Density Function (pdf)

- Shows all values of x in the given interval $[a, b]$, the **density** $p(x)$
- $p(x)$ is a ***probability density function*** (**pdf**)
- Since probabilities need to sum to 1, the corresponding condition is:

$$F(-\infty \leq X \leq \infty) = \int_{-\infty}^{\infty} p(x)dx = 1$$

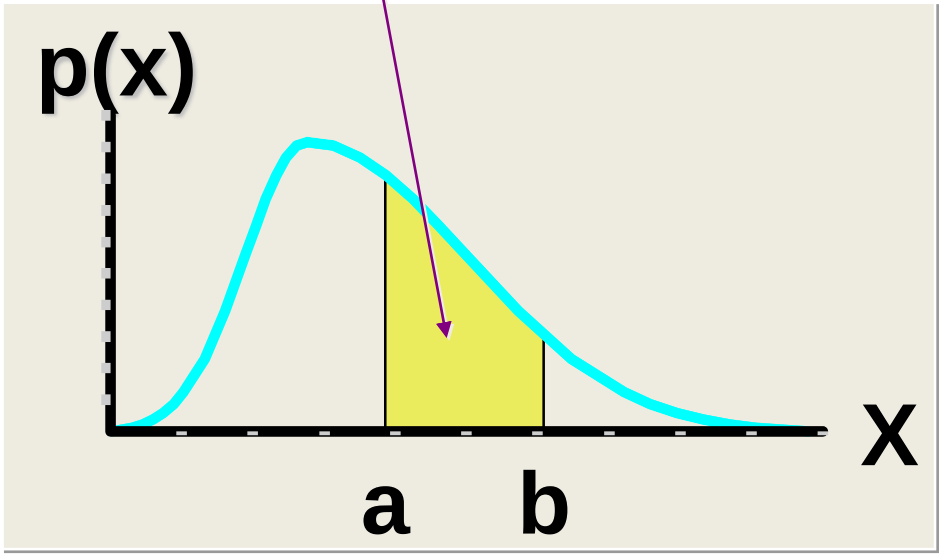


Cumulative density (cdf)

The probability that $\boldsymbol{x}$ lies in the interval $[\boldsymbol{a}, \boldsymbol{b}]$ is given by $\boldsymbol{F}(\boldsymbol{a} \leq \boldsymbol{X} \leq \boldsymbol{b})$.

Cumulative probability is Area Under Curve!

$$\boldsymbol{F}(\boldsymbol{a} \leq \boldsymbol{X} \leq \boldsymbol{b}) = \int_{\boldsymbol{a}}^{\boldsymbol{b}} \boldsymbol{p}(\boldsymbol{x}) \boldsymbol{d}\boldsymbol{x}$$



Some properties

$$\boldsymbol{F(-\infty \leq X \leq \infty) = \int_{-\infty}^{\infty} p(x)dx = 1}$$

$$\boldsymbol{F(X = a) = F(a \leq X \leq a) = \int_{a}^{a} p(x)dx = 0}$$

$$\boldsymbol{F(a) = F(-\infty \leq X \leq a) = \int_{-\infty}^{a} p(x)dx}$$

- or, more generally:

$$\boldsymbol{F(x) = \int p(x)dx}$$

- $\boldsymbol{F(x)}$ is known as the ***cumulative distribution function***. (**CDF**).
- The pdf and cdf are related by:

$$\boldsymbol{\frac{d}{dx}F(x) = p(x)}$$

- The following provides an intuitive understanding:

$$\boldsymbol{F\left(a - \frac{\Delta}{2} \leq X \leq a + \frac{\Delta}{2}\right) = \int_{a-\frac{\Delta}{2}}^{a+\frac{\Delta}{2}} p(x)dx \simeq \Delta\, \mathbf{p}(\mathbf{a})}$$

Expectation and Variance

Weighted Average

$$\boldsymbol{E[X] = \int_{-\infty}^{\infty} x\, p(x)\, dx}$$

Averaged Squared Distance From Mean

$$\boldsymbol{Var(X) = \int_{-\infty}^{\infty} (x - E[X])^2\, p(x)\, dx}$$

Bayes for Continuous random variables

Law of Total Probability

- For **discrete** random variables:

$$p(X = x) = \sum_i p(X = x, Y = y_i)$$
$$= \sum_i p(X = x | Y = y_i) p(Y = y_i)$$

- For **continuous** random variables:

$$p(x) = \int p(x, y) dy$$
$$= \int p(x|y)\, p(y) dy$$

Bayes for Continuous random variables

- From the definition of conditional probability:

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

- Substituting for $\boldsymbol{p(x)}$ we derive.

Bayes' theorem for continuous random variables:

$$\boldsymbol{p(y|x) = \frac{p(x, y)}{p(x)} = \frac{p(x|y)p(y)}{\int p(x|y)\, p(y) dy}}$$

Normal Distribution

- The normal (or Gaussian) distribution gives the familiar bell shaped curve
- **Definition**: A random variable $\boldsymbol{X}$ is normally distributed if its probability density function is given by:

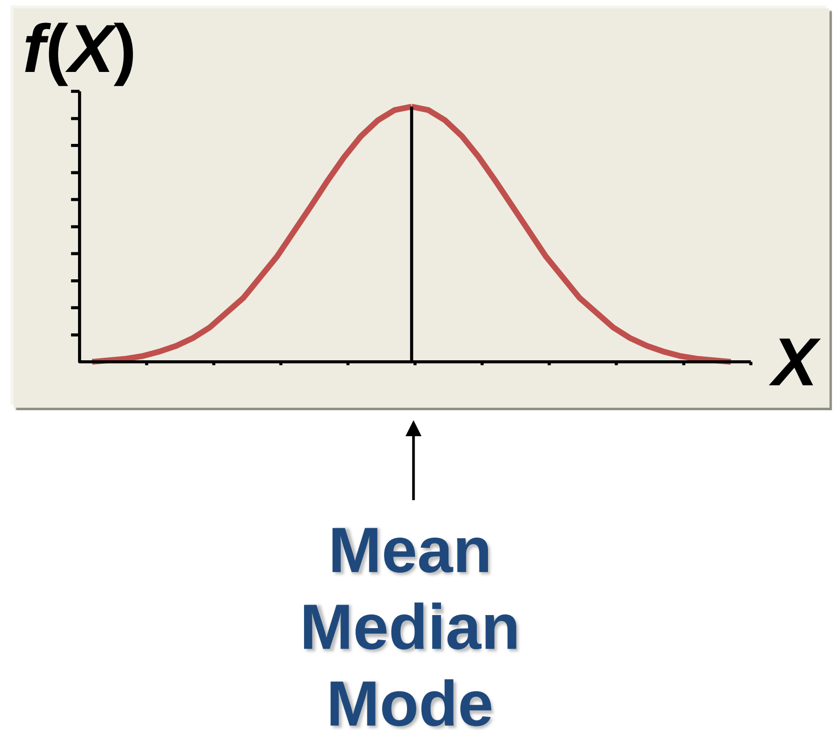
$$\boldsymbol{p(X = x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}}$$

- The parameters $\boldsymbol{\mu}$ and $\boldsymbol{\sigma^2}$ are known as the mean and variance.
- It turns out that:

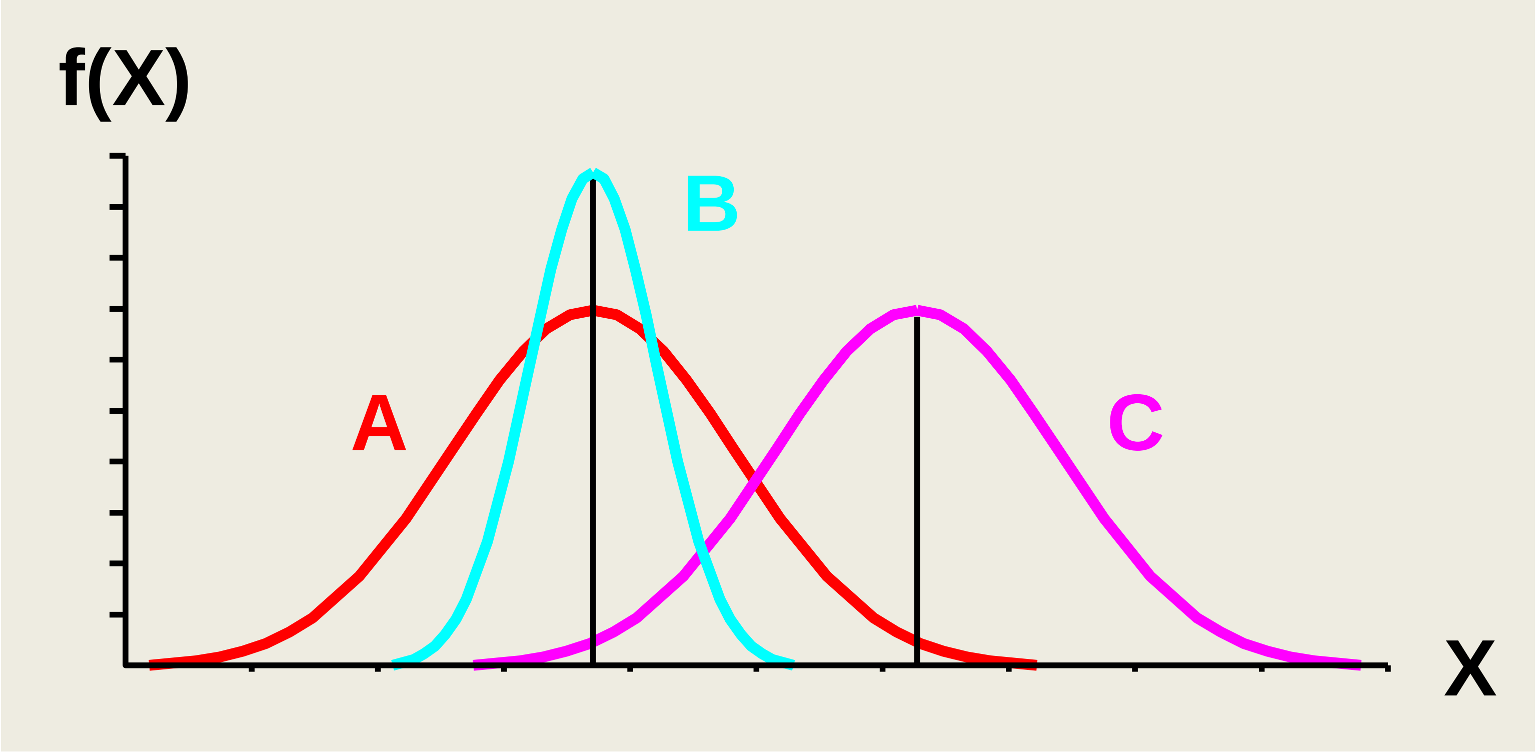
$$\boldsymbol{E[X] = \mu}$$
$$\boldsymbol{\mathrm{Var}(X) = \sigma^2}$$

The Normal Distribution

- 1. ‘Bell-Shaped’ & Symmetrical
- 2. Mean, median, mode are equal
- 3. Random variable has infinite range



Effect of varying μ and σ^2



Some properties of Normal distribution

- **Central Limit Theorem**
  - The sum of a set of independent random variables approaches the Gaussian distribution as the number of variables $\rightarrow\infty$, regardless of the distributions of the individual variables
  - (There are generalisations of the CLT.)
  - Example: the sum of the face value of randomly drawn playing cards has an approximately Gaussian distribution
- Simplicity: specified by only two intuitive parameters
- Mathematically tractable
  - Many analyses turn out very simply with the Normal distribution

Beta Distribution

- Suppose that the coin factory makes coins that are not perfectly fair all the time
- So, most times the coins are fair but less often the coins are slightly unfair, and less-less often the coins are quite unfair
- So, there is a distribution over the bias of the coins
- **Question**: How to model this distribution?
- The output/sample from this distribution will be the **bias** of a coin i.e. a number between $\mathbf{0}$ and $\mathbf{1}$

Gamma function

- The gamma function $\Gamma(N)$ generalises the factorial function to the reals such that:

$$\boldsymbol{\Gamma(N+1) = N!} \quad \text{for natural number } \boldsymbol{N}$$

$$\boldsymbol{\Gamma(1) = 0! = 1}$$

$$\boldsymbol{\Gamma\left(\frac{1}{2}\right) = \left(-\frac{1}{2}\right)! = \sqrt{\pi}}$$

$$\boldsymbol{\Gamma(0) = (-1)! = \frac{\pi}{2}}$$

- The gamma function can be used to ***interpolate*** for values for which the factorial is undefined.

Beta distribution

$$\mathbf{p}(\boldsymbol{\theta}|\boldsymbol{\alpha},\boldsymbol{\beta}) = \frac{\boldsymbol{\theta}^{\boldsymbol{\alpha}-\mathbf{1}}(\mathbf{1}-\boldsymbol{\theta})^{\boldsymbol{\beta}-\mathbf{1}}}{\boldsymbol{B}(\boldsymbol{\alpha},\boldsymbol{\beta})} \sim \boldsymbol{Beta}(\boldsymbol{\alpha},\boldsymbol{\beta})$$

- The $-\mathbf{1}$'s can be thought as mathematical convenience/convention
- To be able to show that such a thing exists, we need to show that:

$$\int_{\mathbf{0}}^{\mathbf{1}} \mathbf{p}(\boldsymbol{\theta}|\boldsymbol{\alpha},\boldsymbol{\beta})\boldsymbol{d\theta} = \int_{\mathbf{0}}^{\mathbf{1}} \frac{\boldsymbol{\theta}^{\boldsymbol{\alpha}-\mathbf{1}}(\mathbf{1}-\boldsymbol{\theta})^{\boldsymbol{\beta}-\mathbf{1}}}{\boldsymbol{B}(\boldsymbol{\alpha},\boldsymbol{\beta})}\boldsymbol{d\theta} = \mathbf{1}$$

- Equivalently, need to show that $\boldsymbol{B}(\boldsymbol{\alpha},\boldsymbol{\beta})$ is well defined:

$$\boldsymbol{B}(\boldsymbol{\alpha},\boldsymbol{\beta}) = \int_{\mathbf{0}}^{\mathbf{1}} \boldsymbol{\theta}^{\boldsymbol{\alpha}-\mathbf{1}}(\mathbf{1}-\boldsymbol{\theta})^{\boldsymbol{\beta}-\mathbf{1}}\boldsymbol{d\theta}$$

Beta function - derivation

Beta function: $\boldsymbol{B(\alpha,\beta) = \int_0^1 \theta^{\alpha-1}(1-\theta)^{\beta-1}d\theta}$

- We will apply '*integration by parts*', with $\boldsymbol{u = \theta^{\alpha-1}}$, and, $\boldsymbol{dv = (1-\theta)^{\beta-1}d\theta}$ so the integral becomes

$$\boldsymbol{B(\alpha,\beta) = \int_0^1 udv = uv - \int_0^1 vdu}$$

- $\boldsymbol{du = (\alpha-1)\theta^{\alpha-2}d\theta}$, and, $\boldsymbol{v = -\frac{1}{\beta}(1-\theta)^{\beta}}$. Thus:

$$\boldsymbol{= \theta^{\alpha-1}\left(-\frac{1}{\beta}(1-\theta)^{\beta}\right)\Bigg|_0^1 - \int_0^1\left(-\frac{1}{\beta}(1-\theta)^{\beta}\right)(\alpha-1)\theta^{\alpha-2}d\theta}$$

$$\boldsymbol{= \frac{(\alpha-1)}{\beta}\int_0^1 (1-\theta)^{\beta}\theta^{\alpha-2}d\theta = \frac{(\alpha-1)}{\beta}B(\alpha-1,\beta+1)}$$

$$\boldsymbol{= \frac{(\alpha-1)(\alpha-2)\ldots 1}{\beta(\beta+1)\ \ldots(\beta+\alpha-2)}B(1,\beta+\alpha-1)}$$

$$\boldsymbol{= \frac{(\alpha-1)(\alpha-2)\ldots 1}{\beta(\beta+1)\ \ldots(\beta+\alpha-2)}\int_0^1 \theta^{1-1}(1-\theta)^{\beta+\alpha-2}d\theta}$$

$$= \frac{(\alpha-1)(\alpha-2)\ldots 1}{\beta(\beta+1)\ \ldots(\beta+\alpha-2)} \int_0^1 (1-\theta)^{\beta+\alpha-2} d\theta$$

$$= \frac{(\alpha-1)(\alpha-2)\ldots 1}{\beta(\beta+1)\ \ldots(\beta+\alpha-2)} \left(-\frac{(1-\theta)^{\beta+\alpha-1}}{\beta+\alpha-1} \right) \Bigg|_0^1$$

$$= \frac{(\alpha-1)(\alpha-2)\ldots 1}{\beta(\beta+1)\ \ldots(\beta+\alpha-2)(\beta+\alpha-1)}$$

$$= \frac{\Gamma(\alpha)}{\beta(\beta+1)\ \ldots(\beta+\alpha-2)(\beta+\alpha-1)}$$

Strictly speaking this proof only applies to positive natural values of α, β

$$= \frac{1\ldots(\beta-2)(\beta-1)\Gamma(\alpha)}{1\ldots(\beta-2)(\beta-1)\beta(\beta+1)\ \ldots(\beta+\alpha-2)(\beta+\alpha-1)}$$

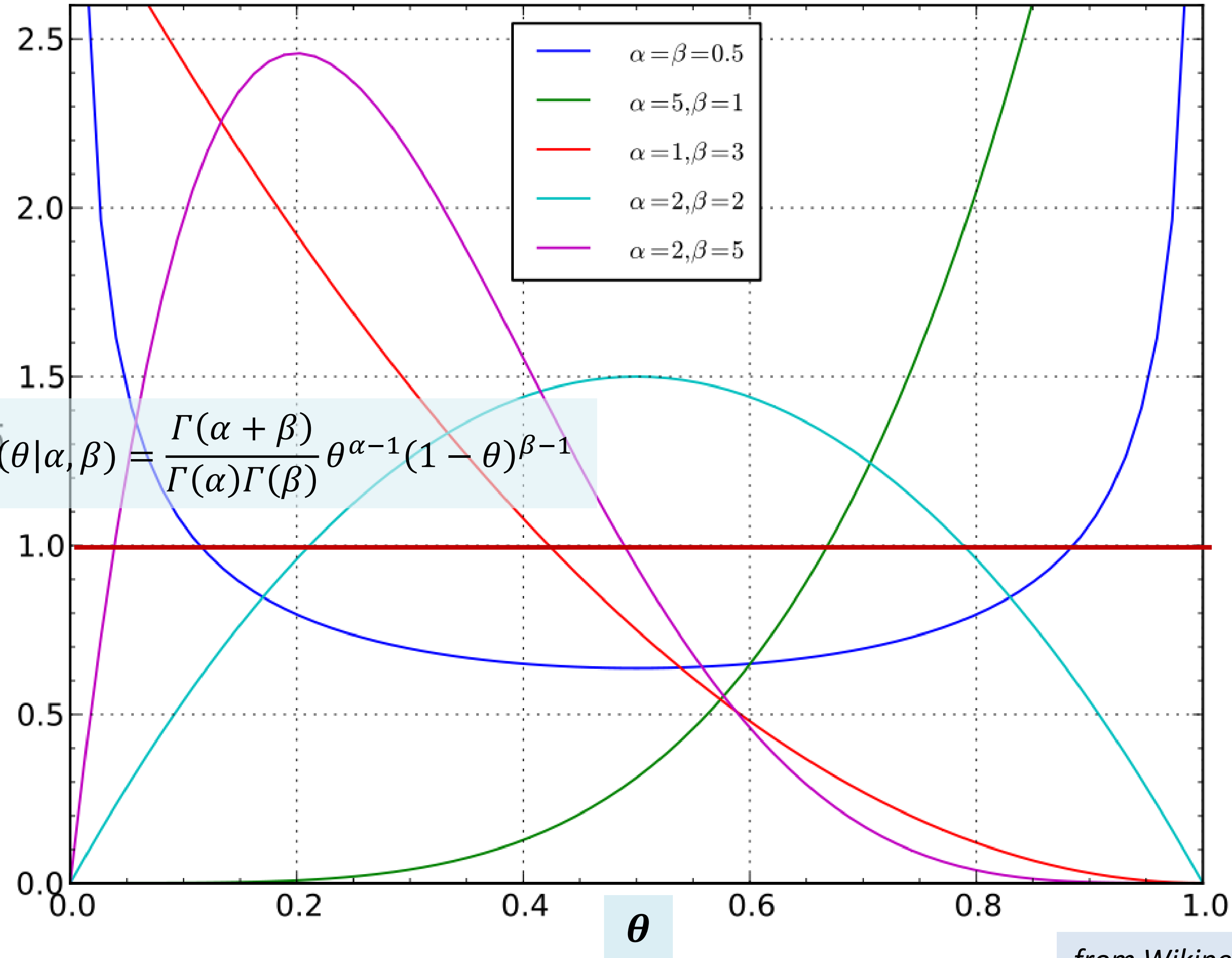
$$= \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$$

- **Thus,** $B(\alpha,\beta) = \int_0^1 \theta^{\alpha-1}(1-\theta)^{\beta-1} d\theta = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$ **hence:**

$$\mathrm{p}(\theta|\alpha,\beta) = \frac{\theta^{\alpha-1}(1-\theta)^{\beta-1}}{B(\alpha,\beta)} = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \theta^{\alpha-1}(1-\theta)^{\beta-1}$$

- For $\alpha = \beta = 1$, the distribution is ***uniform*** with $\mathrm{p}(\theta|\alpha,\beta) = 1$

Shape of Beta distribution



$$p(\theta|\alpha,\beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}$$

from Wikipedia

Modelling the coin factory

- If the factory has a high quality control then choosing $\boldsymbol{\alpha = \beta = 100}$ will give a very peaky pdf
- What would choosing a uniform distribution, $\boldsymbol{\alpha = \beta = 1}$, give? Would that be a good factory?
- Similarly, what about values less than 1?

Modelling typical coins from coin factory

- Suppose, the factory inspector visits the coin factory and picks a coin at random from the factory
- Remember, the factory has beta parameters $\boldsymbol{\alpha}, \boldsymbol{\beta}$
- The inspector flips the coin $\boldsymbol{N}$ times and sees

$$\boldsymbol{c} = (\boldsymbol{c_1}, \boldsymbol{c_2})$$

i.e $\boldsymbol{c_1}$ heads and $\boldsymbol{c_2}$ tails with $\boldsymbol{N} = \boldsymbol{c_1} + \boldsymbol{c_2}$

- What is the probability of $\boldsymbol{c_1}$ heads and $\boldsymbol{c_2}$ tails?
- How do we compute this?
- Can we use Bayes?

N here equals $\boldsymbol{c_1 + c_2}$

Modelling typical coins from coin factory

- What we want is: $\boldsymbol{p(c|\alpha,\beta)}$?

- How do we derive this: (*use law of total probability*)

$$\boldsymbol{p(c|\alpha,\beta) = \int p(c,\theta|\alpha,\beta)\, d\theta}$$

$$\boldsymbol{= \int \underbrace{p(c|\theta)}_{\text{Binomial Likelihood}}\, \underbrace{p(\theta|\alpha,\beta)}_{\text{Beta Prior}}\, d\theta}$$

- Integration can be solved analytically (i.e. by hand) if the two distributions have similar form.
- In this case, we say that that two distributions are **conjugate**.

Modelling coin tosses

- Once, we are happy with the choice of $\boldsymbol{\alpha}, \boldsymbol{\beta}$ for our factory we can plug this in

$$p(c|\alpha,\beta) = \int p(c,\theta|\alpha,\beta)\, d\theta = \int p(c|\theta)\, p(\theta|\alpha,\beta)\, d\theta$$

$$= \int \frac{N!}{c_1!\, c_2!}\, \theta^{c_1}(1-\theta)^{c_2} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \theta^{\alpha-1}(1-\theta)^{\beta-1} d\theta$$

$$= \frac{N!}{c_1!\, c_2!} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \int \theta^{c_1}(1-\theta)^{c_2}\theta^{\alpha-1}(1-\theta)^{\beta-1} d\theta$$

$$= \frac{N!}{c_1!\, c_2!} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \int \theta^{c_1+\alpha-1}(1-\theta)^{c_2+\beta-1} d\theta$$

conjugacy helps

$$= \frac{N!}{c_1!\, c_2!} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \int \theta^{c_1+\alpha-1}(1-\theta)^{c_2+\beta-1} d\theta$$

$$= \frac{N!}{c_1!\, c_2!} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \frac{\Gamma(c_1+\alpha)\Gamma(c_2+\beta)}{\Gamma(c_1+c_2+\alpha+\beta)}$$

from definition of Beta function

N here equals $c_1 + c_2$

The Beta-Binomial Distribution

- What we have just derived is the beta-binomial distribution that gives the (averaged) probability of drawing $\boldsymbol{c_1}$ heads and $\boldsymbol{c_2}$ tails from a coin that has been drawn from a beta distribution with parameters $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$.
- The 'averaged' above means that the coin parameter $\boldsymbol{\theta}$ has been ***integrated out*** (and hence no longer appears in the equations):

$$p(c|\alpha,\beta) = \frac{N!}{c_1!\,c_2!}\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\frac{\Gamma(c_1+\alpha)\Gamma(c_2+\beta)}{\Gamma(c_1+c_2+\alpha+\beta)}$$

- Using the definition of the ***beta function***:

$$B(a,b\,) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$$

$$p(c|\alpha,\beta) = \frac{N!}{c_1!\,c_2!}\frac{B(c_1+\alpha,c_2+\beta)}{B(\alpha,\beta)}$$

- $\frac{N!}{c_1!c_2!} \neq \frac{1}{B(C_1+1,C_2+1)}$ since $\frac{1}{B(C_1+1,C_2+1)} = \frac{\Gamma(C_1+C_2+2)}{\Gamma(C_1+1)\Gamma(C_2+1\,)} = \frac{(N+1)!}{C_1!C_2!}$

Inference

Inference

- The inference problem can be viewed as determining the parameters of your model from observations
- **Example**: If you throw a coin 20 times and you see heads 10 times, is the coin fair?

Inference

- The inference problem can be viewed as determining the parameters of your model from observations
- **Example**: If you throw a coin 20 times and you see heads 10 times, is the coin fair? Or what is the probability that it is fair? How do you even ask such a question?

Inference

- The inference problem can be viewed as determining the parameters of your model from observations
- **Example**: If you throw a coin 20 times and you see heads 10 times, is the coin fair? Or what is the probability that it is fair? How do you even ask such a question?
- The key to Bayesian inference is the mechanism to integrate our prior beliefs into the modelling process and provide mathematically grounded answers to the above questions.

Inference

- **Example**: If you throw a coin 20 times and you see heads 10 times, is the coin fair?

$$p(c = (c_1 = 10, c_2 = 10)|\theta) = \binom{20}{10}\theta^{10}(1-\theta)^{10}$$

- What is $p(\theta|c = (c_1 = 10, c_2 = 10))$?
- Using the definition of conditional probability:

$$p(\theta, c) = p(\theta|c)\,p(c) = p(c|\theta)p(\theta)$$

- Using Law of total probability:

$$p(c) = \int_0^1 p(\theta, c)d\theta = \int_0^1 p(c|\theta)p(\theta)d\theta$$

- Substituting, we get:

$$p(\theta|c) = \frac{p(c|\theta)p(\theta)}{\int_0^1 p(c|\theta)p(\theta)d\theta}$$

Inference

- Thus to find out $\boldsymbol{p(\theta|c = (c_1 = 10, c_2 = 10))}$ we can use:

$$\boldsymbol{p(\theta|c) = \frac{p(c|\theta)p(\theta)}{\int_0^1 p(c|\theta)p(\theta)d\theta}}$$

- However, we do not know what $\boldsymbol{p(\theta)}$ is?
- Hmmm… What might this mean?

Inference

- Thus to find out $\boldsymbol{p(\theta|c=(c_1=10, c_2=10))}$ we can use:

$$\boldsymbol{p(\theta|c) = \frac{p(c|\theta)p(\theta)}{\int_0^1 p(c|\theta)p(\theta)d\theta}}$$

- However, we do not know what $\boldsymbol{p(\theta)}$ is?
- Hmmm… What might this mean?
- $\boldsymbol{p(\theta)}$ is our prior belief about the coin with bias $\boldsymbol{\theta}$
- What does this mean?

Inference

Data:
- $\boldsymbol{c}$ is the ***data***
- ***or observations***

$$\boldsymbol{p(\theta|c)} = \frac{\boldsymbol{p(c|\theta)p(\theta)}}{\int \boldsymbol{p(c|\theta)p(\theta)d\theta}}$$

Posterior:

- $\boldsymbol{p(\theta|c)}$ is the ***posterior distribution***
- Gives the probability of the parameter given data

Prior:

- $\boldsymbol{p(\theta)}$ is the ***prior distribution***
- Gives the probability for different values of $\boldsymbol{\theta}$ and quantifies our belief regarding $\boldsymbol{\theta}$

Likelihood:

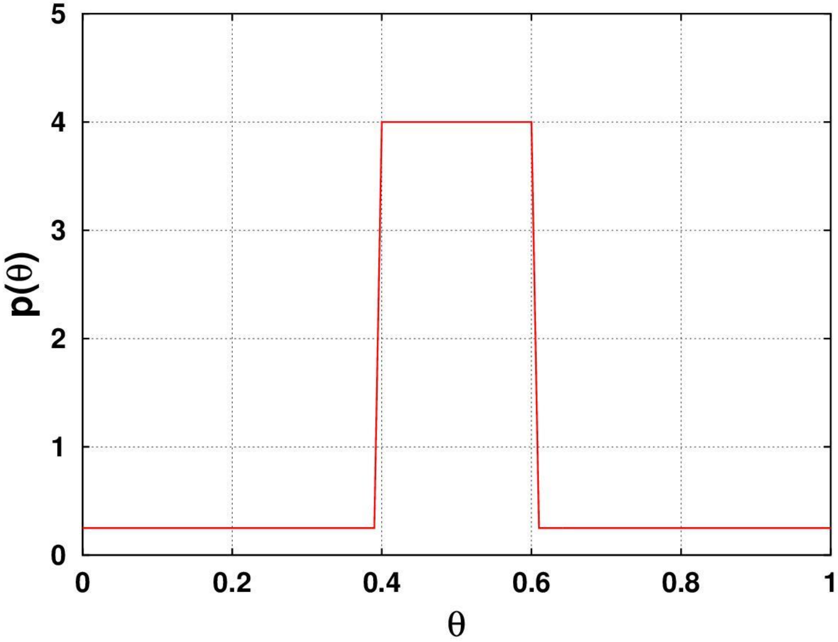
- $\boldsymbol{p(c|\theta)}$ is the ***likelihood*** of the data given by the model
- Gives the probability of the data being generated by the model

Partition function/Normalising constant/Evidence:

- $\int_0^1 \boldsymbol{p(c|\theta)p(\theta)d\theta}$ is the partition function/normalising constant/evidence.
- This is a constant needed to ensure that the probabilities sum to 1.

Inference

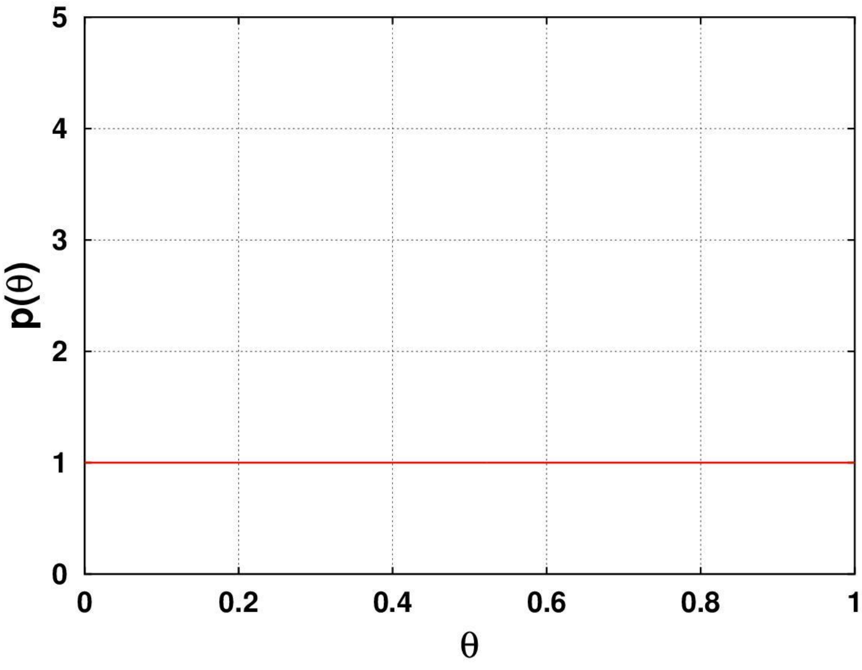
- $\boldsymbol{p(\theta)}$ is our prior belief about the coin with bias $\boldsymbol{\theta}$
- Suppose, we say that, $\boldsymbol{p(\theta)}$ is **piecewise uniform** with:
  - $\boldsymbol{F(0.4 \leq \theta \leq 0.6) = 0.8}$, $\boldsymbol{F(0 \leq \theta < 0.4) = 0.1}$, $\boldsymbol{F(0.6 < \theta \leq 1) = 0.1}$



Inference

- Alternatively suppose that, $\boldsymbol{p(\theta)}$ is **uniform** with:

$$\boldsymbol{p(\theta)} = \begin{cases} \mathbf{1} & if \quad 0 < \theta < 1 \\ \mathbf{0} & otherwise \end{cases}$$



Inference

- Alternatively still, we can assume that $\boldsymbol{p(\theta)}$ is given by a ***beta distribution*** with parameters $\boldsymbol{\alpha, \beta}$

$$\mathbf{p}(\boldsymbol{\theta}|\boldsymbol{\alpha},\boldsymbol{\beta}) = \frac{\boldsymbol{\theta}^{\boldsymbol{\alpha}-1}(1-\boldsymbol{\theta})^{\boldsymbol{\beta}-1}}{B(\boldsymbol{\alpha},\boldsymbol{\beta})} = \frac{\Gamma(\boldsymbol{\alpha}+\boldsymbol{\beta})}{\Gamma(\boldsymbol{\alpha})\Gamma(\boldsymbol{\beta})}\boldsymbol{\theta}^{\boldsymbol{\alpha}-1}(1-\boldsymbol{\theta})^{\boldsymbol{\beta}-1}$$

- The uniform distribution can be recovered by $\boldsymbol{\alpha = 1, \beta = 1}$

$$\mathbf{p}(\boldsymbol{\theta}|1,1) = \frac{\boldsymbol{\theta}^{1-1}(1-\boldsymbol{\theta})^{1-1}}{B(\boldsymbol{\alpha},\boldsymbol{\beta})} = \frac{\Gamma(1+1)}{\Gamma(1)\Gamma(1)} = 1$$

- The piecewise uniform distribution may also be approximated by suitable choices of $\boldsymbol{\alpha, \beta}$
- Choosing the ***beta distribution*** as a **prior** means that we can take advantage of ***conjugacy***.

Inference: Binomial Posterior Distribution

- For the case when $\boldsymbol{p(\theta)}$ is given by the **beta distribution** with parameters $\boldsymbol{\alpha, \beta}$:

$$p(\theta|c,\alpha,\beta) = \frac{p(c|\theta,\alpha,\beta)\, p(\theta|\alpha,\beta)}{\int_0^1 p(c|\theta,\alpha,\beta)\, p(\theta|\alpha,\beta) d\theta}$$

Remember $\boldsymbol{c = (c_1, c_2)}$

$$= \frac{\frac{N!}{c_1!\,c_2!}\,\theta^{c_1}(1-\theta)^{c_2}\,\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\,\theta^{\alpha-1}(1-\theta)^{\beta-1}}{\int_0^1 \frac{N!}{c_1!\,c_2!}\,\theta^{c_1}(1-\theta)^{c_2}\,\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\,\theta^{\alpha-1}(1-\theta)^{\beta-1} d\theta}$$

$$= \frac{\theta^{c_1+\alpha-1}(1-\theta)^{c_2+\beta-1}}{\frac{\Gamma(c_1+\alpha)\Gamma(c_2+\beta)}{\Gamma(c_1+c_2+\alpha+\beta)}}$$

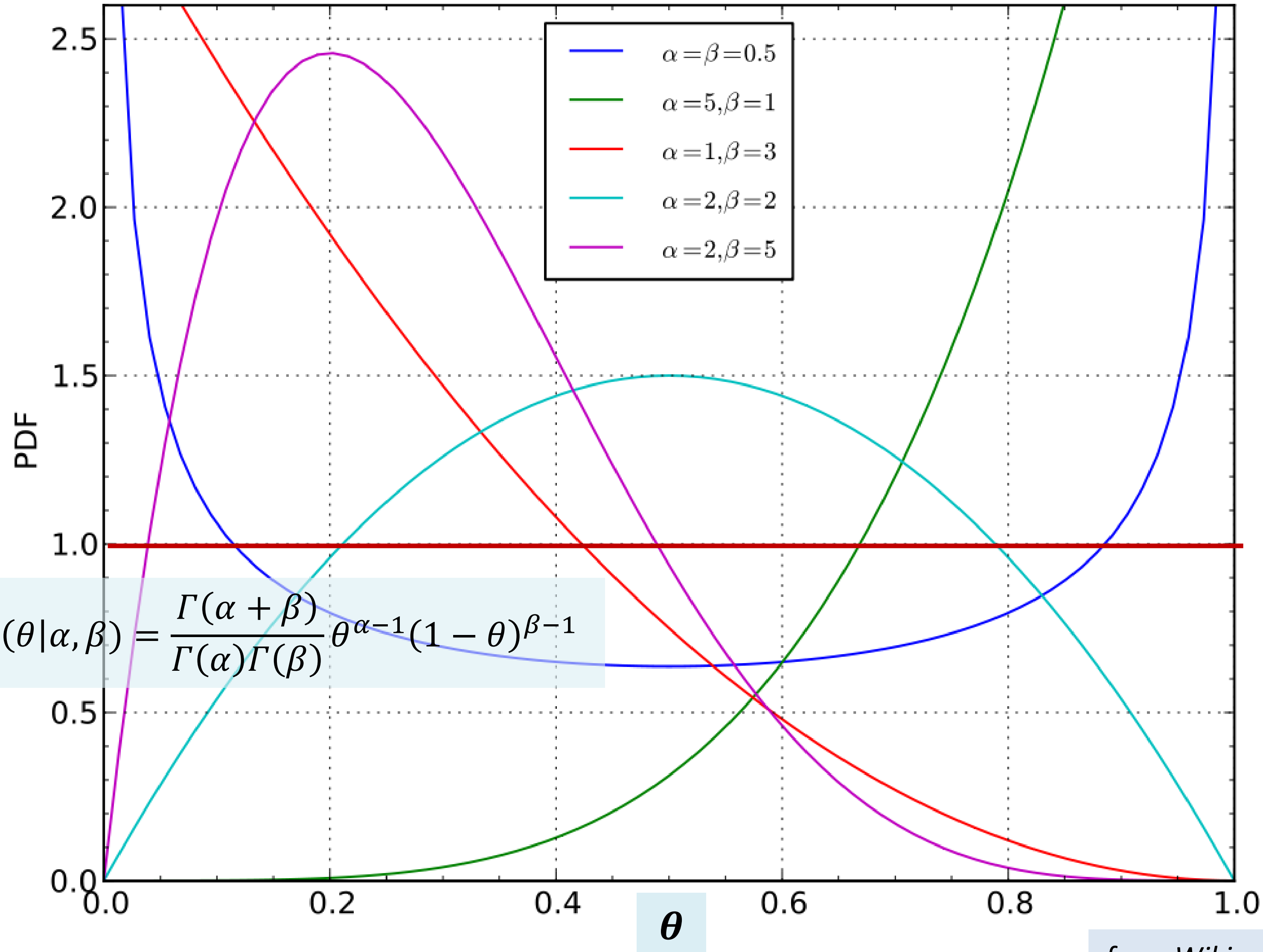
$$= \frac{\Gamma(c_1+c_2+\alpha+\beta)}{\Gamma(c_1+\alpha)\Gamma(c_2+\beta)}\,\theta^{c_1+\alpha-1}(1-\theta)^{c_2+\beta-1}$$

$$= \frac{1}{B(c_1+\alpha, c_2+\beta)}\,\theta^{c_1+\alpha-1}(1-\theta)^{c_2+\beta-1}$$

$$= p(\theta|\alpha+c_1, \beta+c_2) \sim Beta(\alpha+c_1, \beta+c_2)$$

- Thus, posterior of binomial (under beta) is beta with counts added into the parameters

Shape of Beta distribution



$$p(\theta|\alpha,\beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}$$

from Wikipedia

Aside: Inference

- If $\boldsymbol{\theta}$ is ***piecewise uniform*** with, $\boldsymbol{F(0.4 \leq \theta \leq 0.6) = 0.8}$, $\boldsymbol{F(0 \leq \theta < 0.4) = 0.1}$, $\boldsymbol{F(0.6 < \theta \leq 1) = 0.1}$
- We can calculate, $\boldsymbol{p(\theta)}$:

$\boldsymbol{F(0.4 \leq \theta \leq 0.6) = 0.8}$ implies $\boldsymbol{p(\theta) = \frac{0.8}{0.2} = 4}$

$\boldsymbol{F(0 \leq \theta < 0.4) = 0.1}$ implies $\boldsymbol{p(\theta) = \frac{0.1}{0.4} = \frac{1}{4}}$

$\boldsymbol{F(0.6 < \theta \leq 1) = 0.1}$ implies $\boldsymbol{p(\theta) = \frac{0.1}{0.4} = \frac{1}{4}}$

Aside: Inference

$$p(\theta|c) = \frac{p(c|\theta)\, p(\theta)}{\int_0^1 p(c|\theta)\, p(\theta) d\theta}$$

$$= \frac{p(c|\theta)\, p(\theta)}{\int_0^{0.4} p(c|\theta)\, .\frac{1}{4} d\theta + \int_{0.4}^{0.6} p(c|\theta).\, 4\, d\theta + \int_{0.6}^{1} p(c|\theta)\, .\frac{1}{4} d\theta}$$

$$= \frac{\theta^{c_1}(1-\theta)^{c_2}\, p(\theta)}{\frac{1}{4} IB(.4, c_1, c_2) + 4[IB(.6, c_1, c_2) - IB(.4, c_1, c_2)] + \frac{1}{4}[B(c_1, c_2) - IB(.6, c_1, c_2)]}$$

- Where IB is the ***incomplete beta function*** defined by:

$$IB(a, c_1, c_2) = \int_0^a \theta^{c_1 - 1}(1-\theta)^{c_2 - 1} d\theta \qquad\qquad 0 \le a \le 1$$

- Hence: $IB(1, c_1, c_2) = B(c_1, c_2)$
- Most math packages provide implementations of the incomplete beta function.
- Thus, posterior probability for any θ and c can be calculated.

Bayesian Inference

$$\boldsymbol{p(\theta|c)} = \frac{\boldsymbol{p(c|\theta)p(\theta)}}{\int \boldsymbol{p(c|\theta)p(\theta)d\theta}}$$

- ***Being Bayesian*** typically means that we treat $\boldsymbol{p(\theta|c)}$ as a distribution
- This means that we do not get a single value for the model parameter $\boldsymbol{\theta}$
- However, sometimes, we may want to find out what the '**best**' value for the model parameter $\boldsymbol{\theta}$ would be.
- How do define what '**best**' means ?

MLE Inference

- This is the maximum likelihood estimate (where c is the data)

$$\boldsymbol{\theta}_{MLE} = \arg\max_{\boldsymbol{\theta}} p(\boldsymbol{c}|\boldsymbol{\theta})$$

- Thus MLE is simply the maximum/**mode** of the model likelihood.
- We don't even need to define a prior, so (apparently) more "objective".
- Typically deal with situations where there is only one mode, so can talk about **the** MLE.

MAP Inference

$$p(\theta|c) = \frac{p(c|\theta)p(\theta)}{\int p(c|\theta)p(\theta)d\theta}$$

MAP (Maximum a Posteriori):

- Assume that $p(\theta)$ is distributed as per ***some given distribution***

$$\theta_{MAP} = \arg\max_{\theta} p(\theta|c) = \arg\max_{\theta} \frac{p(c|\theta)p(\theta)}{\int_0^1 p(c|\theta)p(\theta)d\theta}$$

- Since the denominator is a constant:

$$\theta_{MAP} = \arg\max_{\theta} p(c|\theta)p(\theta)$$

- Thus MAP estimate is the ***mode*** of the posterior distribution

Multinomial Distribution

- Probability of observing $\boldsymbol{c} = (\boldsymbol{c_1}, \dots, \boldsymbol{c_k})$ heads in all possible ways out of $\boldsymbol{C} = \sum_i \boldsymbol{c_i}$ throws from a k-headed dice with probability of heads $\boldsymbol{\theta} = (\boldsymbol{\theta_1}, \dots, \boldsymbol{\theta_k})$ *s.t.* $\sum_i \boldsymbol{\theta_i} = \mathbf{1}$

$$\boldsymbol{p(heads = c|\theta)} = \frac{\boldsymbol{C}!}{\prod_i \boldsymbol{c_i}!} \prod_{\boldsymbol{i}} \boldsymbol{\theta}^{\boldsymbol{c_i}}$$

- Points to note:
  - Generalises the binomial distribution to $\boldsymbol{k} > \mathbf{2}$
  - Equivalent to the binomial for $\boldsymbol{k} = \mathbf{2}$

The Dirichlet Distribution

- Dirichlet distribution generalises the Beta distribution to the $\boldsymbol{k-1}$ probability simplex
- The Beta distribution:

$$\boldsymbol{p(\theta|\alpha,\beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1} \sim Beta(\alpha,\beta)}$$

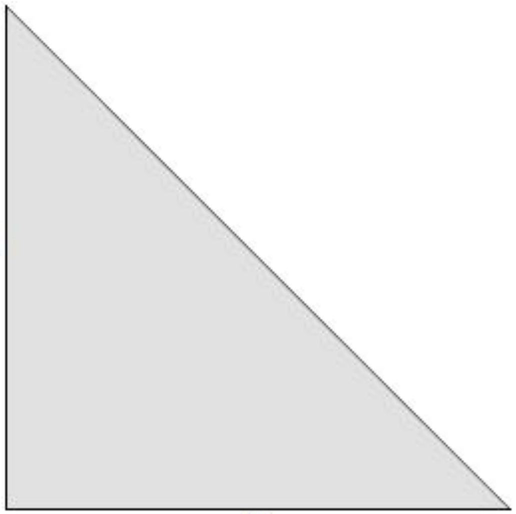
- The Dirichlet distribution:

$$\boldsymbol{p(\theta|\alpha) = \frac{\Gamma(A)}{\prod_i \Gamma(\alpha_i)}\prod_i \theta_i^{\alpha_i-1} \sim Dir(\alpha_1,\dots,\alpha_k)}$$

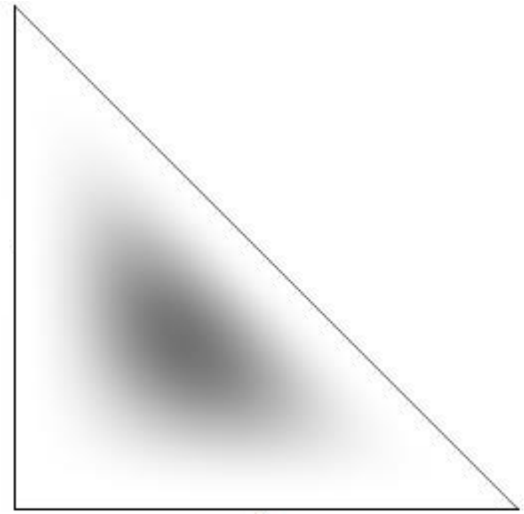
with:

$$\boldsymbol{\alpha = (\alpha_1,\dots,\alpha_k), \theta = (\theta_1,\dots,\theta_k), \sum_i \theta_i = 1, A = \sum_i \alpha_i}$$

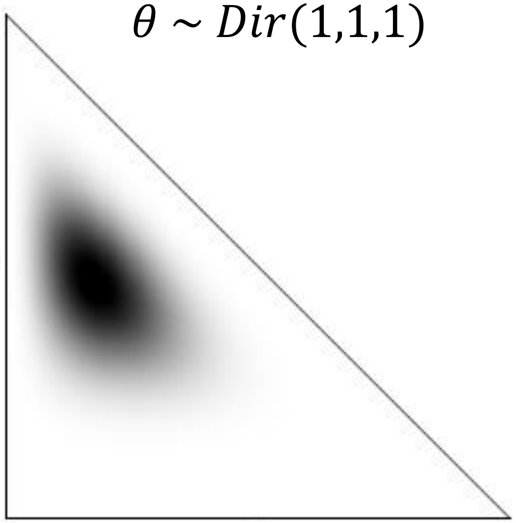
- So, the samples from the Dirichlet distribution can be used to model the bias in a $\boldsymbol{k}$-sided dice.



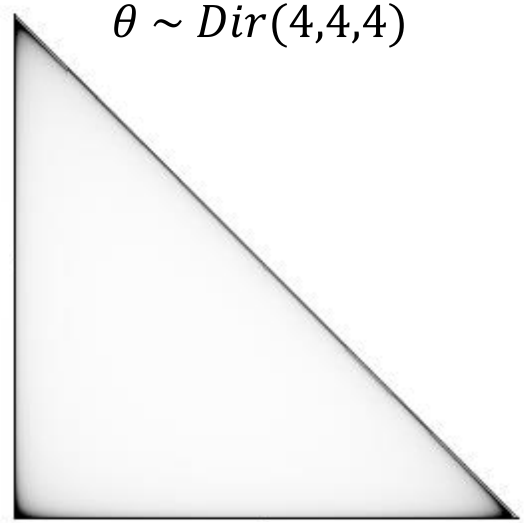
$\theta \sim Dir(1,1,1)$



$\theta \sim Dir(4,4,4)$



$\theta \sim Dir(4,9,7)$



$\theta \sim Dir(0.2, 0.2, 0.2)$

The Dirichlet-Multinomial

- Like for the Beta-Binomial distribution we can integrate out the Multinomial parameters
- Here, we are modelling the case where we need to predict the outcome $\boldsymbol{c} = (\boldsymbol{c_1}, \dots, \boldsymbol{c_k})$ of a dice throw from a dice sampled from a factory with parameter $\boldsymbol{\alpha} = (\boldsymbol{\alpha_1}, \dots, \boldsymbol{\alpha_k})$

$$
\begin{aligned}
\boldsymbol{p(c|\alpha)} &= \int \boldsymbol{p(c, \theta|\alpha)\, d\theta} = \int \boldsymbol{p(c|\theta)\, p(\theta|\alpha)\, d\theta} \\
&= \int \left[\frac{\boldsymbol{C}!}{\prod_i \boldsymbol{c_i}!} \prod_i \boldsymbol{\theta^{c_i}} \right] \frac{\boldsymbol{\Gamma(A)}}{\prod_i \boldsymbol{\Gamma(\alpha_i)}} \prod_i \boldsymbol{\theta_i^{\alpha_i - 1}}\ \boldsymbol{d\theta} \\
&= \frac{\boldsymbol{C}!}{\prod_i \boldsymbol{c_i}!} \frac{\boldsymbol{\Gamma(A)}}{\prod_i \boldsymbol{\Gamma(\alpha_i)}} \int \prod_i \boldsymbol{\theta_i^{c_i + \alpha_i - 1}}\ \boldsymbol{d\theta}
\end{aligned}
$$

- From earlier: $\boldsymbol{B(\alpha, \beta)} = \int_0^1 \boldsymbol{\theta^{\alpha-1}(1-\theta)^{\beta-1} d\theta} = \frac{\boldsymbol{\Gamma(\alpha)\Gamma(\beta)}}{\boldsymbol{\Gamma(\alpha+\beta)}}$
- This generalises to: $\boldsymbol{\Delta(\alpha)} = \int_0^1 \prod_i \boldsymbol{\theta_i^{\alpha_i - 1}}\ \boldsymbol{d\theta} = \frac{\prod_i \boldsymbol{\Gamma(\alpha_i)}}{\boldsymbol{\Gamma(A)}}$

$$
\boldsymbol{p(c|\alpha)} = \frac{\boldsymbol{C}!}{\prod_i \boldsymbol{c_i}!} \frac{\boldsymbol{\Gamma(A)}}{\prod_i \boldsymbol{\Gamma(\alpha_i)}} \int \prod_i \boldsymbol{\theta_i^{c_i + \alpha_i - 1}}\ \boldsymbol{d\theta} = \frac{\boldsymbol{C}!}{\prod_i \boldsymbol{c_i}!} \frac{\boldsymbol{\Gamma(A)}}{\prod_i \boldsymbol{\Gamma(\alpha_i)}} \frac{\prod_i \boldsymbol{\Gamma(c_i + \alpha_i)}}{\boldsymbol{\Gamma(C + A)}}
$$

Posterior Multinomial under Dirichlet prior

- Suppose we observe counts $\boldsymbol{c} = (\boldsymbol{c_1}, \dots, \boldsymbol{c_k})$ from a dice sampled from our factory.
- We would like to predict the most likely parameters $\boldsymbol{\theta}$ for this dice.

$$\boldsymbol{p(\theta|c,\alpha) = \frac{p(c,\theta|\alpha)}{p(c|\alpha)} = \frac{p(c,\theta|\alpha)}{\int p(c,\theta|\alpha)d\theta} = \frac{p(c|\theta)p(\theta|\alpha)}{\int p(c|\theta)p(\theta|\alpha)d\theta}}$$

$$\boldsymbol{= \frac{\frac{C!}{\prod_i c_i!}\prod_i \theta_i^{c_i}\frac{\Gamma(A)}{\prod_i \Gamma(\alpha_i)}\prod_i \theta_i^{\alpha_i - 1}}{\int p(c|\theta)p(\theta|\alpha)d\theta} = \frac{\frac{C!}{\prod_i c_i!}\frac{\Gamma(A)}{\prod_i \Gamma(\alpha_i)}\prod_i \theta_i^{c_i+\alpha_i-1}}{\int \frac{C!}{\prod_i c_i!}\frac{\Gamma(A)}{\prod_i \Gamma(\alpha_i)}\prod_i \theta_i^{c_i+\alpha_i-1}\, d\theta}}$$

$$\boldsymbol{= \frac{\prod_i \theta_i^{c_i+\alpha_i-1}}{\frac{\prod_i \Gamma(c_i+\alpha_i)}{\Gamma(C+A)}} = \frac{\Gamma(C+A)}{\prod_i \Gamma(c_i+\alpha_i)}\prod_i \theta_i^{c_i+\alpha_i-1} \sim Dir(c_1+\alpha_1, \dots, c_k+\alpha_k)}$$

- So, the shape of the posterior is exactly like that of the prior with counts added into the parameters