

# *Artificial Intelligence*

## Probabilistic reasoning: *learning*

Marco Piastra

# Types of learning problems

Consider a number of observations (input data) made by an agent

$$\{D_1, D_2, \dots, D_n\}$$

## ■ **Supervised learning**

Together with the input objects  $\{D_1, D_2, \dots, D_n\}$ , the agent knows a set of corresponding expected values  $\{Y_1, Y_2, \dots, Y_n\}$

The objective is abstracting a *function*  $Y_i = f(D_i)$  that *generalizes* this correspondence

## ■ **Unsupervised learning**

The objective is constructing from a set of observations  $\{D_1, D_2, \dots, D_n\}$  a descriptive *model*

In probabilistic terms, the objective is learning a *joint distribution*  $P$

## ■ **Reinforcement learning**

The observations  $\{D_1, D_2, \dots, D_n\}$  are *states* or *situations*, at each state  $X_i$  the agent must perform an **action**  $a_i$  that produces a **result**  $r_i$ .

The objective is defining a *function*  $a_i = \pi(D_i)$  that describes a strategy that the agent will follow

The strategy should be optimal, in the sense that it should maximize the expected value of a *function*  $v(\langle r_1, r_2, \dots, r_n \rangle)$  of the sequence of *results*

# Events and observations

## ■ Events

An **event** is a subset of *possible worlds*

An event **occurs** when the actual world is known to belong to the subset

## ■ Multiple random variables

A convenient way to define a  $\sigma$ -algebra of events

In the discrete case, each combination of values of the random variables describes an *event*

## ■ Observations (data)

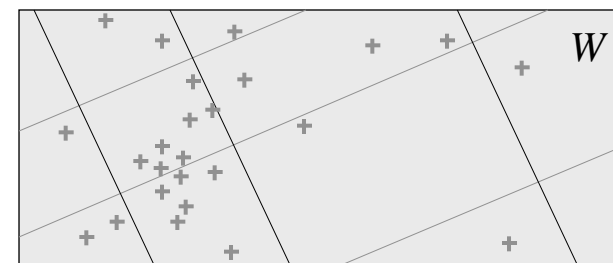
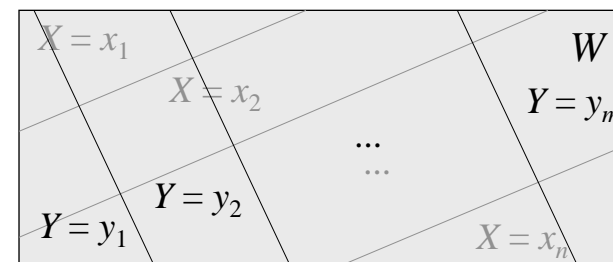
Each *observation* is about one *possible world*

In each possible world, all the values of random variables are determined

Observations could be either *complete* or *partial*

In the sense that not all the relevant values could be actually observed

How do observations (complete or not) connect to *probability*  $P$  ?



# Observations and Independence

## *About notation*

Each observation could be the outcome of an experiment or a test

The outcome of a particular experiment will be represented by a set of *random variables*

For example, if the model adopts the random variable  $\{X, Y\}$ ,  
the  $n$  outcomes of the experiments are  $D_1 = \{X_1, Y_1\}$ ,  $D_2 = \{X_2, Y_2\}$ , ...,  $D_n = \{X_n, Y_n\}$

## ■ *Independent observations, same probability distribution*

*Independent, Identically Distributed (IID) random variables*

### Definition

A sequence or set of random variables  $\{X_1, X_2, \dots, X_n\}$  is IID iff:

- 1)  $\langle X_i \perp X_j \rangle, i \neq j$  (independence)
- 2)  $P(X_i) = P(X_j), i \neq j$  (same distribution)

The extension to sequences of subsets of random variable is immediate

### CAUTION:

*Being IID is not an obvious property of observations*

Example: different measurements on different patients *may* be IID,  
but different measurements over time on the same patient are not IID

# Maximum Likelihood Estimation (MLE)

A probabilistic model  $P(X)$ , with parameters  $\theta$

$\theta$  is a vector of values that characterizes  $P(X)$  *completely*: once  $\theta$  is defined,  $P(X)$  is also defined.

A set of IID observations  $D = \{D_1, D_2, \dots, D_n\}$

## ■ Likelihood function

A function, or a conditional probability, derived from the model  $P(X)$

$$L(\theta | D) = P(D | \theta) = P(D_1, D_2, \dots, D_n | \theta)$$

where  $P(D | \theta)$  is the conditional probability that the parameter  $\theta$ , considered as a random variable, could generate the observations  $D$

When the observations  $\{D_1, D_2, \dots, D_n\}$  are IID:

$$P(D | \theta) = P(D_1 | \theta)P(D_2 | \theta) \dots P(D_n | \theta) = \prod_i P(D_i | \theta)$$

## ■ Maximum Likelihood Estimation

$$\theta_{ML}^* = \arg \max_{\theta} L(\theta | D)$$

When the observations are IID, the *Log-Likelihood* could ease computations:

$$\ell(\theta | D) = \log L(\theta | D) = \log \prod_i P(D_i | \theta) = \sum_i \log P(D_i | \theta)$$

$$\theta_{ML}^* = \arg \max_{\theta} \ell(\theta | D)$$

# Example: coin tossing (*Bernoulli Trials*)

Test: tossing a coin  $X$ , not necessarily *fair*. ( $X = 1$  head,  $X = 0$  tail)

Model:  $P(X = 1) = \theta$ ,  $P(X = 0) = 1 - \theta$

Observations: a sequence  $\langle D_1, D_2, \dots, D_n \rangle$

(i.e.  $D = \{D_1 = \{X_1 = 1\}, D_2 = \{X_2 = 1\}, D_3 = \{X_3 = 0\} \dots\}$ )

## ■ (Log-)Likelihood Function

$$\ell(\theta | D) = \log P(D | \theta) = \log P(\{X_i\} | \theta) = \log \prod_i P(X_i | \theta) = \sum_i \log P(X_i | \theta)$$

Likelihood for  $P$ : (Algebraic Follies!)

$$P(X | \theta) = \theta^{[X=1]} (1 - \theta)^{[X=0]} \quad \text{where:} \quad [X_i = v] = \begin{cases} 1 & \text{se } X_i = v \\ 0 & \text{se } X_i \neq v \end{cases}$$

$$\ell(\theta | D) = \sum_i \log \left( \theta^{[X_i=1]} (1 - \theta)^{[X_i=0]} \right) = \log \theta \sum_i [X_i = 1] + \log (1 - \theta) \sum_i [X_i = 0]$$

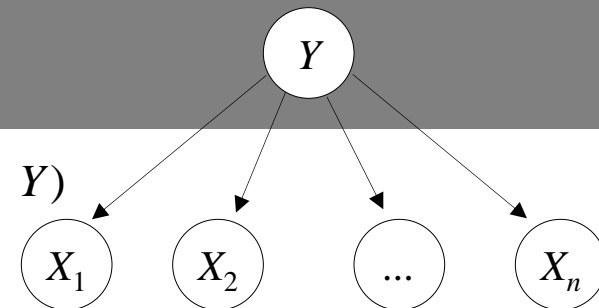
$$= N_{X=1} \log \theta + N_{X=0} \log (1 - \theta) \quad (\text{where } N_{X=1} \text{ is the number of } X_i = 1 \text{ in the sequence } D)$$

## ■ Maximum Likelihood Estimation

$$\frac{\partial \ell}{\partial \theta} = \frac{N_{X=1}}{\theta} - \frac{N_{X=0}}{(1 - \theta)} \quad \frac{\partial \ell}{\partial \theta} = 0 \quad \Rightarrow \quad \theta_{ML}^* = \frac{N_{X=1}}{N_{X=1} + N_{X=0}}$$

# Anti-spam filter

$$P(Y, \{X_i\}) = P(Y) \prod_{i=1}^n P(X_i | Y)$$



Model: the *conditional probability tables* in the graphical model

$$P(Y = k) = \pi_k, \quad P(X_i = j | Y = k) = \eta_{ijk}$$

Observations: a set of messages, with classification

$$D = \{D_1 = \{Y_1 = 1, X_{11} = 1, X_{12} = 1, \dots, X_{1n} = 0\}, D_2 = \{Y_2 = 0, X_{21} = 0, X_{22} = 1, \dots, X_{2n} = 1\}, \dots\}$$

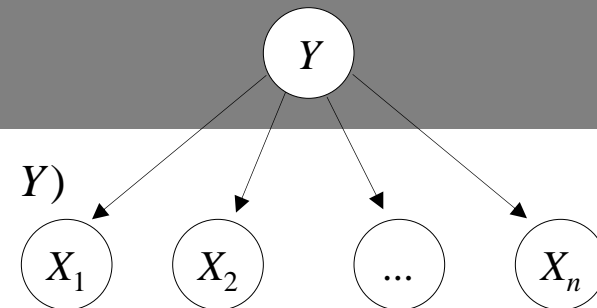
## ■ Likelihood Function

Sequence of messages

$$\begin{aligned} L(\{\pi_k, \eta_{ijk}\} | D) &= P(D | \theta) = P(\{D_m\} | \{\pi_k, \eta_{ijk}\}) = \prod_m P(D_m | \{\pi_k, \eta_{ijk}\}) && \text{(messages are IID)} \\ &= \prod_m P(\{Y_m = y_m, X_{mi} = x_{mi}\} | \{\pi_k, \eta_{ijk}\}) \\ &= \prod_m P(Y_m = y_m | \{\pi_k, \eta_{ijk}\}) P(\{X_{mi} = x_{mi}\} | Y_m = y_m, \{\pi_k, \eta_{ijk}\}) && \text{(factorization)} \\ &= \prod_m P(Y_m = y_m | \{\pi_k\}) P(\{X_{mi} = x_{mi}\} | Y_m = y_m, \{\eta_{ijk}\}) && \text{(cond. independence)} \\ &= \prod_m P(Y_m = y_m | \{\pi_k\}) \prod_i P(X_{mi} = x_{mi} | Y_m = y_m, \{\eta_{ijk}\}) && (\langle X_i \perp X_j, Y \rangle) \end{aligned}$$

# Anti-spam filter

$$P(Y, \{X_i\}) = P(Y) \prod_{i=1}^n P(X_i | Y)$$



## ■ Log-Likelihood Function

$$\ell(\{\pi_k, \eta_{ijk}\} | D) = \sum_m \log P(Y_m = y_m | \{\pi_k\}) + \sum_m \sum_i \log P(X_{mi} = x_{mi} | Y_m = y_m, \{\eta_{ijk}\})$$

Alternative form for  $P$ :

$$P(Y = k | \{\pi_k\}) = \pi_k = \prod_k \pi_k^{[Y=k]} \quad (\text{Algebraic Follies!})$$

$$P(X_i = j | Y_m = k, \{\eta_{ijk}\}) = \eta_{ijk} = \prod_j \prod_k \eta_{ijk}^{[X_i=j][Y_m=k]}$$

$$\ell(\{\pi_k, \eta_{ijk}\} | D) = \sum_m \sum_k [Y_m = k] \log \pi_k + \sum_m \sum_i \sum_j \sum_k [X_{mi} = j][Y_m = k] \log \eta_{ijk}$$

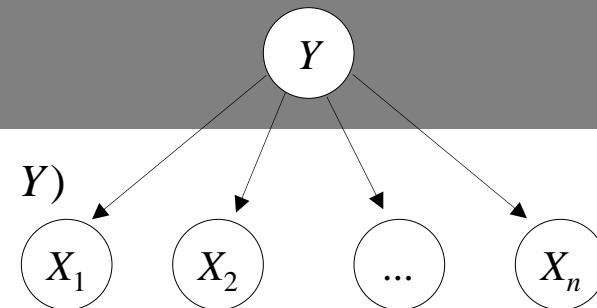
## ■ Maximum Likelihood Estimation

Being both positive, the two terms of the function can be optimized separately



# Anti-spam filter

$$P(Y, \{X_i\}) = P(Y) \prod_{i=1}^n P(X_i | Y)$$



## Maximum Likelihood Estimation

$$\ell(\{\pi_k, \eta_{ijk}\} | D) = \sum_m \sum_k [Y_m = k] \log \pi_k + \sum_m \sum_i \sum_j \sum_k [X_{mi} = j][Y_m = k] \log \eta_{ijk}$$

First term:

$$\ell^*(\{\pi_k\} | D) = \sum_m \sum_k [Y_m = k] \log \pi_k + \lambda (1 - \sum_k \pi_k)$$

Lagrange multiplier

$$\frac{\partial \ell^*}{\partial \pi_k} = \frac{\sum [Y_m = k]}{\pi_k} - \lambda$$

number of messages in  $D$  classified as  $k$

$$\frac{\partial \ell^*}{\partial \pi_k} = 0 \Rightarrow \pi_k = \frac{N_{Y=k}}{\lambda}$$

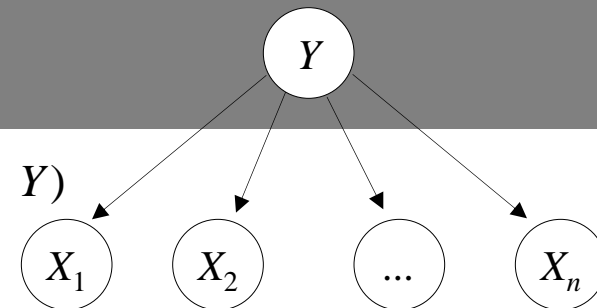
$$\sum_k \pi_k = 1 \Rightarrow \sum_k \frac{N_{Y=k}}{\lambda} = 1 \Rightarrow \lambda = \sum_k N_{Y=k} = N_D$$

number of messages in  $D$

$$\pi_k^* = \frac{N_{Y=k}}{N_D} \quad (\text{Maximum Likelihood Estimator of } \pi_k)$$

# Anti-spam filter

$$P(Y, \{X_i\}) = P(Y) \prod_{i=1}^n P(X_i | Y)$$



## ■ Maximum Likelihood Estimation

$$\ell(\{\pi_k, \eta_{ijk}\} | D) = \sum_m \sum_k [Y_m = k] \log \pi_k + \sum_m \sum_i \sum_j \sum_k [X_{mi} = j][Y_m = k] \log \eta_{ijk}$$

Second term:

$$\ell^*(\{\eta_{ijk}\} | D) = \sum_m \sum_i \sum_j \sum_k [X_{mi} = j][Y_m = k] \log \eta_{ijk} + \sum_i \sum_k \lambda_{ik} (1 - \sum_j \eta_{ijk})$$

$$\frac{\partial \ell^*}{\partial \eta_{ijk}} = \frac{\sum_m [X_{mi} = j][Y_m = k]}{\eta_{ijk}} - \lambda_{ik}$$

$$\frac{\partial \ell^*}{\partial \eta_{ijk}} = 0 \Rightarrow \eta_{ijk} = \frac{N_{X_i=j, Y=k}}{\lambda_{ik}}$$

$$\sum_j \eta_{ijk} = 1 \Rightarrow \sum_j \frac{N_{X_i=j, Y=k}}{\lambda_{ik}} = 1 \Rightarrow \lambda_{ik} = \sum_j N_{X_i=j, Y=k} = N_{Y=k}$$

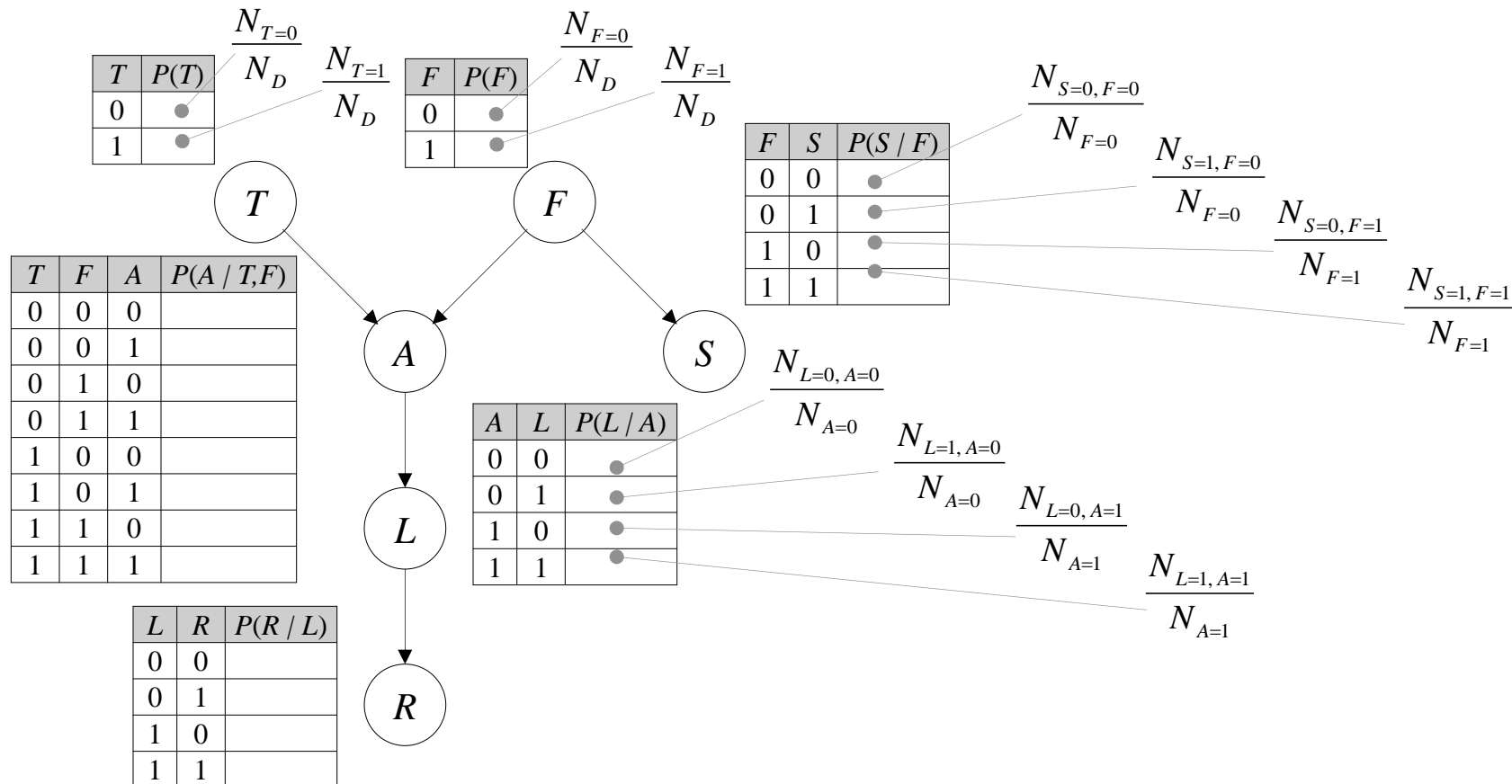
$$\eta_{ijk}^* = \frac{N_{X_i=j, Y=k}}{N_{Y=k}} \quad (\text{Maximum Likelihood Estimator of } \eta_{ijk})$$

# Learning CPTs for a graphical model

As Maximum Likelihood Estimation

Model: the graphical model of the *fire alarm* example, with CPTs as parameters

Observations: sequence of sets  $d_i$  values, from completely observed situations



# Bayesian learning

## ■ *Maximum a Posteriori Estimation (MAP)*

Instead of a *likelihood function*, the a posteriori probability is maximized

$$P(\theta | D) = \frac{P(D | \theta)P(\theta)}{P(D)} = \frac{P(D | \theta)P(\theta)}{\sum_{\theta} P(D | \theta)P(\theta)}$$

Which is equivalent to optimize:

$$P(D | \theta) P(\theta)$$

Advantages:

- Regularization: not all possible combinations of values might be present in  $D$
- A formula for incremental learning:  
*a priori* terms could represent what was known *before* observations  $D$

**Problem:**

- Which *prior* distribution  $P(\theta)$  ?

# Beta distribution

Gamma function ( $n$  integer  $> 0$ )

$$\Gamma(n) := (n-1)!$$

Beta function ( $\alpha$  and  $\beta$  integers  $> 0$ )

$$B(\alpha, \beta) := \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha + \beta)} = \frac{(\alpha-1)!(\beta-1)!}{(\alpha + \beta - 1)!}$$

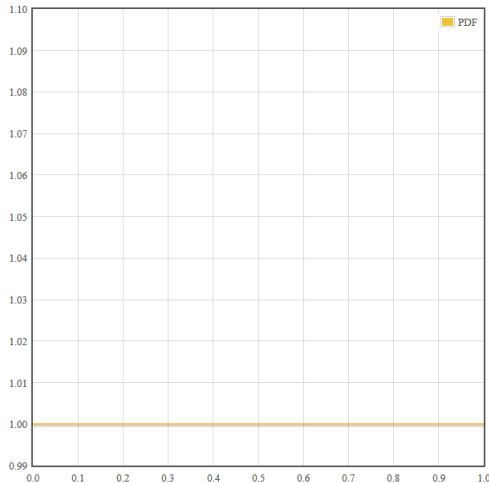
The definition is more complex when  $\alpha$  and  $\beta$  are not integers (see Wikipedia)

- Beta probability density function (pdf) ( $\alpha$  and  $\beta$  integers  $> 0$ )

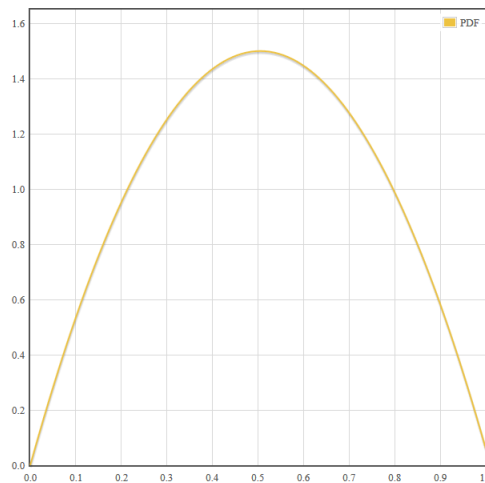
$$\text{Beta}(x; \alpha, \beta) := \frac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha, \beta)}$$

The maximum occurs at:  $x = \frac{\alpha-1}{\alpha+\beta-2}$

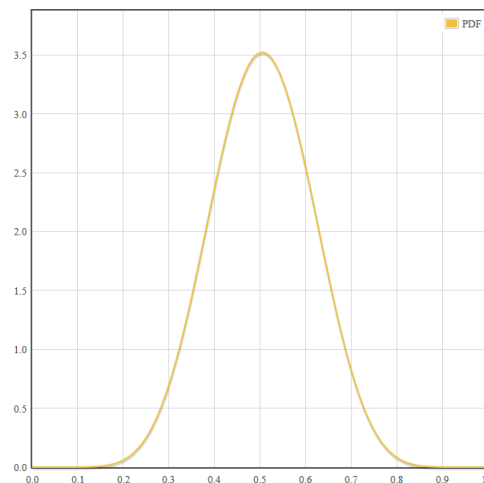
Beta( $x;1,1$ )



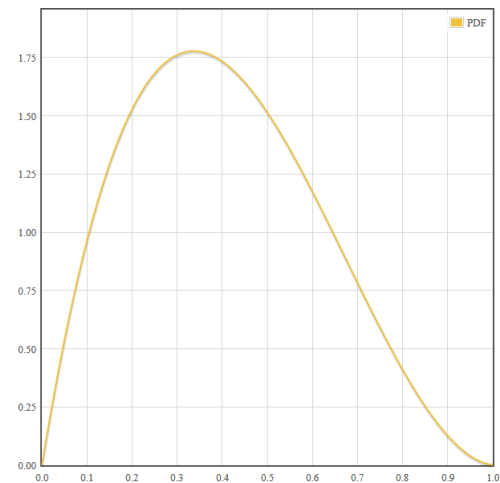
Beta( $x;2,2$ )



Beta( $x;10,10$ )



Beta( $x;2,3$ )



# Conjugate prior distributions

## Coin tossing

$$P(D_i | \theta) = \theta^{[X_i=1]} (1-\theta)^{[X_i=0]}$$

(a.k.a. the Bernoulli distribution)

## Coin tossing (repeated experiments)

$\alpha_D$  and  $\beta_D$  are the result counts (i.e. heads and tails)

$$P(D | \theta) = P(\{D_i\} | \theta) = \prod_i P(D_i | \theta) = \theta^{\alpha_D} (1-\theta)^{\beta_D}$$

## A posteriori probability with Beta prior

$\alpha_P$  and  $\beta_P$  are the hyperparameters of the prior

$$\begin{aligned} P(D | \theta) P(\theta) &= \theta^{\alpha_D} (1-\theta)^{\beta_D} \cdot \text{Beta}(\theta; \alpha_P, \beta_P) = \theta^{\alpha_D} (1-\theta)^{\beta_D} \cdot \frac{\theta^{\alpha_P-1} (1-\theta)^{\beta_P-1}}{\text{B}(\alpha_P, \beta_P)} \\ &= \frac{\theta^{\alpha_D+\alpha_P-1} (1-\theta)^{\beta_D+\beta_P-1}}{\text{B}(\alpha_P, \beta_P)} = \frac{\text{B}(\alpha_D + \alpha_P, \beta_D + \beta_P)}{\text{B}(\alpha_P, \beta_P)} \cdot \text{Beta}(\theta; \alpha_D + \alpha_P, \beta_D + \beta_P) \end{aligned}$$

Moral:

this factor is a positive number

$$P(D | \theta) P(\theta) \propto \text{Beta}(\theta; \alpha_D + \alpha_P, \beta_D + \beta_P)$$

Therefore

$$\theta_{MAP}^* = \arg \max_{\theta} \text{Beta}(\theta; \alpha_D + \alpha_P, \beta_D + \beta_P) = \frac{\alpha_D + \alpha_P - 1}{\alpha_D + \alpha_P + \beta_D + \beta_P - 2}$$

It is the same result as MLE but with the addition of  $\alpha_P + \beta_P - 2$  pseudo-observations

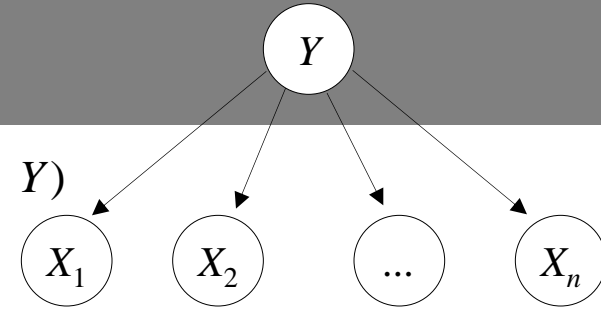
Being a **conjugate prior**  $P(\theta)$  of a distribution  $P(D | \theta)$

in the above sense

means that the posterior  $P(D | \theta) P(\theta)$  is in the same family of  $P(\theta)$

# Anti-spam filter

$$P(Y, \{X_i\}) = P(Y) \prod_{i=1}^n P(X_i | Y)$$



## ■ Maximum a Posteriori (MAP) Estimation

The adapted computations for:

$$\theta_{MAP}^* = \arg \max_{\theta} P(D | \theta) P(\theta)$$

yield:

$$\pi_k^* = \frac{\alpha_k + N_{Y=k} - 1}{\alpha_k + \beta_k + N_D - 2} \quad (\text{MAP Estimator of } \pi_k)$$

$$\eta_{ijk}^* = \frac{\alpha_{ijk} + N_{X_i=j, Y=k} - 1}{\alpha_{ijk} + \beta_{ijk} + N_{Y=k} - 2} \quad (\text{MAP Estimator of } \eta_{ijk})$$

where the

$$\alpha_k, \beta_k, \alpha_{ijk}, \beta_{ijk}$$

are the *hyperparameters* of the prior distribution representing the *pseudo-observations* made *before* the arrival of new, actual observations  $D$