

# Probability recall

Course of Machine Learning  
Master Degree in Computer Science  
University of Rome “Tor Vergata”  
a.a. 2024-2025

Giorgio Gambosi

## 1 Probability

### Discrete random variables

A discrete **random variable**  $X$  can take values from some finite or countably infinite set  $\mathcal{X}$ . A **probability mass function** (pmf) associates to each event  $X = x$  a probability  $p(X = x)$ .

#### Properties

- $0 \leq p(x) \leq 1$  for all  $x \in \mathcal{X}$
- $\sum_{x \in \mathcal{X}} p(x) = 1$

Note: we shall denote as  $x$  the event  $X = x$

### Discrete random variables

#### Joint and conditional probabilities

Given two events  $x, y$ , it is possible to define:

- the probability  $p(x, y) = p(x \wedge y)$  of their joint occurrence
- the conditional probability  $p(x|y)$  of  $x$  under the hypothesis that  $y$  has occurred

#### Union of events

Given two events  $x, y$ , the probability of  $x$  or  $y$  is defined as

$$p(x \vee y) = p(x) + p(y) - p(x, y)$$

in particular,

$$p(x \vee y) = p(x) + p(y)$$

The same definitions hold for probability distributions.

## Discrete random variables

### Product rule

The product rule relates joint and conditional probabilities

$$p(x, y) = p(x|y)p(y) = p(y|x)p(x)$$

where  $p(x)$  is the **marginal** probability.

In general,

$$\begin{aligned} p(x_1, \dots, x_n) &= p(x_2, \dots, x_n|x_1)p(x_1) \\ &= p(x_3, \dots, x_n|x_1, x_2)p(x_2|x_1)p(x_1) \\ &= \dots \\ &= p(x_n|x_1, \dots, x_{n-1})p(x_{n-1}|x_1 \dots x_{n-2}) \cdots p(x_2|x_1)p(x_1) \end{aligned}$$

## Discrete random variables

### Sum rule and marginalization

The sum rule relates the joint probability of two events  $x, y$  and the probability of one such events  $p(y)$  (or  $p(y)$ )

$$p(x) = \sum_{y \in \mathcal{Y}} p(x, y) = \sum_{y \in \mathcal{Y}} p(x|y)p(y)$$

Applying the sum rule to derive a marginal probability from a joint probability is usually called **marginalization**

## Discrete random variables

### Bayes rule

Since

$$\begin{aligned} p(x, y) &= p(x|y)p(y) \\ p(x, y) &= p(y|x)p(x) \\ p(y) &= \sum_{x \in \mathcal{X}} p(x, y) = \sum_{x \in \mathcal{X}} p(y|x)p(x) \end{aligned}$$

it results

$$p(x|y) = \frac{p(y|x)p(x)}{p(y)} = \frac{p(y|x)p(x)}{\sum_{x \in \mathcal{X}} p(y|x)p(x)}$$

## Terminology

- $p(x)$ : **Prior** probability of  $x$  (before knowing that  $y$  occurred)
- $p(x|y)$ : **Posterior** of  $x$  (if  $y$  has occurred)
- $p(y|x)$ : **Likelihood** of  $y$  given  $x$
- $p(y)$ : **Evidence** of  $y$

## Independence

### Definition

Two random variables  $X, Y$  are **independent** ( $X \perp\!\!\!\perp Y$ ) if their joint probability is equal to the product of their marginals

$$p(x, y) = p(x)p(y)$$

or, equivalently,

$$p(x|y) = p(x) \quad p(y|x) = p(y)$$

The condition  $p(x|y) = p(x)$ , in particular, states that, if two variables are independent, knowing the value of one does not add any knowledge about the other one.

## Independence

### Conditional independence

Two random variables  $X, Y$  are **conditionally independent** w.r.t. a third r.v.  $Z$  ( $X \perp\!\!\!\perp Y | Z$ ) if

$$p(x, y|z) = p(x|z)p(y|z)$$

Conditional independence does not imply (absolute) independence, and vice versa.

## Continuous random variables

A continuous random variable  $X$  can take values from a continuous infinite set  $\mathcal{X}$ . Its probability is defined as **cumulative distribution function** (cdf)  $F(x) = p(X \leq x)$ .

The probability that  $X$  is in an interval  $(a, b]$  is then  $p(a < X \leq b) = F(b) - F(a)$ .

### Probability density function

The probability density function (pdf) is defined as  $f(x) = \frac{dF(x)}{dx}$ . As a consequence,

$$p(a < X \leq b) = \int_a^b f(x)dx$$

and

$$p(x < X \leq x + dx) \approx f(x)dx$$

for a sufficiently small  $dx$ .

### Sum rule and continuous random variables

In the case of continuous random variables, their probability density functions relate as follows.

$$f(x) = \int_{\mathcal{Y}} f(x, y)dy = \int_{y \in \mathcal{Y}} p(x|y)p(y)dy$$

## Expectation

### Definition

Let  $x$  be a discrete random variable with distribution  $p(x)$ , and let  $g : \mathbb{R} \mapsto \mathbb{R}$  be any function: the expectation of  $g(x)$  w.r.t.  $p(x)$  is

$$E_p[g(x)] = \sum_{x \in V_x} g(x)p(x)$$

If  $x$  is a continuous r.v., with probability density  $f(x)$ , then

$$E_f[g(x)] = \int_{-\infty}^{\infty} g(x)f(x)dx$$

### Mean value

Particular case:  $g(x) = x$

$$E_p[x] = \sum_{x \in V_x} xp(x) \quad E_f[x] = \int_{-\infty}^{\infty} xf(x)dx$$

## Elementary properties of expectation

- $E[a] = a$  for each  $a \in \mathbb{R}$
- $E[af(x)] = aE[f(x)]$  for each  $a \in \mathbb{R}$
- $E[f(x) + g(x)] = E[f(x)] + E[g(x)]$

## Variance

### Definition

$$\text{Var}[X] = E[(x - E[x])^2]$$

We may easily derive:

$$\begin{aligned} E[(x - E[x])^2] &= E[x^2 - 2E[x]x + E[x]^2] \\ &= E[x^2] - 2E[x]E[x] + E[x]^2 \\ &= E[x^2] - E[x]^2 \end{aligned}$$

Some elementary properties:

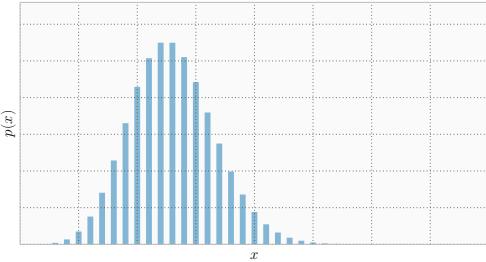
- $\text{Var}[a] = 0$  for each  $a \in \mathbb{R}$
- $\text{Var}[af(x)] = a^2\text{Var}[f(x)]$  for each  $a \in \mathbb{R}$

## Probability distributions

### Probability distribution

Given a discrete random variable  $X \in V_X$ , the corresponding **probability distribution** is a function  $p(x) = P(X = x)$  such that

- $0 \leq p(x) \leq 1$
- $\sum_{x \in V_X} p(x) = 1$
- $\sum_{x \in A} p(x) = P(x \in A)$ , with  $A \subseteq V_X$

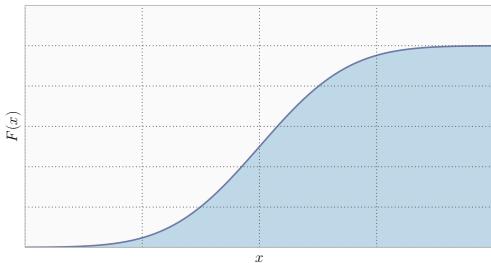


## Some definitions

### Cumulative distribution

Given a continuous random variable  $X \in \mathbb{R}$ , the corresponding **cumulative probability distribution** is a function  $F(x) = P(X \leq x)$  such that:

- $0 \leq F(x) \leq 1$
- $\lim_{x \rightarrow -\infty} F(x) = 0$
- $\lim_{x \rightarrow \infty} F(x) = 1$
- $x \leq y \Rightarrow F(x) \leq F(y)$



## Some definitions

### Probability density

Given a continuous random variable  $X \in \mathbb{R}$  with derivable cumulative distribution  $F(x)$ , the **probability density** is defined as

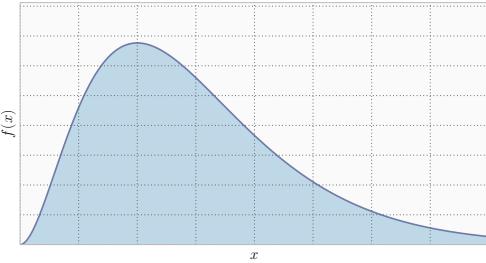
$$f(x) = \frac{dF(x)}{dx}$$

By definition of derivative, for a sufficiently small  $\Delta x$ ,

$$Pr(x \leq X \leq x + \Delta x) \approx f(x)\Delta x$$

The following properties hold:

- $f(x) \geq 0$
- $\int_{-\infty}^{\infty} f(x)dx = 1$
- $\int_{x \in A} f(x)dx = P(X \in A)$



## Bernoulli distribution

### Definition

Let  $x \in \{0, 1\}$ , then  $x \sim \text{Bernoulli}(p)$ , with  $0 \leq p \leq 1$ , if

$$p(x) = \begin{cases} p & \text{se } x = 1 \\ 1 - p & \text{se } x = 0 \end{cases}$$

or, equivalently,

$$p(x) = p^x(1 - p)^{1-x}$$

Probability that, given a coin with head (H) probability  $p$  (and tail probability (T)  $1 - p$ ), a coin toss result into  $x \in \{H, T\}$ .

### Mean and variance

$$E[x] = p \quad Var[x] = p(1 - p)$$

## Extension to multiple outcomes

Assume  $k$  possible outcomes (for example a die toss).

In this case, a generalization of the Bernoulli distribution is considered, usually named **categorical** distribution.

$$p(x) = \prod_{j=1}^k p_j^{x_j}$$

where  $(p_1, \dots, p_k)$  are the probabilities of the different outcomes ( $\sum_{j=1}^k p_j = 1$ ) and  $x_j = 1$  iff the  $j$ -th outcome occurs.

## Binomial distribution

### Definition

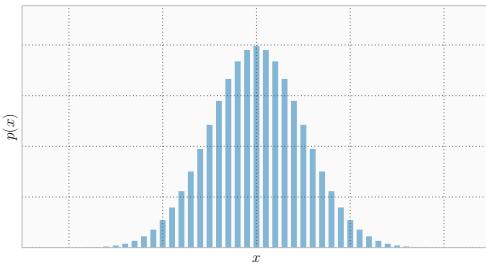
Let  $x \in \mathbb{N}$ , then  $x \sim \text{Binomial}(n, p)$ , with  $0 \leq p \leq 1$ , if

$$p(x) = \binom{n}{x} p^x (1-p)^{n-x} = \frac{n!}{x!(n-x)!} p^x (1-p)^{n-x}$$

Probability that, given a coin with head (H) probability  $p$ , a sequence of  $n$  independent coin tosses result into  $x$  heads.

### Mean and variance

$$\begin{aligned} E[x] &= np \\ \text{Var}[x] &= np(1-p) \end{aligned}$$



## Poisson distribution

### Definition

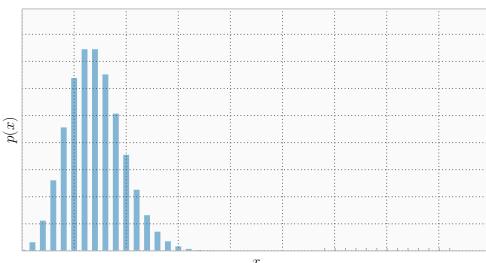
Let  $x_i \in \mathbb{N}$ , then  $x \sim \text{Poisson}(\lambda)$ , with  $\lambda > 0$ , if

$$p(x) = e^{-\lambda} \frac{\lambda^x}{x!}$$

Probability that an event with average frequency  $\lambda$  occurs  $x$  times in the next time unit.

### Mean and variance

$$\begin{aligned} E[x] &= \lambda \\ \text{Var}[x] &= \lambda \end{aligned}$$



## Normal (gaussian) distribution

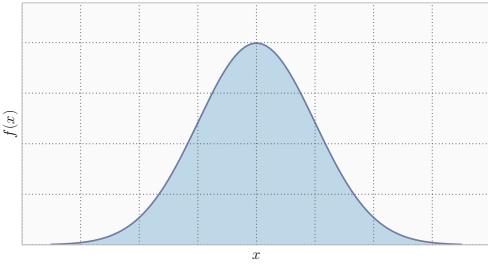
### Definition

Let  $x \in \mathbb{R}$ , then  $x \sim \text{Normal}(\mu, \sigma^2)$ , with  $\mu, \sigma \in \mathbb{R}, \sigma \geq 0$ , if

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

### Mean and variance

$$\begin{aligned} E[x] &= \mu \\ \text{Var}[x] &= \sigma^2 \end{aligned}$$



## Beta distribution

### Definition

Let  $x \in [0, 1]$ , then  $x \sim \text{Beta}(\alpha, \beta)$ , with  $\alpha, \beta > 0$ , if

$$f(x) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1} (1-x)^{\beta-1}$$

where

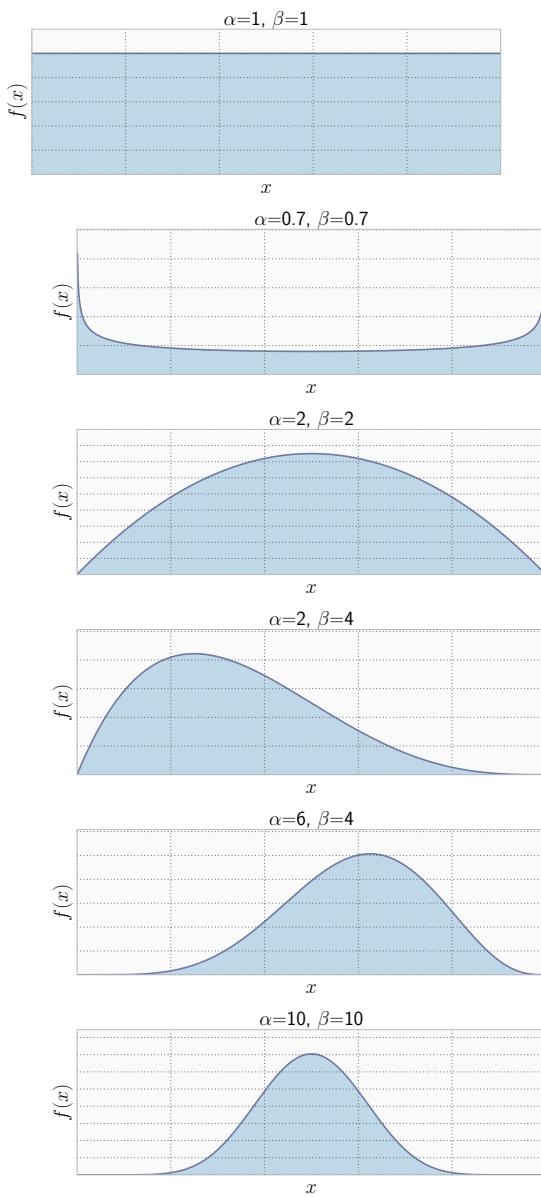
$$\Gamma(x) = \int_0^\infty u^{x-1} e^u du$$

is a generalization of the factorial to the real field  $\mathbb{R}$ : in particular,  $\Gamma(n) = (n-1)!$  if  $n \in \mathbb{N}$

### Mean and variance

$$\begin{aligned} E[x] &= \frac{\beta}{\alpha + \beta} \\ \text{Var}[x] &= \frac{\alpha\beta}{(\alpha + \beta)^2(\alpha + \beta + 1)} \end{aligned}$$

## Beta distribution



## Multivariate distributions

Definition for  $k = 2$  discrete variables

Given two discrete r.v.  $X, Y$ , their **joint** distribution is

$$p(x, y) = P(X = x, Y = y)$$

The following properties hold:

1.  $0 \leq p(x, y) \leq 1$
2.  $\sum_{x \in V_X} \sum_{y \in V_Y} p(x, y) = 1$

## Multivariate distributions

### Definition for $k = 2$ variables

Given two continuous r.v.  $X, Y$ , their cumulative joint distribution is defined as

$$F(x, y) = P(X \leq x, Y \leq y)$$

The following properties hold:

1.  $0 \leq F(x, y) \leq 1$
2.  $\lim_{x,y \rightarrow \infty} F(x, y) = 1$
3.  $\lim_{x,y \rightarrow -\infty} F(x, y) = 0$

If  $F(x, y)$  is derivable everywhere w.r.t. both  $x$  and  $y$ , **joint probability density** is

$$f(x, y) = \frac{\partial^2 F(x, y)}{\partial x \partial y}$$

The following property derives

$$\int \int_{(x,y) \in A} f(x, y) dx dy = P((X, Y) \in A)$$

## Covariance

### Definition

$$\text{Cov}[X, Y] = E[(X - E[X])(Y - E[Y])]$$

As for the variance, we may derive

$$\begin{aligned} \text{Cov}[X, Y] &= E[(X - E[X])(Y - E[Y])] \\ &= E[XY - XE[Y] - YE[X] + E[X]E[Y]] \\ &= E[XY] - E[X]E[Y] - E[Y]E[X] + E[E[X]E[Y]] \\ &= E[XY] - E[X]E[Y] \end{aligned}$$

Moreover, the following properties hold:

1.  $\text{Var}[X + Y] = \text{Var}[X] + \text{Var}[Y] + 2\text{Cov}[X, Y]$
2. If  $X \perp\!\!\!\perp Y$  then  $\text{Cov}[X, Y] = 0$

## Random vectors

### Definition

Let  $X_1, X_2, \dots, X_n$  be a set of r.v.: we may then define a random vector as

$$\mathbf{x} = \begin{pmatrix} X_1 \\ \vdots \\ X_n \end{pmatrix}$$

## Expectation and random vectors

### Definition

Let  $g : \mathbb{R}^n \mapsto \mathbb{R}^m$  be any function. It may be considered as a vector of functions

$$g(\mathbf{x}) = \begin{pmatrix} g_1(\mathbf{x}) \\ \vdots \\ g_m(\mathbf{x}) \end{pmatrix}$$

where  $\mathbf{x} \in \mathbb{R}^n$ .

The expectation of  $g$  is the vector of the expectations of all functions  $g_i$ ,

$$E[g(\mathbf{x})] = \begin{pmatrix} E[g_1(\mathbf{x})] \\ \vdots \\ E[g_m(\mathbf{x})] \end{pmatrix}$$

## Covariance matrix

### Definition

Let  $\mathbf{x} \in \mathbb{R}^n$  be a random vector: its covariance matrix  $\Sigma$  is a matrix  $n \times n$  such that, for each  $1 \leq i, j \leq n$ ,  $\Sigma_{ij} = \text{Cov}[X_i, X_j] = E[(X_i - \mu_i)(X_j - \mu_j)]$ , where  $\mu_i = E[X_i]$ ,  $\mu_j = E[X_j]$ .

Hence,

$$\begin{aligned} \Sigma &= \begin{bmatrix} \text{Cov}[X_1, X_1] & \text{Cov}[X_1, X_2] & \cdots & \text{Cov}[X_1, X_n] \\ \text{Cov}[X_2, X_1] & \text{Cov}[X_2, X_2] & \cdots & \text{Cov}[X_2, X_n] \\ \vdots & \vdots & \ddots & \vdots \\ \text{Cov}[X_n, X_1] & \text{Cov}[X_n, X_2] & \cdots & \text{Cov}[X_n, X_n] \end{bmatrix} \\ &= \begin{bmatrix} \text{Var}[X_1] & \cdots & \text{Cov}[X_1, X_n] \\ \vdots & \ddots & \vdots \\ \text{Cov}[X_n, X_1] & \cdots & \text{Var}[X_n] \end{bmatrix} \end{aligned}$$

## Covariance matrix

By definition of covariance,

$$\begin{aligned} \Sigma &= \begin{bmatrix} E[X_1^2] - E[X_1]^2 & \cdots & E[X_1 X_n] - E[X_1]E[X_n] \\ \vdots & \ddots & \vdots \\ E[X_n X_1] - E[X_n]E[X_1] & \cdots & E[X_n^2] - E[X_n]E[X_n] \end{bmatrix} \\ &= E[\mathbf{XX}^T] - \boldsymbol{\mu}\boldsymbol{\mu}^T \end{aligned}$$

where  $\boldsymbol{\mu} = (\mu_1, \dots, \mu_n)^T$  is the vector of expectations of the random variables  $X_1, \dots, X_n$ .

### Properties

The covariance matrix is necessarily:

- semidefinite positive: that is,  $\mathbf{z}^T \Sigma \mathbf{z} \geq 0$  for any  $\mathbf{z} \in \mathbb{R}^n$
- symmetric:  $\text{Cov}[X_i, X_j] = \text{Cov}[X_j, X_i]$  for  $1 \leq i, j \leq n$

## Correlation

For any pair of r.v.  $X, Y$ , the Pearson correlation coefficient is defined as

$$\rho_{X,Y} = \frac{\text{Cov}[X, Y]}{\sqrt{\text{Var}[X]\text{Var}[Y]}}$$

Note that, if  $Y = aX + b$  for some pair  $a, b$ , then

$$\text{Cov}[X, Y] = E[(X - \mu)(aX + b - a\mu - b)] = E[a(X - \mu)^2] = a\text{Var}[X]$$

and, since

$$\text{Var}[Y] = (aX - a\mu)^2 = a^2\text{Var}[X]$$

it results  $\rho_{X,Y} = 1$ . As a corollary,  $\rho_{X,X} = 1$ .

Observe that if  $X$  and  $Y$  are independent,  $p(X, Y) = p(X)p(Y)$ : as a consequence,  $\text{Cov}[X, Y] = 0$  and  $\rho_{X,Y} = 0$ . That is, independent variables have null covariance and correlation.

The contrary is not true: null correlation does not imply independence: see for example  $X$  uniform in  $[-1, 1]$  and  $Y = X^2$ .

## Correlation matrix

The correlation matrix of  $(X_1, \dots, X_n)^T$  is defined as

$$\Sigma = \begin{bmatrix} \rho_{X_1, X_1} & \rho_{X_1, X_2} & \cdots & \rho_{X_1, X_n} \\ \vdots & \ddots & & \vdots \\ \rho_{X_n, X_1} & \rho_{X_n, X_2} & \cdots & \rho_{X_n, X_n} \\ 1 & \rho_{X_1, X_2} & \cdots & \rho_{X_1, X_n} \\ \vdots & \ddots & & \vdots \\ \rho_{X_n, X_1} & \rho_{X_n, X_2} & \cdots & 1 \end{bmatrix}$$

## Multinomial distribution

### Definition

Let  $x_i \in \mathbb{N}$  for  $i = 1, \dots, k$ , then  $(x_1, \dots, x_k) \sim \text{Mult}(n, p_1, \dots, p_k)$  with  $0 \leq p \leq 1$ , if

$$p(x_1, \dots, x_k) = \frac{n!}{x_1! \dots x_k!} \prod_{i=1}^k p_i^{x_i} \quad \text{con } \sum_{i=1}^k x_i = n$$

Generalization of the binomial distribution to  $k \geq 2$  possible toss results  $t_1, \dots, t_k$  with probabilities  $p_1, \dots, p_k$  ( $\sum_{i=1}^k p_i = 1$ ).

Probability that in a sequence of  $n$  independent tosses  $p_1, \dots, p_k$ , exactly  $x_i$  tosses have result  $t_i$  ( $i = 1, \dots, k$ ).

### Mean and variance

$$E[x_i] = np_i \quad \text{Var}[x_i] = np_i(1 - p_i) \quad i = 1, \dots, k$$

## Dirichlet distribution

### Definition

Let  $x_i \in [0, 1]$  for  $i = 1, \dots, k$ , then  $(x_1, \dots, x_k) \sim \text{Dirichlet}(\alpha_1, \alpha_2, \dots, \alpha_k)$  if

$$f(x_1, \dots, x_k) = \frac{\Gamma(\sum_{i=1}^k \alpha_i)}{\prod_{i=1}^k \Gamma(\alpha_i)} \prod_{i=1}^k x_i^{\alpha_i-1} = \frac{1}{\Delta(\alpha_1, \dots, \alpha_k)} \prod_{i=1}^k x_i^{\alpha_i-1}$$

with  $\sum_{i=1}^k x_i = 1$ .

Generalization of the Beta distribution to the multinomial case  $k \geq 2$ .

A random variable  $\phi = (\phi_1, \dots, \phi_K)$  with Dirichlet distribution takes values on the  $K - 1$  dimensional simplex (set of points  $\mathbf{x} \in \mathbb{R}^K$  such that  $x_i \geq 0$  for  $i = 1, \dots, K$  and  $\sum_{i=1}^K x_i = 1$ )

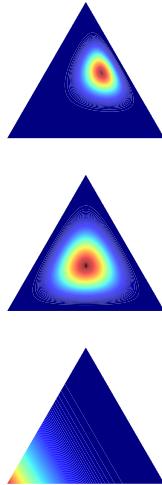
### Mean and variance

$$E[x_i] = \frac{\alpha_i}{\alpha_0} \quad \text{Var}[x_i] = \frac{\alpha_i(\alpha_0 - \alpha_i)}{\alpha_0^2(\alpha_0 + 1)} \quad i = 1, \dots, k$$

with  $\alpha_0 = \sum_{j=1}^k \alpha_j$

## Dirichlet distribution

Examples of Dirichlet distributions with  $k = 3$



## Dirichlet distribution

### Symmetric Dirichlet distribution

Particular case, where  $\alpha_i = \alpha$  for  $i = 1, \dots, K$

$$p(\phi_1, \dots, \phi_K | \alpha, K) = \text{Dir}(\phi | \alpha, K) = \frac{\Gamma(K\alpha)}{\Gamma(\alpha)^K} \prod_{i=1}^K \phi_i^{\alpha-1} = \frac{1}{\Delta_K(\alpha)} \prod_{i=1}^K \phi_i^{\alpha-1}$$

### Mean and variance

In this case,

$$E[x_i] = \frac{1}{K} \quad \text{Var}[x_i] = \frac{K-1}{K^2(\alpha+1)} \quad i = 1, \dots, K$$

## 2 The normal distribution

### Gaussian distribution

- Properties
  - Analytically tractable
  - Completely specified by the first two moments
  - A number of processes are asymptotically gaussian (theorem of the Central Limit)
  - Linear transformation of gaussians result in a gaussian

### Univariate gaussian

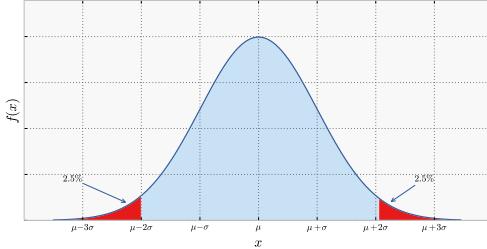
For  $x \in \mathbb{R}$ :

$$\begin{aligned} p(x) &= \mathcal{N}(\mu, \sigma^2) \\ &= \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \end{aligned}$$

with

$$\begin{aligned} \mu &= E[x] = \int_{-\infty}^{\infty} xp(x)dx \\ \sigma^2 &= E[(x - \mu)^2] = \int_{-\infty}^{\infty} (x - \mu)^2 p(x)dx \end{aligned}$$

### Univariate gaussian



A univariate gaussian distribution has about 95% of its probability in the interval  $|x - \mu| \geq 2\sigma$ .

### Multivariate gaussian

For  $\mathbf{x} \in \mathbb{R}^d$ :

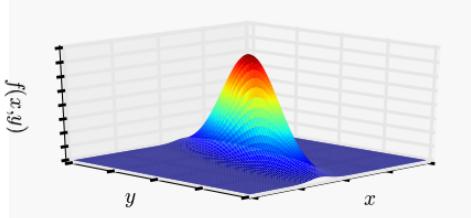
$$\begin{aligned} p(\mathbf{x}) &= \mathcal{N}(\boldsymbol{\mu}, \Sigma) \\ &= \frac{1}{(2\pi)^{d/2} |\Sigma|^{1/2}} e^{-\frac{1}{2} (\mathbf{x} - \boldsymbol{\mu})^T \Sigma^{-1} (\mathbf{x} - \boldsymbol{\mu})} \end{aligned}$$

where

$$\begin{aligned} \boldsymbol{\mu} &= E[\mathbf{x}] = \int \mathbf{x} p(\mathbf{x}) d\mathbf{x} \\ \Sigma &= E[(\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})^T] = \int (\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})^T p(\mathbf{x}) d\mathbf{x} \end{aligned}$$

## Multivariate gaussian

- $\mu$ : expectation (vector of size  $d$ )
- $\Sigma$ : matrix  $d \times d$  of covariance.  $\sigma_{ij} = E[(X_i - \mu_i)(X_j - \mu_j)]$



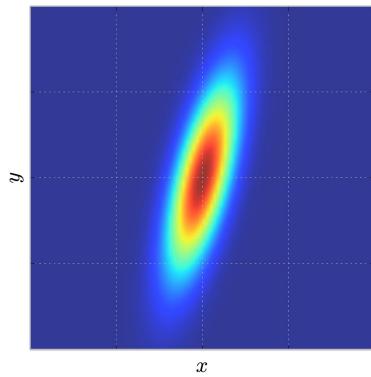
## Multivariate gaussian

### Mahalanobis distance

- Probability is a function of  $\mathbf{x}$  through the **quadratic form**

$$\Delta^2 = (\mathbf{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu})$$

- $\Delta$  is the **Mahalanobis distance** from  $\boldsymbol{\mu}$  to  $\mathbf{x}$ : it reduces to the euclidean distance if  $\boldsymbol{\Sigma} = \mathbf{I}$ .
- Constant probability on the curves (ellipsis) at constant  $\Delta$ .



## Multivariate gaussian

In general,

$$\mathbf{x}^T \mathbf{A} \mathbf{x} = (\mathbf{x}^T \mathbf{A} \mathbf{x})^T = \mathbf{x}^T \mathbf{A}^T \mathbf{x}$$

this implies that

$$\mathbf{x}^T \mathbf{A} \mathbf{x} = \frac{1}{2} \mathbf{x}^T \mathbf{A} \mathbf{x} + \frac{1}{2} \mathbf{x}^T \mathbf{A}^T \mathbf{x} = \mathbf{x}^T \left( \frac{1}{2} \mathbf{A} + \frac{1}{2} \mathbf{A}^T \right) \mathbf{x}$$

- $\mathbf{A} + \mathbf{A}^T$  is necessarily symmetric, as a consequence,  $\boldsymbol{\Sigma}$  is symmetric
- as a consequence, its inverse  $\boldsymbol{\Sigma}^{-1}$  does exist.

## Diagonal covariance matrix

Assume a diagonal covariance matrix:

$$\Sigma = \begin{bmatrix} \sigma_1^2 & 0 & \cdots & 0 \\ 0 & \sigma_2^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_n^2 \end{bmatrix}$$

then,  $|\Sigma| = \sigma_1^2 \sigma_2^2 \dots \sigma_n^2$  and

$$\Sigma^{-1} = \begin{bmatrix} \frac{1}{\sigma_1^2} & 0 & \cdots & 0 \\ 0 & \frac{1}{\sigma_2^2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \frac{1}{\sigma_n^2} \end{bmatrix}$$

## Diagonal covariance matrix

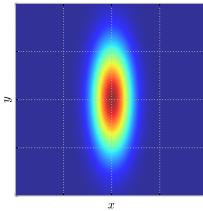
Easy to verify that

$$(\mathbf{x} - \boldsymbol{\mu})^T \Sigma^{-1} (\mathbf{x} - \boldsymbol{\mu}) = \sum_{i=1}^n \frac{(x_i - \mu_i)^2}{\sigma_i^2}$$

and

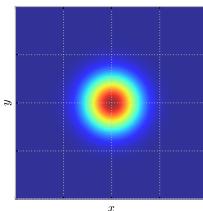
$$f(\mathbf{x} | \boldsymbol{\mu}, \Sigma) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi}\sigma_i} \exp\left(-\frac{1}{2} \frac{(x_i - \mu_i)^2}{\sigma_i^2}\right)$$

The multivariate distribution turns out to be the product of  $d$  univariate gaussians, one for each coordinate  $x_i$ .



## Identity covariance matrix

The distribution is the product of  $d$  “copies” of the same univariate gaussian, one copy for each coordinate  $x_i$ .



## Spectral properties of $\Sigma$

$\Sigma$  is real and symmetric: then,

1. all its eigenvalues  $\lambda_i$  are in  $\mathbb{R}$

2. there exists a corresponding set of orthonormal eigenvectors  $i$  (i.e. such that  $(i^T j = 1$  if  $i = j$  and 0 otherwise)

Let us define the  $d \times d$  matrix  $\mathbf{U}$  whose columns correspond to the orthonormal eigenvectors

$$\mathbf{U} = \begin{pmatrix} & & \\ | & \dots & | \\ 1 & & 2 \\ | & & | \end{pmatrix} d$$

and the diagonal  $d \times d$  matrix  $\mathbf{\Lambda}$  with eigenvalues on the diagonal

$$\mathbf{\Lambda} = \begin{bmatrix} \lambda_1 & & & & \\ & \lambda_2 & & 0 & \\ & & \lambda_3 & & \\ 0 & & & \ddots & \\ & & & & \lambda_d \end{bmatrix}$$

## Multivariate gaussian

### Decomposition of $\Sigma$

By the definition of  $\mathbf{U}$  and  $\mathbf{\Lambda}$ , and since  $\Sigma i = i\lambda_i$  for all  $i = 1, \dots, d$ , we may write

$$\Sigma \mathbf{U} = \mathbf{U} \mathbf{\Lambda}$$

Since the eigenvectors  $u_i$  are orthonormal,  $\mathbf{U}^{-1} = \mathbf{U}^T$  by the properties of orthonormal matrices: as a consequence,

$$\Sigma = \mathbf{U} \mathbf{\Lambda} \mathbf{U}^{-1} = \mathbf{U} \mathbf{\Lambda} \mathbf{U}^T = \sum_{i=1}^d \lambda_i \mathbf{u}_i \mathbf{u}_i^T$$

Then, its inverse matrix is a diagonal matrix itself

$$\Sigma^{-1} = \sum_{i=1}^d \frac{1}{\lambda_i} \mathbf{u}_i \mathbf{u}_i^T$$

## Multivariate gaussian

### Density as a function of eigenvalues and eigenvectors

As shown before,

$$\begin{aligned} \Delta^2 &= (\mathbf{x} - \boldsymbol{\mu})^T \Sigma^{-1} (\mathbf{x} - \boldsymbol{\mu}) = (\mathbf{x} - \boldsymbol{\mu})^T \sum_{i=1}^d \frac{1}{\lambda_i} i i^T (\mathbf{x} - \boldsymbol{\mu}) \\ &= \sum_{i=1}^d \frac{1}{\lambda_i} (\mathbf{x} - \boldsymbol{\mu})^T i i^T (\mathbf{x} - \boldsymbol{\mu}) = \sum_{i=1}^d \frac{1}{\lambda_i} (i^T (\mathbf{x} - \boldsymbol{\mu}))^T i^T (\mathbf{x} - \boldsymbol{\mu}) \\ &= \sum_{i=1}^d \frac{(i^T (\mathbf{x} - \boldsymbol{\mu}))^2}{\lambda_i} \end{aligned}$$

Let  $y_i = i^T(\mathbf{x} - \boldsymbol{\mu})$ : then

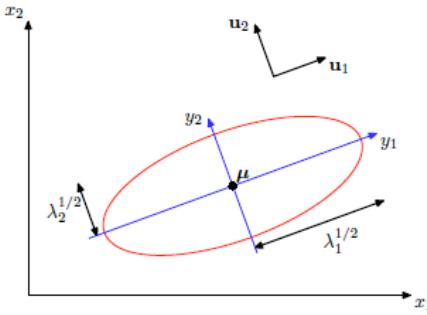
$$(\mathbf{x} - \boldsymbol{\mu})^T \Sigma^{-1} (\mathbf{x} - \boldsymbol{\mu}) = \sum_{i=1}^n \frac{y_i^2}{\lambda_i}$$

and

$$f(\mathbf{x} | \boldsymbol{\mu}, \Sigma) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi\lambda_i}} \exp\left(-\frac{1}{2} \frac{y_i^2}{\lambda_i}\right)$$

### Multivariate gaussian

$y_i$  is the scalar product of  $\mathbf{x} - \boldsymbol{\mu}$  and the  $i$ -th eigenvector  $i$ , that is the length of the projection of  $\mathbf{x} - \boldsymbol{\mu}$  along the direction of the eigenvector. Since eigenvectors are orthonormal, they are the basis of a new space, and for each vector  $\mathbf{x} = (x_1, \dots, x_d)$ , the values  $(y_1, \dots, y_d)$  are the coordinates of  $\mathbf{x}$  in the eigenvector space.



Eigenvectors of  $\Sigma$  correspond to the axes of the distribution; each eigenvalue is a scale factor along the axis of the corresponding eigenvector.

### Linear transformations

Let  $\mathbf{x} \in \mathbb{R}^d$ ,  $\mathbf{A} \in \mathbb{R}^{d \times k}$ ,  $\mathbf{y} = \mathbf{A}^T \mathbf{x} \in \mathbb{R}^k$ : then, if  $\mathbf{x}$  is normally distributed, so is  $\mathbf{y}$ .

In particular, if the distribution of  $\mathbf{x}$  has mean  $\boldsymbol{\mu}$  and covariance matrix  $\Sigma$ , the distribution of  $\mathbf{y}$  has mean  $\mathbf{A}^T \boldsymbol{\mu}$  and covariance matrix  $\mathbf{A}^T \Sigma \mathbf{A}$ .

$$\mathbf{x} \sim \mathcal{N}(\boldsymbol{\mu}, \Sigma) \Rightarrow \mathbf{y} \sim \mathcal{N}(\mathbf{A}^T \boldsymbol{\mu}, \mathbf{A}^T \Sigma \mathbf{A})$$

### Marginal and conditional of a joint gaussian

Let  $\mathbf{x}_1 \in \mathbb{R}^h$ ,  $\mathbf{x}_2 \in \mathbb{R}^k$  be such that  $\begin{bmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \end{bmatrix} \sim \mathcal{N}(\boldsymbol{\mu}, \Sigma)$  and let

$$\bullet \quad \boldsymbol{\mu} = \begin{bmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{bmatrix} \text{ with } \boldsymbol{\mu}_1 \in \mathbb{R}^h, \boldsymbol{\mu}_2 \in \mathbb{R}^k$$

$$\bullet \quad \Sigma = \left[ \begin{array}{c|c} \Sigma_{11} & \Sigma_{12} \\ \hline \Sigma_{21} & \Sigma_{22} \end{array} \right] \text{ with } \Sigma_{11} \in \mathbb{R}^{h \times h}, \Sigma_{12} \in \mathbb{R}^{h \times k}, \Sigma_{21} \in \mathbb{R}^{k \times h}, \Sigma_{22} \in \mathbb{R}^{k \times k}$$

then

- the marginal distribution of  $\mathbf{x}_1$  is  $\mathbf{x}_1 \sim \mathcal{N}(\boldsymbol{\mu}_1, \Sigma_{11})$

- the conditional distribution of  $\mathbf{x}_1$  given  $\mathbf{x}_2$  is  $\mathbf{x}_1|\mathbf{x}_2 \sim \mathcal{N}(\boldsymbol{\mu}_{1|2}, \Sigma_{1|2})$  with

$$\begin{aligned}\boldsymbol{\mu}_{1|2} &= \boldsymbol{\mu}_1 - \Sigma_{12}\Sigma_{22}^{-1}(\mathbf{x}_2 - \boldsymbol{\mu}_2) \\ \Sigma_{1|2} &= \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}\end{aligned}$$

## Bayes' formula and gaussians

Let  $\mathbf{x}, \mathbf{y}$  be such that

$$\mathbf{x} \sim \mathcal{N}(\boldsymbol{\mu}, \Sigma_1) \quad \text{and} \quad \mathbf{y}|\mathbf{x} \sim \mathcal{N}(A\mathbf{x} + \mathbf{b}, \Sigma_2)$$

That is, the marginal distribution of  $\mathbf{x}$  (the prior) is a gaussian and the conditional distribution of  $\mathbf{y}$  w.r.t.  $\mathbf{x}$  (the likelihood) is also a gaussian with (conditional) mean given by a linear combination on  $\mathbf{x}$ . Then, both the the conditional distribution of  $\mathbf{x}$  w.r.t.  $\mathbf{y}$  (the posterior) and the marginal distribution of  $\mathbf{y}$  (the evidence) are gaussian.

$$\begin{aligned}\mathbf{y} &\sim \mathcal{N}(A\boldsymbol{\mu} + \mathbf{b}, \Sigma_2 + A\Sigma_1 A^T) \\ \mathbf{x}|\mathbf{y} &\sim \mathcal{N}(\hat{\boldsymbol{\mu}}, \hat{\Sigma})\end{aligned}$$

where

$$\begin{aligned}\hat{\boldsymbol{\mu}} &= (\Sigma_1^{-1} + A^T\Sigma_2^{-1}A)^{-1}(A^T\Sigma_2^{-1}(\mathbf{y} - \mathbf{b}) + \Sigma_1^{-1}\boldsymbol{\mu}) \\ \hat{\Sigma} &= (\Sigma_1^{-1} + A^T\Sigma_2^{-1}A)^{-1}\end{aligned}$$

## 3 Bayesian statistics

### Bayesian statistics

#### Classical (frequentist) statistics

- Interpretation of probability as frequency of an event over a sufficiently long sequence of reproducible experiments.
- Parameters seen as constants to determine

#### Bayesian statistics

- Interpretation of probability as degree of belief that an event may occur.
- Parameters seen as random variables

### Bayes' rule

Cornerstone of bayesian statistics is Bayes' rule

$$p(X = x|\Theta = \theta) = \frac{p(\Theta = \theta|X = x)p(X = x)}{p(\Theta = \theta)}$$

Given two random variables  $X, \Theta$ , it relates the conditional probabilities  $p(X = x|\Theta = \theta)$  and  $p(\Theta = \theta|X = x)$ .

## Bayesian inference

Given an observed dataset  $\mathbf{X}$  and a family of probability distributions  $p(x|\Theta)$  with parameter  $\Theta$  (a probabilistic model), we wish to find the parameter value which best allows to describe  $\mathbf{X}$  through the model.

In the bayesian framework, we deal with the distribution probability  $p(\Theta)$  of the parameter  $\Theta$  considered here as a random variable. Bayes' rule states that

$$p(\Theta|\mathbf{X}) = \frac{p(\mathbf{X}|\Theta)p(\Theta)}{p(\mathbf{X})}$$

## Bayesian inference

### Interpretation

- $p(\Theta)$  stands as the knowledge available about  $\Theta$  before  $\mathbf{X}$  is observed (a.k.a. **prior distribution**)
- $p(\Theta|\mathbf{X})$  stands as the knowledge available about  $\Theta$  after  $\mathbf{X}$  is observed (a.k.a. **posterior distribution**)
- $p(\mathbf{X}|\Theta)$  measures how much the observed data are coherent to the model, assuming a certain value  $\Theta$  of the parameter (a.k.a. **likelihood**)
- $p(\mathbf{X}) = \sum_{\Theta'} p(\mathbf{X}|\Theta')p(\Theta')$  is the probability that  $\mathbf{X}$  is observed, considered as a mean w.r.t. all possible values of  $\Theta$  (a.k.a. **evidence**)

## Conjugate distributions

### Definition

Given a likelihood function  $p(y|x)$ , a (prior) distribution  $p(x)$  is **conjugate** to  $p(y|x)$  if the posterior distribution  $p(x|y)$  is of the same type as  $p(x)$ .

### Consequence

If we look at  $p(x)$  as our knowledge of the random variable  $x$  before knowing  $y$  and with  $p(x|y)$  our knowledge once  $y$  is known, the new knowledge can be expressed as the old one.

### Examples of conjugate distributions: beta-bernoulli

The Beta distribution is conjugate to the Bernoulli distribution. In fact, given  $x \in [0, 1]$  and  $y \in \{0, 1\}$ , if

$$\begin{aligned} p(\phi|\alpha, \beta) &= \text{Beta}(\phi|\alpha, \beta) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} \phi^{\alpha-1} (1-\phi)^{\beta-1} \\ p(x|\phi) &= \phi^x (1-\phi)^{1-x} \end{aligned}$$

then

$$p(\phi|x) = \frac{1}{Z} \phi^{\alpha-1} (1-\phi)^{\beta-1} \phi^x (1-\phi)^{1-x} = \text{Beta}(x|\alpha + x - 1, \beta - x)$$

where  $Z$  is the normalization coefficient

$$Z = \int_0^1 \phi^{\alpha+x-1} (1-\phi)^{\beta-x} d\phi = \frac{\Gamma(\alpha + \beta + 1)}{\Gamma(\alpha + x)\Gamma(\beta - x + 1)}$$

## Examples of conjugate distributions: beta-binomial

The Beta distribution is also conjugate to the Binomial distribution. In fact, given  $x \in [0, 1]$  and  $y \in \{0, 1\}$ , if

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

$$p(k|\phi, N) = \binom{N}{k} \phi^k (1-\phi)^{N-k} = \frac{N!}{(N-k)!k!} \phi^N (1-\phi)^{N-k}$$

then

$$p(\phi|k, N, \alpha, \beta) = \frac{1}{Z} \phi^{\alpha-1} (1-\phi)^{\beta-1} \phi^k (1-\phi)^{N-k} = \text{Beta}(\phi|\alpha+k-1, \beta+N-k-1)$$

with the normalization coefficient

$$Z = \int_0^1 \phi^{\alpha+k-1} (1-\phi)^{\beta+N-k-1} d\phi = \frac{\Gamma(\alpha+\beta+N)}{\Gamma(\alpha+k)\Gamma(\beta+N-k)}$$

## Multivariate distributions

### Multinomial

Generalization of the binomial

$$p(n_1, \dots, n_K | \phi_1, \dots, \phi_K, n) = \frac{n!}{\prod_{i=1}^K n_i!} \prod_{i=1}^K \phi_i^{n_i} \quad \sum_{i=1}^k n_i = n, \sum_{i=1}^k \phi_i = 1$$

the case  $n = 1$  is a generalization of the Bernoulli distribution

$$p(x_1, \dots, x_K | \phi_1, \dots, \phi_K) = \prod_{i=1}^K \phi_i^{x_i} \quad \forall i : x_i \in \{0, 1\}, \sum_{i=1}^K x_i = 1, \sum_{i=1}^K \phi_i = 1$$

### Likelihood of a multinomial

$$p(\mathbf{X} | \phi_1, \dots, \phi_K) \propto \prod_{i=1}^N \prod_{j=1}^K \phi_j^{x_{ij}} = \prod_{j=1}^K \phi_j^{N_j}$$

### Conjugate of the multinomial

#### Dirichlet distribution

The conjugate of the multinomial is the Dirichlet distribution, generalization of the Beta to the case  $K > 2$

$$p(\phi_1, \dots, \phi_K | \alpha_1, \dots, \alpha_K) = \text{Dir}(\phi | \boldsymbol{\alpha}) = \frac{\Gamma(\sum_{i=1}^K \alpha_i)}{\prod_{i=1}^K \Gamma(\alpha_i)} \prod_{i=1}^K \phi_i^{\alpha_i-1}$$

$$= \frac{1}{Z'} \prod_{i=1}^K \phi_i^{\alpha_i-1}$$

with  $\alpha_i > 0$  for  $i = 1, \dots, K$

## Random variables and Dirichlet distribution

A random variable  $\boldsymbol{\phi} = (\phi_1, \dots, \phi_K)$  with Dirichlet distribution takes values on the  $K - 1$  dimensional simplex (set of points  $\mathbf{x} \in \mathbb{R}^K$  such that  $x_i \geq 0$  for  $i = 1, \dots, K$  and  $\sum_{i=1}^K x_i = 1$ )

### Examples of conjugate distributions: dirichlet-multinomial

Assume  $\boldsymbol{\phi} \sim \text{Dir}(\boldsymbol{\phi}|\boldsymbol{\alpha})$  and  $z \sim \text{Mult}(z|\boldsymbol{\phi})$ . Then,

$$\begin{aligned} p(\boldsymbol{\phi}|z, \boldsymbol{\alpha}) &= \frac{p(z|\boldsymbol{\phi})p(\boldsymbol{\phi}|\boldsymbol{\alpha})}{p(z|\boldsymbol{\alpha})} = \frac{1}{Z} \frac{1}{Z'} \frac{1}{Z''} \prod_{i=1}^K \phi_i^{z_i} \prod_{i=1}^K \phi_i^{\alpha_i-1} \\ &= \frac{1}{Z'''} \prod_{i=1}^K \phi_i^{\alpha_i+z_i-1} = \text{Dir}(\boldsymbol{\phi}|\boldsymbol{\alpha}') \end{aligned}$$

where  $\boldsymbol{\alpha}' = (\alpha_1 + z_1, \dots, \alpha_K + z_K)$

## Text modeling

### Unigram model

Collection  $\mathbf{W}$  of  $N$  term occurrences:  $N$  observations of a same random variable, with multinomial distribution over a dictionary  $\mathbf{V}$  of size  $V$ .

$$p(\mathbf{W}|\boldsymbol{\phi}) = L(\boldsymbol{\phi}|\mathbf{W}) = \prod_{i=1}^V \phi_i^{N_i} \quad \sum_{i=1}^V \phi_i = 1, \sum_{i=1}^V N_i = N$$

### Parameter model

Use of a Dirichlet distribution, conjugate to the multinomial

$$\begin{aligned} p(\boldsymbol{\phi}|\boldsymbol{\alpha}) &= \text{Dir}(\boldsymbol{\phi}|\boldsymbol{\alpha}) \\ p(\boldsymbol{\phi}|\mathbf{W}, \boldsymbol{\alpha}) &= \text{Dir}(\boldsymbol{\phi}|\boldsymbol{\alpha} + \mathbf{N}) \end{aligned}$$

## Information theory

Let  $X$  be a discrete random variable:

- define a measure  $h(x)$  of the information (surprise) of observing  $X = x$
- requirements:
  - likely events provide low surprise, while rare events provide high surprise:  $h(x)$  is inversely proportional to  $p(x)$
  - $X, Y$  independent: the event  $X = x, Y = y$  has probability  $p(x)p(y)$ . Its surprise is the sum of the surprise for  $X = x$  and for  $Y = y$ , that is,  $h(x, y) = h(x) + h(y)$  (information is additive)

this results into  $h(x) = -\log x$  (usually base 2)

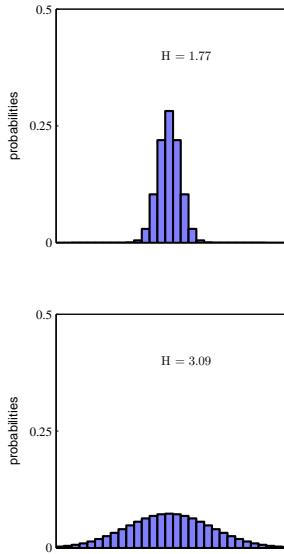
## Entropy

A sender transmits the value of  $X$  to a receiver: the expected amount of information transmitted (w.r.t.  $p(x)$ ) is the **entropy** of  $X$

$$H(X) = - \sum_x p(x) \log_2 p(x)$$

- lower entropy results from more sharply peaked distributions
- the uniform distribution provides the highest entropy

Entropy is a measure of disorder.



## Entropy, some properties

- $p(x) \in [0, 1]$  implies  $p(x) \log_2 p(x) \leq 0$  and  $H(X) \geq 0$
- $H(X) = 0$  if there exists  $x$  such that  $p(x) = 1$

## Maximum entropy

Given a fixed number  $k$  of outcomes, the distribution  $p_1, \dots, p_k$  with maximum entropy is derived by maximizing  $H(X)$  under the constraint  $\sum_{i=1}^k p_i = 1$ . By using Lagrange multipliers, this amounts to maximizing

$$-\sum_{i=1}^k p_i \log_2 p_i + \lambda \left( \sum_{i=1}^k p_i - 1 \right)$$

Setting the derivative of each  $p_i$  to 0,

$$0 = -\log_2 p_i - \log_2 e + \lambda$$

results into  $p_i = 2^\lambda - e$  for each  $i$ , that is into the uniform distribution  $p_i = \frac{1}{k}$  and  $H(X) = \log_2 k$

## Entropy, some properties

$H(X)$  is a lower bound on the expected number of bits needed to encode the values of  $X$

- trivial approach: code of length  $\log_2 k$  (assuming uniform distribution of values for  $X$ )
- for non-uniform distributions, better coding schemes by associating shorter codes to likely values of  $X$

## Conditional entropy

Let  $X, Y$  be discrete r.v. : for a pair of values  $x, y$  the additional information needed to specify  $y$  if  $x$  is known is  $-\ln p(y|x)$ .

The expected additional information needed to specify the value of  $Y$  if we assume the value of  $X$  is known is the **conditional entropy** of  $Y$  given  $X$

$$H(Y|X) = - \sum_x \sum_y p(x, y) \ln p(y|x)$$

Clearly, since  $\ln p(y|x) = \ln p(x, y) - \ln p(x)$

$$H(X, Y) = H(Y|X) + H(X)$$

that is, the information needed to describe (on the average) the values of  $X$  and  $Y$  is the sum of the information needed to describe the value of  $X$  plus that needed to describe the value of  $Y$  if  $X$  is known.

## KL divergence

Assume the distribution  $p(x)$  of  $X$  is unknown, and we have modeled it as an approximation  $q(x)$ .

If we use  $q(x)$  to encode values of  $X$  we need an average length  $-\sum_x p(x) \ln q(x)$ , while the minimum (known  $p(x)$ ) is  $-\sum_x p(x) \ln p(x)$ .

The additional amount of information needed, due to the approximation of  $p(x)$  through  $q(x)$  is the **Kullback-Leibler divergence**

$$\begin{aligned} KL(p||q) &= - \sum_x p(x) \ln q(x) + \sum_x p(x) \ln p(x) \\ &= - \sum_x p(x) \ln \frac{q(x)}{p(x)} \end{aligned}$$

$KL(p||q)$  measures the difference between the distributions  $p$  and  $q$ .

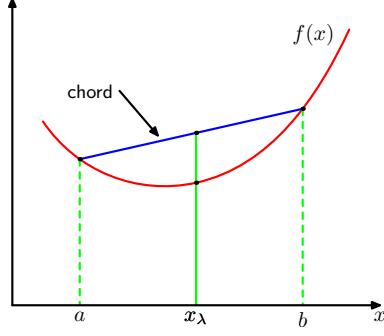
- $KL(p||p) = 0$
- $KL(p||q) \neq KL(q||p)$ : the function is not symmetric, it is not a distance (it would be  $d(x, y) = d(y, x)$ )

## Convexity

A function is convex (in an interval  $[a, b]$ ) if, for all  $0 \leq \lambda \leq 1$ , the following inequality holds

$$f(\lambda a + (1 - \lambda)b) \leq \lambda f(a) + (1 - \lambda)f(b)$$

- $\lambda a + (1 - \lambda)b$  is a point  $x \in [a, b]$  and  $f(\lambda a + (1 - \lambda)b)$  is the corresponding value of the function
- $\lambda f(a) + (1 - \lambda)f(b) = f(x)$  is the value at  $\lambda a + (1 - \lambda)b$  of the chord from  $(a, f(a))$  to  $(b, f(b))$ .



### Jensen's inequality and KL divergence

- If  $f(x)$  is a convex function, the **Jensen's inequality** holds for any set of points  $x_1, \dots, x_M$

$$f\left(\sum_{i=1}^M \lambda_i x_i\right) \leq \sum_{i=1}^M \lambda_i f(x_i)$$

where  $\lambda_i \geq 0$  for all  $i$  and  $\sum_{i=1}^M \lambda_i = 1$ .

- In particular, if  $\lambda_i = p(x_i)$ ,

$$f(E[x]) \leq E[f(x)]$$

- if  $x$  is a continuous variable, this results into

$$f\left(\int x p(x) dx\right) \leq \int f(x) p(x) dx$$

- applying the inequality to  $KL(p||q)$ , since the logarithm is convex,

$$KL(p||q) = - \int p(x) \ln \frac{q(x)}{p(x)} dx \geq - \ln \int q(x) dx = 0$$

thus proving the  $KL$  is always non-negative.

### Applying KL divergence

- $\mathbf{x} = (x_1, \dots, x_n)$ , dataset generated by a unknown distribution  $p(x)$
- we want to infer the parameters of a probabilistic model  $q_\theta(x|\theta)$
- approach: minimize

$$\begin{aligned} KL(p||q_\theta) &= - \sum_x p(x) \ln \frac{q(x|\theta)}{p(x)} \\ &\approx -\frac{1}{n} \sum_{i=1}^n \ln \frac{q(x_i|\theta)}{p(x_i)} \\ &= \frac{1}{n} \sum_{i=1}^n (\ln p(x_i) - \ln q(x_i|\theta)) \end{aligned}$$

First term is independent of  $\theta$ , while the second one is the negative log-likelihood of  $\mathbf{x}$ . The value of  $\theta$  which minimizes  $KL(p||q_\theta)$  also maximizes the log-likelihood.

## Mutual information

- Measure of the independence between  $X$  and  $Y$

$$I(X, Y) = KL(p(X, Y) || p(X)p(Y)) = - \sum_x \sum_y p(x, y) \ln \frac{p(x)p(y)}{p(x, y)}$$

additional encoding length if independence is assumed

- We have:

$$\begin{aligned} I(X, Y) &= - \sum_x \sum_y p(x, y) \ln \frac{p(x)p(y)}{p(x, y)} \\ &= - \sum_x \sum_y p(x, y) \ln \frac{p(x)p(y)}{p(x|y)p(y)} \\ &= - \sum_x \sum_y p(x, y) \ln \frac{p(x)}{p(x|y)} \\ &= - \sum_x \sum_y p(x, y) \ln p(x) + \sum_x \sum_y p(x, y) \ln p(x|y) = H(X) - H(X|Y) \end{aligned}$$

- Similarly, it derives  $I(X, Y) = H(Y) - H(Y|X)$