

# Chapter 2

## Probability

### 2.1 Frequentists vs. Bayesians

what is probability?

One is called the **frequentist** interpretation. In this view, probabilities represent long run frequencies of events. For example, the above statement means that, if we flip the coin many times, we expect it to land heads about half the time.

The other interpretation is called the **Bayesian** interpretation of probability. In this view, probability is used to quantify our **uncertainty** about something; hence it is fundamentally related to information rather than repeated trials (Jaynes 2003). In the Bayesian view, the above statement means we believe the coin is equally likely to land heads or tails on the next toss

One big advantage of the Bayesian interpretation is that it can be used to model our uncertainty about events that do not have long term frequencies. For example, we might want to compute the probability that the polar ice cap will melt by 2020 CE. This event will happen zero or one times, but cannot happen repeatedly. Nevertheless, we ought to be able to quantify our uncertainty about this event. To give another machine learning oriented example, we might have observed a blip on our radar screen, and want to compute the probability distribution over the location of the corresponding target (be it a bird, plane, or missile). In all these cases, the idea of repeated trials does not make sense, but the Bayesian interpretation is valid and indeed quite natural. We shall therefore adopt the Bayesian interpretation in this book. Fortunately, the basic rules of probability theory are the same, no matter which interpretation is adopted.

## 2.2 A brief review of probability theory

### 2.2.1 Basic concepts

We denote a random event by defining a **random variable**  $X$ .

**Discrete random variable:**  $X$  can take on any value from a finite or countably infinite set.

**Continuous random variable:** the value of  $X$  is real-valued.

#### 2.2.1.1 CDF

$$F(x) \triangleq P(X \leq x) = \begin{cases} \sum_{u \leq x} P(u) & , \text{ discrete} \\ \int_{-\infty}^x f(u) du & , \text{ continuous} \end{cases} \quad (2.1)$$

#### 2.2.1.2 PMF and PDF

For discrete random variable, We denote the probability of the event that  $X = x$  by  $P(X = x)$ , or just  $p(x)$  for short. Here  $p(x)$  is called a **probability mass function** or **PMF**. A probability mass function is a function that gives the probability that a discrete random variable is exactly equal to some value<sup>4</sup>. This satisfies the properties  $0 \leq p(x) \leq 1$  and  $\sum_{x \in \mathcal{X}} p(x) = 1$ .

For continuous variable, in the equation  $F(x) = \int_{-\infty}^x f(u) du$ , the function  $f(x)$  is called a **probability density function** or **PDF**. A probability density function is a function that describes the relative likelihood for this random variable to take on a given value<sup>5</sup>. This satisfies the properties  $f(x) \geq 0$  and  $\int_{-\infty}^{\infty} f(x) dx = 1$ .

### 2.2.2 Multivariate random variables

#### 2.2.2.1 Joint CDF

We denote joint CDF by  $F(x, y) \triangleq P(X \leq x \cap Y \leq y) = P(X \leq x, Y \leq y)$ .

$$F(x, y) \triangleq P(X \leq x, Y \leq y) = \begin{cases} \sum_{u \leq x, v \leq y} P(u, v) \\ \int_{-\infty}^x \int_{-\infty}^y f(u, v) du dv \end{cases} \quad (2.2)$$

**product rule:**

$$p(X, Y) = P(X|Y)P(Y) \quad (2.3)$$

**Chain rule:**

<sup>4</sup> [http://en.wikipedia.org/wiki/Probability\\_mass\\_function](http://en.wikipedia.org/wiki/Probability_mass_function)

<sup>5</sup> [http://en.wikipedia.org/wiki/Probability\\_density\\_function](http://en.wikipedia.org/wiki/Probability_density_function)

$$p(X_{1:N}) = p(X_1)p(X_2|X_1)\dots p(X_N|X_{1:N-1}) \quad (2.4)$$

## 2.2.4 Independence and conditional independence

### 2.2.2.2 Marginal distribution

**Marginal CDF:**

$$F_X(x) \triangleq F(x, +\infty) = \begin{cases} \sum_{x_i \leq x} P(X = x_i) = \sum_{x_i \leq x} \sum_{j=1}^{+\infty} P(X = x_i, Y = y_j) \\ \int_{-\infty}^x f_X(u) du = \int_{-\infty}^x \int_{-\infty}^{+\infty} f(u, v) du dv \end{cases} \quad (2.5)$$

$$F_Y(y) \triangleq F(+\infty, y) = \begin{cases} \sum_{y_j \leq y} p(Y = y_j) = \sum_{i=1}^{+\infty} \sum_{y_j \leq y} P(X = x_i, Y = y_j) \\ \int_{-\infty}^y f_Y(v) dv = \int_{-\infty}^{+\infty} \int_{-\infty}^y f(u, v) du dv \end{cases} \quad (2.6)$$

**Marginal PMF and PDF:**

$$\begin{cases} P(X = x_i) = \sum_{j=1}^{+\infty} P(X = x_i, Y = y_j) & , \text{ discrete} \\ f_X(x) = \int_{-\infty}^{+\infty} f(x, y) dy & , \text{ continuous} \end{cases} \quad (2.7)$$

$$\begin{cases} p(Y = y_j) = \sum_{i=1}^{+\infty} P(X = x_i, Y = y_j) & , \text{ discrete} \\ f_Y(y) = \int_{-\infty}^{+\infty} f(x, y) dx & , \text{ continuous} \end{cases} \quad (2.8)$$

### 2.2.2.3 Conditional distribution

**Conditional PMF:**

$$p(X = x_i | Y = y_j) = \frac{p(X = x_i, Y = y_j)}{p(Y = y_j)} \text{ if } p(Y) > 0 \quad (2.9)$$

The pmf  $p(X|Y)$  is called **conditional probability**.

**Conditional PDF:**

$$f_{X|Y}(x|y) = \frac{f(x, y)}{f_Y(y)} \quad (2.10)$$

### 2.2.3 Bayes rule

$$\begin{aligned} p(Y = y | X = x) &= \frac{p(X = x, Y = y)}{p(X = x)} \\ &= \frac{p(X = x | Y = y)p(Y = y)}{\sum_{y'} p(X = x | Y = y')p(Y = y')} \end{aligned} \quad (2.11)$$

We say  $X$  and  $Y$  are unconditionally independent or marginally independent, denoted  $X \perp Y$ , if we can represent the joint as the product of the two marginals, i.e.,

$$X \perp Y = P(X, Y) = P(X)P(Y) \quad (2.12)$$

We say  $X$  and  $Y$  are conditionally independent(CI) given  $Z$  if the conditional joint can be written as a product of conditional marginals:

$$X \perp Y | Z = P(X, Y | Z) = P(X | Z)P(Y | Z) \quad (2.13)$$

### 2.2.5 Quantiles

Since the cdf  $F$  is a monotonically increasing function, it has an inverse; let us denote this by  $F^{-1}$ . If  $F$  is the cdf of  $X$ , then  $F^{-1}(\alpha)$  is the value of  $x_\alpha$  such that  $P(X \leq x_\alpha) = \alpha$ ; this is called the  $\alpha$  quantile of  $F$ . The value  $F^{-1}(0.5)$  is the **median** of the distribution, with half of the probability mass on the left, and half on the right. The values  $F^{-1}(0.25)$  and  $F^{-1}(0.75)$  are the lower and upper **quartiles**.

### 2.2.6 Mean and variance

The most familiar property of a distribution is its **mean**, or **expected value**, denoted by  $\mu$ . For discrete rvs, it is defined as  $\mathbb{E}[X] \triangleq \sum_{x \in \mathcal{X}} xp(x)$ , and for continuous rvs, it is defined as  $\mathbb{E}[X] \triangleq \int_{\mathcal{X}} xp(x)dx$ . If this integral is not finite, the mean is not defined (we will see some examples of this later).

The **variance** is a measure of the spread of a distribution, denoted by  $\sigma^2$ . This is defined as follows:

$$\begin{aligned} \text{var}[X] &= \mathbb{E}[(X - \mu)^2] \\ &= \int (x - \mu)^2 p(x) dx \\ &= \int x^2 p(x) dx + \mu^2 \int p(x) dx - 2\mu \int xp(x) dx \\ &= \mathbb{E}[X^2] - \mu^2 \end{aligned} \quad (2.14)$$

from which we derive the useful result

$$\mathbb{E}[X^2] = \sigma^2 + \mu^2 \quad (2.16)$$

The **standard deviation** is defined as

$$\text{std}[X] \triangleq \sqrt{\text{var}[X]} \quad (2.17)$$

This is useful since it has the same units as  $X$  itself.

## 2.3 Some common discrete distributions

In this section, we review some commonly used parametric distributions defined on discrete state spaces, both finite and countably infinite.

### 2.3.1 The Bernoulli and binomial distributions

**Definition 2.1.** Now suppose we toss a coin only once. Let  $X \in \{0, 1\}$  be a binary random variable, with probability of success or heads of  $\theta$ . We say that  $X$  has a **Bernoulli distribution**. This is written as  $X \sim \text{Ber}(\theta)$ , where the pmf is defined as

$$\text{Ber}(x|\theta) \triangleq \theta^{\mathbb{I}(x=1)}(1-\theta)^{\mathbb{I}(x=0)} \quad (2.18)$$

**Definition 2.2.** Suppose we toss a coin  $n$  times. Let  $X \in \{0, 1, \dots, n\}$  be the number of heads. If the probability of heads is  $\theta$ , then we say  $X$  has a **binomial distribution**, written as  $X \sim \text{Bin}(n, \theta)$ . The pmf is given by

$$\text{Bin}(k|n, \theta) \triangleq \binom{n}{k} \theta^k (1-\theta)^{n-k} \quad (2.19)$$

### 2.3.2 The multinoulli and multinomial distributions

**Definition 2.3.** The Bernoulli distribution can be used to model the outcome of one coin tosses. To model the outcome of tossing a  $K$ -sided dice, let  $\mathbf{x} = (\mathbb{I}(x=1), \dots, \mathbb{I}(x=K)) \in \{0, 1\}^K$  be a random vector (this is called **dummy encoding** or **one-hot encoding**), then we say  $X$  has a **multinoulli distribution** (or **categorical distribution**), written as  $X \sim \text{Cat}(\theta)$ . The pmf is given by:

$$p(\mathbf{x}) \triangleq \prod_{k=1}^K \theta_k^{\mathbb{I}(x_k=1)} \quad (2.20)$$

**Definition 2.4.** Suppose we toss a  $K$ -sided dice  $n$  times. Let  $\mathbf{x} = (x_1, x_2, \dots, x_K) \in \{0, 1, \dots, n\}^K$  be a random vector, where  $x_j$  is the number of times side  $j$  of the dice occurs, then we say  $X$  has a **multinomial distribution**,

written as  $X \sim \text{Mu}(n, \theta)$ . The pmf is given by

$$p(\mathbf{x}) \triangleq \binom{n}{x_1 \dots x_K} \prod_{k=1}^K \theta_k^{x_k} \quad (2.21)$$

$$\text{where } \binom{n}{x_1 \dots x_K} \triangleq \frac{n!}{x_1! x_2! \dots x_K!}$$

Bernoulli distribution is just a special case of a Binomial distribution with  $n = 1$ , and so is multinoulli distribution as to multinomial distribution. See Table 2.1 for a summary.

Table 2.1: Summary of the multinomial and related distributions.

Name	$K$	$n$	$X$
Bernoulli	1	1	$x \in \{0, 1\}$
Binomial	1	-	$\mathbf{x} \in \{0, 1, \dots, n\}$
Multinoulli	-	1	$\mathbf{x} \in \{0, 1\}^K, \sum_{k=1}^K x_k = 1$
Multinomial	-	-	$\mathbf{x} \in \{0, 1, \dots, n\}^K, \sum_{k=1}^K x_k = n$

### 2.3.3 The Poisson distribution

**Definition 2.5.** We say that  $X \in \{0, 1, 2, \dots\}$  has a **Poisson distribution** with parameter  $\lambda > 0$ , written as  $X \sim \text{Poi}(\lambda)$ , if its pmf is

$$p(x|\lambda) = e^{-\lambda} \frac{\lambda^x}{x!} \quad (2.22)$$

The first term is just the normalization constant, required to ensure the distribution sums to 1.

The Poisson distribution is often used as a model for counts of rare events like radioactive decay and traffic accidents.

### 2.3.4 The empirical distribution

The **empirical distribution function**<sup>6</sup>, or **empirical cdf**, is the cumulative distribution function associated with the empirical measure of the sample. Let  $\mathcal{D} = \{x_1, x_2, \dots, x_N\}$  be a sample set, it is defined as

$$F_n(x) \triangleq \frac{1}{N} \sum_{i=1}^N \mathbb{I}(x_i \leq x) \quad (2.23)$$

<sup>6</sup> [http://en.wikipedia.org/wiki/Empirical\\_distribution\\_function](http://en.wikipedia.org/wiki/Empirical_distribution_function)

Table 2.2: Summary of Bernoulli, binomial multinoulli and multinomial distributions.

Name	Written as	X	$p(x)$ (or $p(\mathbf{x})$ )	$\mathbb{E}[X]$	$\text{var}[X]$
Bernoulli	$X \sim \text{Ber}(\theta)$	$x \in \{0, 1\}$	$\theta^{\mathbb{I}(x=1)}(1-\theta)^{\mathbb{I}(x=0)}$	$\theta$	$\theta(1-\theta)$
Binomial	$X \sim \text{Bin}(n, \theta)$	$x \in \{0, 1, \dots, n\}$	$\binom{n}{k} \theta^k (1-\theta)^{n-k}$	$n\theta$	$n\theta(1-\theta)$
Multinoulli	$X \sim \text{Cat}(\boldsymbol{\theta})$	$\mathbf{x} \in \{0, 1\}^K, \sum_{k=1}^K x_k = 1$	$\prod_{k=1}^K \theta_j^{\mathbb{I}(x_j=1)}$	-	-
Multinomial	$X \sim \text{Mu}(n, \boldsymbol{\theta})$	$\mathbf{x} \in \{0, 1, \dots, n\}^K, \sum_{k=1}^K x_k = n$	$\binom{n}{x_1 \dots x_K} \prod_{k=1}^K \theta_k^{x_k}$	-	-
Poisson	$X \sim \text{Poi}(\lambda)$	$x \in \{0, 1, 2, \dots\}$	$e^{-\lambda} \frac{\lambda^x}{x!}$	$\lambda$	$\lambda$

## 2.4 Some common continuous distributions

In this section we present some commonly used univariate (one-dimensional) continuous probability distributions.

### 2.4.1 Gaussian (normal) distribution

Table 2.3: Summary of Gaussian distribution.

Written as	$f(x)$	$\mathbb{E}[X]$	mode	$\text{var}[X]$
$X \sim \mathcal{N}(\mu, \sigma^2)$	$\frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2\sigma^2}(x-\mu)^2}$	$\mu$	$\mu$	$\sigma^2$

If  $X \sim \mathcal{N}(0, 1)$ , we say  $X$  follows a **standard normal** distribution.

The Gaussian distribution is the most widely used distribution in statistics. There are several reasons for this.

1. First, it has two parameters which are easy to interpret, and which capture some of the most basic properties of a distribution, namely its mean and variance.
2. Second, the central limit theorem (Section TODO) tells us that sums of independent random variables have an approximately Gaussian distribution, making it a good choice for modeling residual errors or noise.
3. Third, the Gaussian distribution makes the least number of assumptions (has maximum entropy), subject to the constraint of having a specified mean and variance, as we show in Section TODO; this makes it a good default choice in many cases.
4. Finally, it has a simple mathematical form, which results in easy to implement, but often highly effective, methods, as we will see.

See (Jaynes 2003, ch 7) for a more extensive discussion of why Gaussians are so widely used.

### 2.4.2 Student's t-distribution

Table 2.4: Summary of Student's t-distribution.

Written as	$f(x)$	$\mathbb{E}[X]$	mode	$\text{var}[X]$
$X \sim \mathcal{T}(\mu, \sigma^2, \nu)$	$\frac{\Gamma(\frac{\nu+1}{2})}{\sqrt{\nu\pi}\Gamma(\frac{\nu}{2})} \left[ 1 + \frac{1}{\nu} \left( \frac{x-\mu}{\sigma} \right)^2 \right]^{-\frac{\nu+1}{2}}$	$\mu$	$\mu$	$\frac{\nu\sigma^2}{\nu-2}$

where  $\Gamma(x)$  is the gamma function:

$$\Gamma(x) \triangleq \int_0^{\infty} t^{x-1} e^{-t} dt \quad (2.24)$$

$\mu$  is the mean,  $\sigma^2 > 0$  is the scale parameter, and  $\nu > 0$  is called the **degrees of freedom**. See Figure 2.1 for some plots.

The variance is only defined if  $\nu > 2$ . The mean is only defined if  $\nu > 1$ .

As an illustration of the robustness of the Student distribution, consider Figure 2.2. We see that the Gaussian is affected a lot, whereas the Student distribution hardly changes. This is because the Student has heavier tails, at least for small  $\nu$  (see Figure 2.1).

If  $\nu = 1$ , this distribution is known as the **Cauchy** or **Lorentz** distribution. This is notable for having such heavy tails that the integral that defines the mean does not converge.

To ensure finite variance, we require  $\nu > 2$ . It is common to use  $\nu = 4$ , which gives good performance in a range of problems (Lange et al. 1989). For  $\nu \gg 5$ , the Student distribution rapidly approaches a Gaussian distribution and loses its robustness properties.

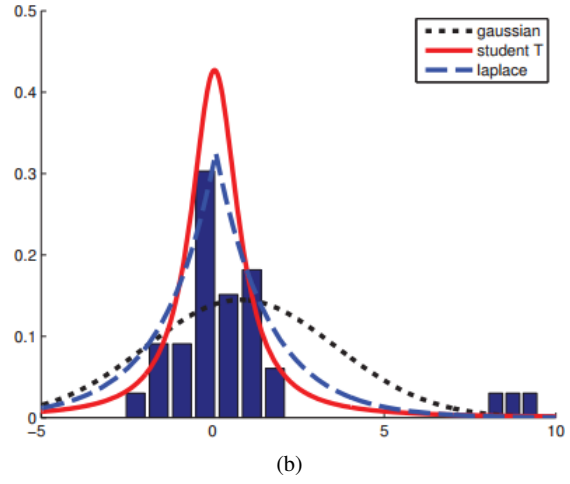
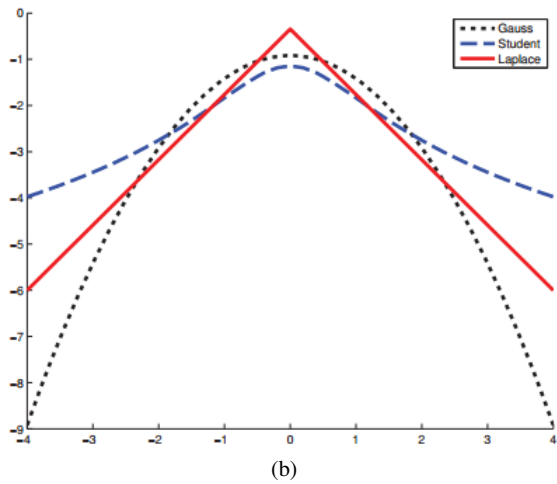
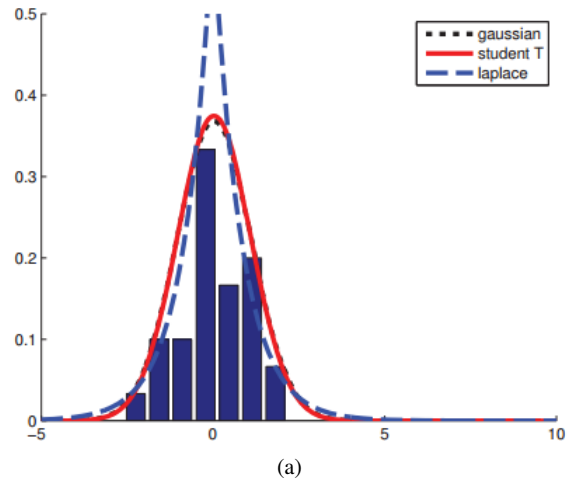
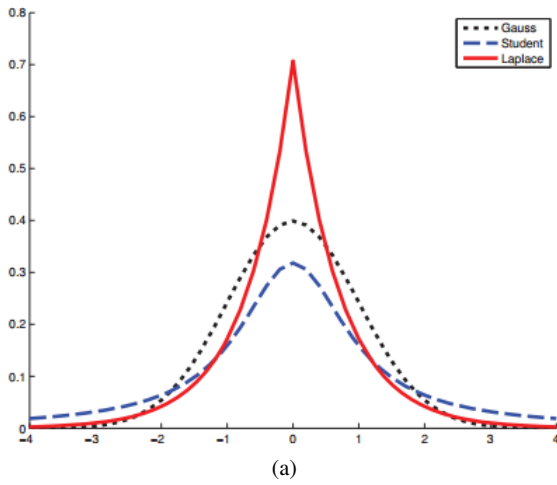


Fig. 2.1: (a) The pdfs for a  $\mathcal{N}(0, 1)$ ,  $\mathcal{T}(0, 1, 1)$  and  $Lap(0, 1/\sqrt{2})$ . The mean is 0 and the variance is 1 for both the Gaussian and Laplace. The mean and variance of the Student is undefined when  $\nu = 1$ . (b) Log of these pdfs. Note that the Student distribution is not log-concave for any parameter value, unlike the Laplace distribution, which is always log-concave (and log-convex...) Nevertheless, both are unimodal.

Fig. 2.2: Illustration of the effect of outliers on fitting Gaussian, Student and Laplace distributions. (a) No outliers (the Gaussian and Student curves are on top of each other). (b) With outliers. We see that the Gaussian is more affected by outliers than the Student and Laplace distributions.

Table 2.5: Summary of Laplace distribution.

Written as	$f(x)$	$\mathbb{E}[X]$	mode	$\text{var}[X]$
$X \sim \text{Lap}(\mu, b)$	$\frac{1}{2b} \exp\left(-\frac{ x-\mu }{b}\right)$	$\mu$	$\mu$	$2b^2$

### 2.4.3 The Laplace distribution

Here  $\mu$  is a location parameter and  $b > 0$  is a scale parameter. See Figure 2.1 for a plot.

Its robustness to outliers is illustrated in Figure 2.2. It also puts more probability density at 0 than the Gaussian. This property is a useful way to encourage sparsity in a model, as we will see in Section TODO.

Table 2.6: Summary of gamma distribution

Written as	$X$	$f(x)$	$\mathbb{E}[X]$	mode	$\text{var}[X]$
$X \sim \text{Ga}(a, b)$	$x \in \mathbb{R}^+$	$\frac{b^a}{\Gamma(a)} x^{a-1} e^{-xb}$	$\frac{a}{b}$	$\frac{a-1}{b}$	$\frac{a}{b^2}$

## 2.4.4 The gamma distribution

Here  $a > 0$  is called the shape parameter and  $b > 0$  is called the rate parameter. See Figure 2.3 for some plots.

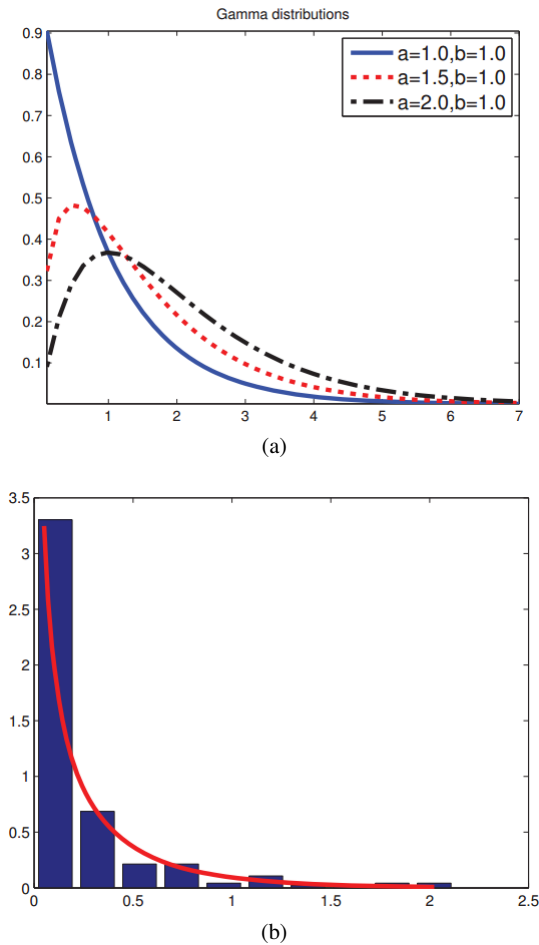


Fig. 2.3: Some  $\text{Ga}(a, b = 1)$  distributions. If  $a \leq 1$ , the mode is at 0, otherwise it is  $> 0$ . As we increase the rate  $b$ , we reduce the horizontal scale, thus squeezing everything leftwards and upwards. (b) An empirical pdf of some rainfall data, with a fitted Gamma distribution superimposed.

## 2.4.5 The beta distribution

Here  $B(a, b)$  is the beta function,

$$B(a, b) \triangleq \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)} \quad (2.25)$$

See Figure 2.4 for plots of some beta distributions. We require  $a, b > 0$  to ensure the distribution is integrable (i.e., to ensure  $B(a, b)$  exists). If  $a = b = 1$ , we get the uniform distribution. If  $a$  and  $b$  are both less than 1, we get a bimodal distribution with spikes at 0 and 1; if  $a$  and  $b$  are both greater than 1, the distribution is unimodal.

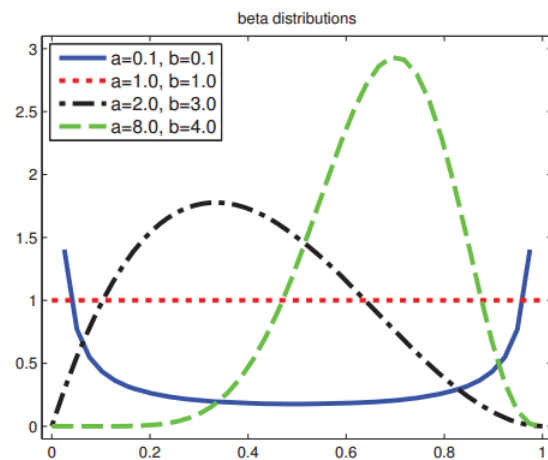


Fig. 2.4: Some beta distributions.

## 2.4.6 Pareto distribution

The **Pareto distribution** is used to model the distribution of quantities that exhibit **long tails**, also called **heavy tails**.

As  $k \rightarrow \infty$ , the distribution approaches  $\delta(x - m)$ . See Figure 2.5(a) for some plots. If we plot the distribution on a log-log scale, it forms a straight line, of the form  $\log p(x) = a \log x + c$  for some constants  $a$  and  $c$ . See Figure 2.5(b) for an illustration (this is known as a **power law**).

Table 2.7: Summary of Beta distribution

Name	Written as	$X$	$f(x)$	$\mathbb{E}[X]$	mode	$\text{var}[X]$
Beta distribution	$X \sim \text{Beta}(a, b)$	$x \in [0, 1]$	$\frac{1}{B(a, b)} x^{a-1} (1-x)^{b-1}$	$\frac{a}{a+b}$	$\frac{a-1}{a+b-2}$	$\frac{ab}{(a+b)^2(a+b+1)}$

Table 2.8: Summary of Pareto distribution

Name	Written as	$X$	$f(x)$	$\mathbb{E}[X]$	mode	$\text{var}[X]$
Pareto distribution	$X \sim \text{Pareto}(k, m)$	$x \geq m$	$km^k x^{-(k+1)} \mathbb{I}(x \geq m)$	$\frac{km}{k-1}$ if $k > 1$	$m$	$\frac{m^2 k}{(k-1)^2(k-2)}$ if $k > 2$

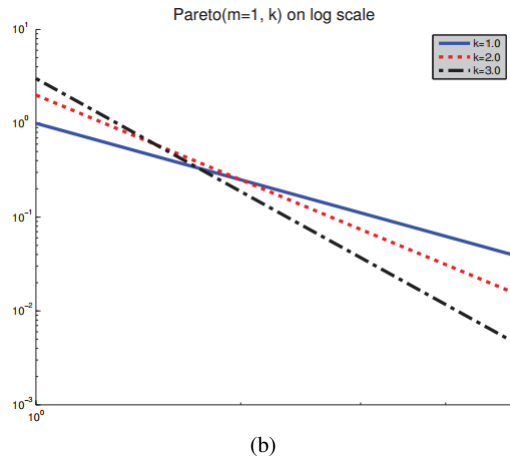
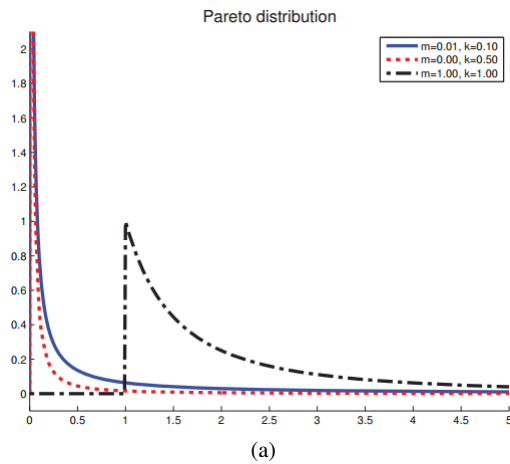


Fig. 2.5: (a) The Pareto distribution  $\text{Pareto}(x|m, k)$  for  $m = 1$ . (b) The pdf on a log-log scale.

## 2.5 Joint probability distributions

Given a **multivariate random variable** or **random vector**  $X \in \mathbb{R}^D$ , the **joint probability distribution**<sup>8</sup> is a probability distribution that gives the probability that each of  $X_1, X_2, \dots, X_D$  falls in any particular range or discrete set of values specified for that variable. In the case of only two random variables, this is called a **bivariate distribution**, but the concept generalizes to any number of random variables, giving a **multivariate distribution**.

The joint probability distribution can be expressed either in terms of a **joint cumulative distribution function** or in terms of a **joint probability density function** (in the case of continuous variables) or **joint probability mass function** (in the case of discrete variables).

### 2.5.1 Covariance and correlation

**Definition 2.6.** The **covariance** between two rvs  $X$  and  $Y$  measures the degree to which  $X$  and  $Y$  are (linearly) related. Covariance is defined as

$$\begin{aligned} \text{cov}[X, Y] &\triangleq \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])] \\ &= \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y] \end{aligned} \quad (2.26)$$

**Definition 2.7.** If  $X$  is a  $D$ -dimensional random vector, its **covariance matrix** is defined to be the following symmetric, positive definite matrix:

<sup>7</sup> [http://en.wikipedia.org/wiki/Multivariate\\_random\\_variable](http://en.wikipedia.org/wiki/Multivariate_random_variable)

<sup>8</sup> [http://en.wikipedia.org/wiki/Joint\\_probability\\_distribution](http://en.wikipedia.org/wiki/Joint_probability_distribution)

$$\begin{aligned} \text{cov}[X] &\triangleq \mathbb{E}[(X - \mathbb{E}[X])(X - \mathbb{E}[X])^T] \\ &= \begin{pmatrix} \text{var}[X_1] & \text{Cov}[X_1, X_2] & \cdots & \text{Cov}[X_1, X_D] \\ \text{Cov}[X_2, X_1] & \text{var}[X_2] & \cdots & \text{Cov}[X_2, X_D] \\ \vdots & \vdots & \ddots & \vdots \\ \text{Cov}[X_D, X_1] & \text{Cov}[X_D, X_2] & \cdots & \text{var}[X_D] \end{pmatrix} \end{aligned} \quad (2.27)$$

$$(2.28)$$

**Definition 2.8.** The (Pearson) **correlation coefficient** between  $X$  and  $Y$  is defined as

$$\text{corr}[X, Y] \triangleq \frac{\text{Cov}[X, Y]}{\sqrt{\text{var}[X], \text{var}[Y]}} \quad (2.29)$$

A **correlation matrix** has the form

$$\mathbf{R} \triangleq \begin{pmatrix} \text{corr}[X_1, X_1] & \text{corr}[X_1, X_2] & \cdots & \text{corr}[X_1, X_D] \\ \text{corr}[X_2, X_1] & \text{corr}[X_2, X_2] & \cdots & \text{corr}[X_2, X_D] \\ \vdots & \vdots & \ddots & \vdots \\ \text{corr}[X_D, X_1] & \text{corr}[X_D, X_2] & \cdots & \text{corr}[X_D, X_D] \end{pmatrix} \quad (2.30)$$

The correlation coefficient can be viewed as a degree of linearity between  $X$  and  $Y$ , see Figure 2.6.

**Uncorrelated does not imply independent.** For example, let  $X \sim U(-1, 1)$  and  $Y = X^2$ . Clearly  $Y$  is dependent on  $X$  (in fact,  $Y$  is uniquely determined by  $X$ ), yet one can show that  $\text{corr}[X, Y] = 0$ . Some striking examples of this fact are shown in Figure 2.6. This shows several data sets where there is clear dependence between  $X$  and  $Y$ , and yet the correlation coefficient is 0. A more general measure of dependence between random variables is mutual information, see Section TODO.

## 2.5.2 Multivariate Gaussian distribution

The **multivariate Gaussian** or **multivariate normal** (MVN) is the most widely used joint probability density function for continuous variables. We discuss MVNs in detail in Chapter 4; here we just give some definitions and plots.

The pdf of the MVN in  $D$  dimensions is defined by the following:

$$\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) \triangleq \frac{1}{(2\pi)^{D/2} |\boldsymbol{\Sigma}|^{1/2}} \exp\left[-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu})\right] \quad (2.31)$$

where  $\boldsymbol{\mu} = \mathbb{E}[X] \in \mathbb{R}^D$  is the mean vector, and  $\boldsymbol{\Sigma} = \text{Cov}[X]$  is the  $D \times D$  covariance matrix. The normalization constant  $(2\pi)^{D/2} |\boldsymbol{\Sigma}|^{1/2}$  just ensures that the pdf integrates to 1.

Figure 2.7 plots some MVN densities in 2d for three different kinds of covariance matrices. A full covariance matrix has  $D(D+1)/2$  parameters (we divide by 2 since

$\boldsymbol{\Sigma}$  is symmetric). A diagonal covariance matrix has  $D$  parameters, and has 0s in the off-diagonal terms. A spherical or isotropic covariance,  $\boldsymbol{\Sigma} = \sigma^2 \mathbf{I}_D$ , has one free parameter.

## 2.5.3 Multivariate Student's t-distribution

A more robust alternative to the MVN is the multivariate Student's t-distribution, whose pdf is given by

$$\begin{aligned} \mathcal{T}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}, \nu) &\triangleq \frac{\Gamma(\frac{\nu+D}{2})}{\Gamma(\frac{\nu}{2})} \frac{|\boldsymbol{\Sigma}|^{-\frac{1}{2}}}{(\nu\pi)^{\frac{D}{2}}} \left[1 + \frac{1}{\nu}(\mathbf{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu})\right]^{-\frac{\nu+D}{2}} \\ &= \frac{\Gamma(\frac{\nu+D}{2})}{\Gamma(\frac{\nu}{2})} \frac{|\boldsymbol{\Sigma}|^{-\frac{1}{2}}}{(\nu\pi)^{\frac{D}{2}}} \left[1 + (\mathbf{x} - \boldsymbol{\mu})^T \mathbf{V}^{-1}(\mathbf{x} - \boldsymbol{\mu})\right]^{-\frac{\nu+D}{2}} \end{aligned} \quad (2.32)$$

$$(2.33)$$

where  $\boldsymbol{\Sigma}$  is called the scale matrix (since it is not exactly the covariance matrix) and  $\mathbf{V} = \nu\boldsymbol{\Sigma}$ . This has fatter tails than a Gaussian. The smaller  $\nu$  is, the fatter the tails. As  $\nu \rightarrow \infty$ , the distribution tends towards a Gaussian. The distribution has the following properties

$$\text{mean} = \boldsymbol{\mu}, \text{ mode} = \boldsymbol{\mu}, \text{ Cov} = \frac{\nu}{\nu-2} \boldsymbol{\Sigma} \quad (2.34)$$

## 2.5.4 Dirichlet distribution

A multivariate generalization of the beta distribution is the **Dirichlet distribution**, which has support over the probability simplex, defined by

$$S_K = \left\{ \mathbf{x} : 0 \leq x_k \leq 1, \sum_{k=1}^K x_k = 1 \right\} \quad (2.35)$$

The pdf is defined as follows:

$$\text{Dir}(\mathbf{x}|\boldsymbol{\alpha}) \triangleq \frac{1}{B(\boldsymbol{\alpha})} \prod_{k=1}^K x_k^{\alpha_k - 1} \mathbb{I}(\mathbf{x} \in S_K) \quad (2.36)$$

where  $B(\alpha_1, \alpha_2, \dots, \alpha_K)$  is the natural generalization of the beta function to  $K$  variables:

$$B(\boldsymbol{\alpha}) \triangleq \frac{\prod_{k=1}^K \Gamma(\alpha_k)}{\Gamma(\alpha_0)} \quad \text{where } \alpha_0 \triangleq \sum_{k=1}^K \alpha_k \quad (2.37)$$

Figure 2.8 shows some plots of the Dirichlet when  $K = 3$ , and Figure 2.9 for some sampled probability vectors. We see that  $\alpha_0$  controls the strength of the dis-



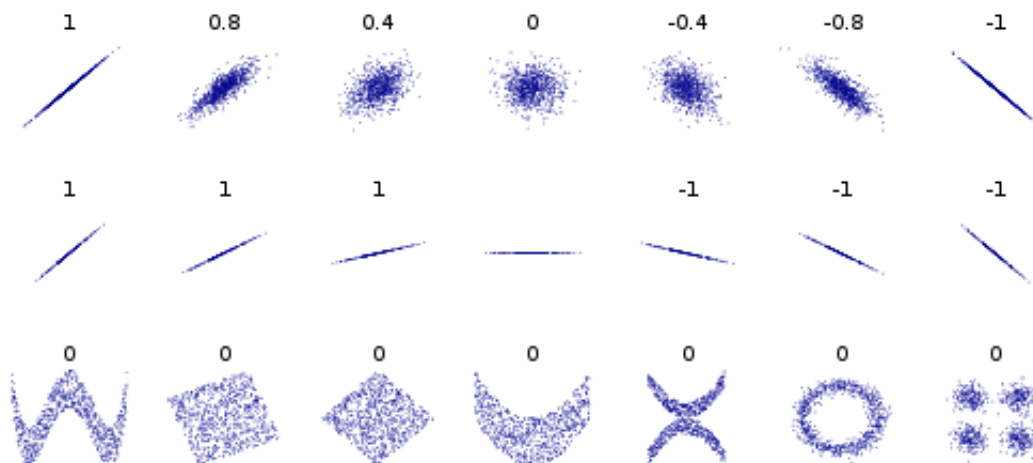


Fig. 2.6: Several sets of  $(x,y)$  points, with the Pearson correlation coefficient of  $x$  and  $y$  for each set. Note that the correlation reflects the noisiness and direction of a linear relationship (top row), but not the slope of that relationship (middle), nor many aspects of nonlinear relationships (bottom). N.B.: the figure in the center has a slope of 0 but in that case the correlation coefficient is undefined because the variance of  $Y$  is zero. Source: <http://en.wikipedia.org/wiki/Correlation>

tribution (how peaked it is), and the control where the peak occurs. For example,  $\text{Dir}(1, 1, 1)$  is a uniform distribution,  $\text{Dir}(2, 2, 2)$  is a broad distribution centered at  $(1/3, 1/3, 1/3)$ , and  $\text{Dir}(20, 20, 20)$  is a narrow distribution centered at  $(1/3, 1/3, 1/3)$ . If  $\alpha_k < 1$  for all  $k$ , we get spikes at the corner of the simplex.

For future reference, the distribution has these properties

$$\mathbb{E}(x_k) = \frac{\alpha_k}{\alpha_0}, \text{mode}[x_k] = \frac{\alpha_k - 1}{\alpha_0 - K}, \text{var}[x_k] = \frac{\alpha_k(\alpha_0 - \alpha_k)}{\alpha_0^2(\alpha_0 + 1)} \quad (2.38)$$

## 2.6 Transformations of random variables

If  $\mathbf{x} \sim P()$  is some random variable, and  $\mathbf{y} = f(\mathbf{x})$ , what is the distribution of  $Y$ ? This is the question we address in this section.

### 2.6.1 Linear transformations

Suppose  $g()$  is a linear function:

$$g(\mathbf{x}) = A\mathbf{x} + b \quad (2.39)$$

First, for the mean, we have

$$\mathbb{E}[\mathbf{y}] = \mathbb{E}[A\mathbf{x} + b] = A\mathbb{E}[\mathbf{x}] + b \quad (2.40)$$

this is called the **linearity of expectation**.

For the covariance, we have

$$\text{Cov}[\mathbf{y}] = \text{Cov}[A\mathbf{x} + b] = A\mathbf{\Sigma}A^T \quad (2.41)$$

### 2.6.2 General transformations

If  $X$  is a discrete rv, we can derive the pmf for  $y$  by simply summing up the probability mass for all the  $x$ s such that  $f(x) = y$ :

$$p_Y(y) = \sum_{x:g(x)=y} p_X(x) \quad (2.42)$$

If  $X$  is continuous, we cannot use Equation 2.42 since  $p_X(x)$  is a density, not a pmf, and we cannot sum up densities. Instead, we work with cdfs, and write

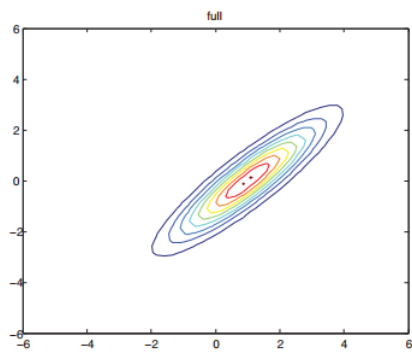
$$F_Y(y) = P(Y \leq y) = P(g(X) \leq y) = \int_{g(X) \leq y} f_X(x) dx \quad (2.43)$$

We can derive the pdf of  $Y$  by differentiating the cdf:

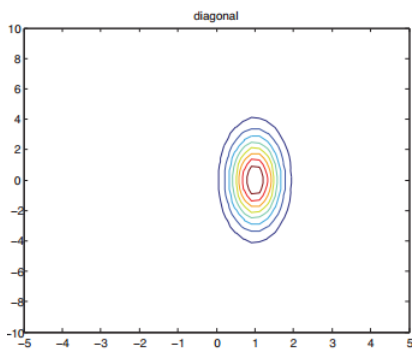
$$f_Y(y) = f_X(x) \left| \frac{dx}{dy} \right| \quad (2.44)$$

This is called **change of variables** formula. We leave the proof of this as an exercise.

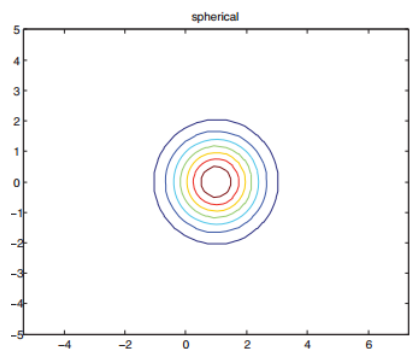
For example, suppose  $X \sim U(1, 1)$ , and  $Y = X^2$ . Then  $p_Y(y) = \frac{1}{2}y^{-\frac{1}{2}}$ .



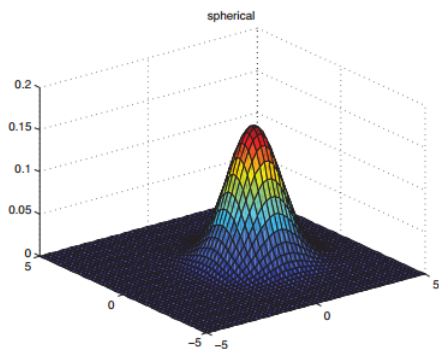
(a)



(b)

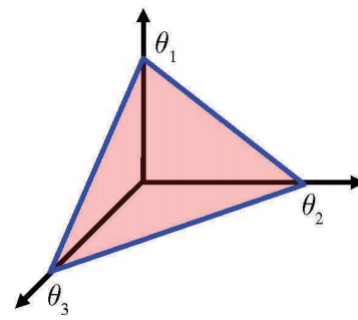


(c)

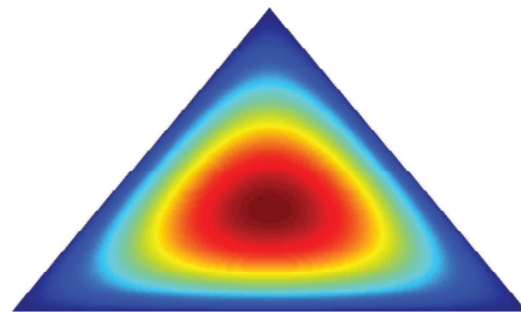


(d)

Fig. 2.7: We show the level sets for 2d Gaussians. (a) A full covariance matrix has elliptical contours. (b) A diagonal covariance matrix is an axis aligned ellipse. (c) A spherical covariance matrix has a circular shape. (d) Surface plot for the spherical Gaussian in (c).



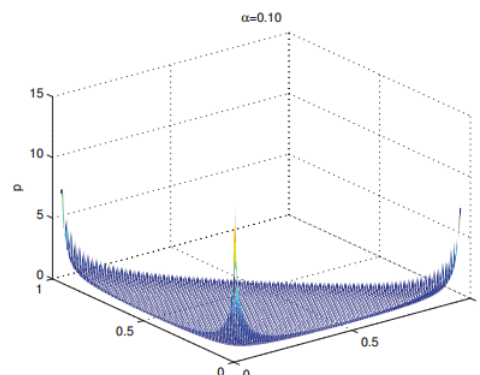
(a)



(b)



(c)



(d)

Fig. 2.8: (a) The Dirichlet distribution when  $K = 3$  defines a distribution over the simplex, which can be represented by the triangular surface. Points on this surface satisfy  $0 \leq \theta_k \leq 1$  and  $\sum_{k=1}^K \theta_k = 1$ . (b) Plot of the Dirichlet density when  $\alpha = (2, 2, 2)$ . (c)  $\alpha = (20, 2, 2)$ .

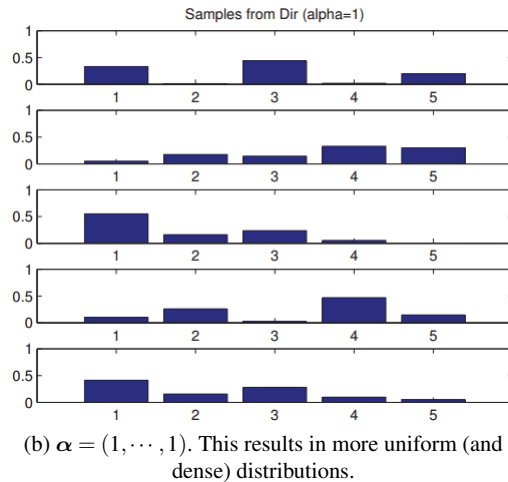
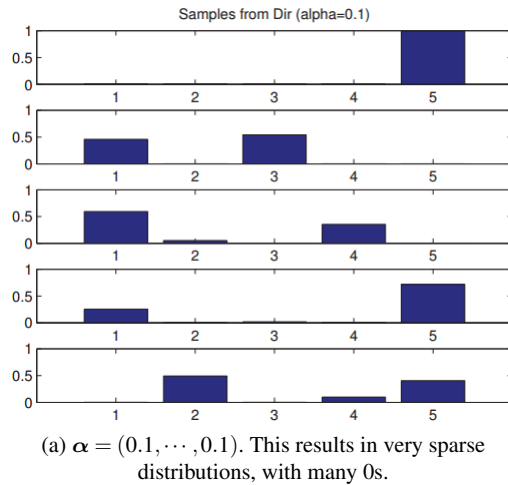


Fig. 2.9: Samples from a 5-dimensional symmetric Dirichlet distribution for different parameter values.

### 2.6.2.1 Multivariate change of variables \*

Let  $f$  be a function  $f: \mathbb{R}^n \rightarrow \mathbb{R}^n$ , and let  $\mathbf{y} = f(\mathbf{x})$ . Then its Jacobian matrix  $\mathbf{J}$  is given by

$$\mathbf{J}_{\mathbf{x} \rightarrow \mathbf{y}} \triangleq \frac{\partial \mathbf{y}}{\partial \mathbf{x}} \triangleq \begin{pmatrix} \frac{\partial y_1}{\partial x_1} & \dots & \frac{\partial y_1}{\partial x_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial y_n}{\partial x_1} & \dots & \frac{\partial y_n}{\partial x_n} \end{pmatrix} \quad (2.45)$$

$|\det(\mathbf{J})|$  measures how much a unit cube changes in volume when we apply  $f$ .

If  $f$  is an invertible mapping, we can define the pdf of the transformed variables using the Jacobian of the inverse mapping  $\mathbf{y} \rightarrow \mathbf{x}$ :

$$p_{\mathbf{y}}(\mathbf{y}) = p_{\mathbf{x}}(\mathbf{x}) \left| \det \left( \frac{\partial \mathbf{x}}{\partial \mathbf{y}} \right) \right| = p_{\mathbf{x}}(\mathbf{x}) \left| \det(\mathbf{J}_{\mathbf{y} \rightarrow \mathbf{x}}) \right| \quad (2.46)$$

## 2.6.3 Central limit theorem

Given  $N$  random variables  $X_1, X_2, \dots, X_N$ , each variable is **independent and identically distributed**<sup>9</sup> (**iid** for short), and each has the same mean  $\mu$  and variance  $\sigma^2$ , then

$$\frac{\sum_{i=1}^n X_i - N\mu}{\sqrt{N}\sigma} \sim \mathcal{N}(0, 1) \quad (2.47)$$

this can also be written as

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{N}} \sim \mathcal{N}(0, 1) \quad , \text{ where } \bar{X} \triangleq \frac{1}{N} \sum_{i=1}^n X_i \quad (2.48)$$

## 2.7 Monte Carlo approximation

In general, computing the distribution of a function of a rv using the change of variables formula can be difficult. One simple but powerful alternative is as follows. First we generate  $S$  samples from the distribution, call them  $x_1, \dots, x_S$ . (There are many ways to generate such samples; one popular method, for high dimensional distributions, is called Markov chain Monte Carlo or MCMC; this will be explained in Chapter TODO.) Given the samples, we can approximate the distribution of  $f(X)$  by using the empirical distribution of  $\{f(x_s)\}_{s=1}^S$ . This is called a **Monte Carlo approximation**<sup>10</sup>, named after a city in Europe known for its plush gambling casinos.

We can use Monte Carlo to approximate the expected value of any function of a random variable. We simply draw samples, and then compute the arithmetic mean of the function applied to the samples. This can be written as follows:

$$\mathbb{E}[g(X)] = \int g(x)p(x)dx \approx \frac{1}{S} \sum_{s=1}^S g(x_s) \quad (2.49)$$

where  $x_s \sim p(X)$ .

This is called **Monte Carlo integration**<sup>11</sup>, and has the advantage over numerical integration (which is based on evaluating the function at a fixed grid of points) that the function is only evaluated in places where there is non-negligible probability.

<sup>9</sup> [http://en.wikipedia.org/wiki/Independent\\_identically\\_distributed](http://en.wikipedia.org/wiki/Independent_identically_distributed)

<sup>10</sup> [http://en.wikipedia.org/wiki/Monte\\_Carlo\\_method](http://en.wikipedia.org/wiki/Monte_Carlo_method)

<sup>11</sup> [http://en.wikipedia.org/wiki/Monte\\_Carlo\\_integration](http://en.wikipedia.org/wiki/Monte_Carlo_integration)

## 2.8 Information theory

### 2.8.1 Entropy

The entropy of a random variable  $X$  with distribution  $p$ , denoted by  $\mathbb{H}(X)$  or sometimes  $\mathbb{H}(p)$ , is a measure of its uncertainty. In particular, for a discrete variable with  $K$  states, it is defined by

$$\mathbb{H}(X) \triangleq - \sum_{k=1}^K p(X=k) \log_2 p(X=k) \quad (2.50)$$

Usually we use log base 2, in which case the units are called **bits** (short for binary digits). If we use log base  $e$ , the units are called **nats**.

The discrete distribution with maximum entropy is the uniform distribution (see Section XXX for a proof). Hence for a  $K$ -ary random variable, the entropy is maximized if  $p(x=k) = 1/K$ ; in this case,  $\mathbb{H}(X) = \log_2 K$ .

Conversely, the distribution with minimum entropy (which is zero) is any **delta-function** that puts all its mass on one state. Such a distribution has no uncertainty.

### 2.8.2 KL divergence

One way to measure the dissimilarity of two probability distributions,  $p$  and  $q$ , is known as the **Kullback-Leibler divergence (KL divergence)** or **relative entropy**. This is defined as follows:

$$\mathbb{KL}(P||Q) \triangleq \sum_x p(x) \log_2 \frac{p(x)}{q(x)} \quad (2.51)$$

where the sum gets replaced by an integral for pdfs<sup>12</sup>. The KL divergence is only defined if  $P$  and  $Q$  both sum to 1 and if  $q(x) = 0$  implies  $p(x) = 0$  for all  $x$  (absolute continuity). If the quantity  $0 \ln 0$  appears in the formula, it is interpreted as zero because  $\lim_{x \rightarrow 0} x \ln x$ . We can rewrite this as

$$\begin{aligned} \mathbb{KL}(p||q) &\triangleq \sum_x p(x) \log_2 p(x) - \sum_{k=1}^K p(x) \log_2 q(x) \\ &= \mathbb{H}(p, q) - \mathbb{H}(p) \end{aligned} \quad (2.52)$$

where  $\mathbb{H}(p, q)$  is called the **cross entropy**,

<sup>12</sup> The KL divergence is not a distance, since it is asymmetric. One symmetric version of the KL divergence is the **Jensen-Shannon divergence**, defined as  $JS(p_1, p_2) = 0.5\mathbb{KL}(p_1||q) + 0.5\mathbb{KL}(p_2||q)$ , where  $q = 0.5p_1 + 0.5p_2$

$$\mathbb{H}(p, q) = - \sum_x p(x) \log_2 q(x) \quad (2.53)$$

One can show (Cover and Thomas 2006) that the cross entropy is the average number of bits needed to encode data coming from a source with distribution  $p$  when we use model  $q$  to define our codebook. Hence the regular entropy  $\mathbb{H}(p) = \mathbb{H}(p, p)$ , defined in section §2.8.1, is the expected number of bits if we use the true model, so the KL divergence is the difference between these. In other words, the KL divergence is the average number of *extra* bits needed to encode the data, due to the fact that we used distribution  $q$  to encode the data instead of the true distribution  $p$ .

The extra number of bits interpretation should make it clear that  $\mathbb{KL}(p||q) \geq 0$ , and that the KL is only equal to zero if  $q = p$ . We now give a proof of this important result.

**Theorem 2.1. (Information inequality)**  $\mathbb{KL}(p||q) \geq 0$  with equality iff  $p = q$ .

One important consequence of this result is that *the discrete distribution with the maximum entropy is the uniform distribution*.

### 2.8.3 Mutual information

**Definition 2.9. Mutual information** or **MI**, is defined as follows:

$$\begin{aligned} \mathbb{I}(X;Y) &\triangleq \mathbb{KL}(P(X,Y)||P(X)P(Y)) \\ &= \sum_x \sum_y p(x,y) \log \frac{p(x,y)}{p(x)p(y)} \end{aligned} \quad (2.54)$$

We have  $\mathbb{I}(X;Y) \geq 0$  with equality if  $P(X,Y) = P(X)P(Y)$ . That is, the MI is zero if the variables are independent.

To gain insight into the meaning of MI, it helps to re-express it in terms of joint and conditional entropies. One can show that the above expression is equivalent to the following:

$$\mathbb{I}(X;Y) = \mathbb{H}(X) - \mathbb{H}(X|Y) \quad (2.55)$$

$$= \mathbb{H}(Y) - \mathbb{H}(Y|X) \quad (2.56)$$

$$= \mathbb{H}(X) + \mathbb{H}(Y) - \mathbb{H}(X,Y) \quad (2.57)$$

$$= \mathbb{H}(X,Y) - \mathbb{H}(X|Y) - \mathbb{H}(Y|X) \quad (2.58)$$

where  $\mathbb{H}(X)$  and  $\mathbb{H}(Y)$  are the **marginal entropies**,  $\mathbb{H}(X|Y)$  and  $\mathbb{H}(Y|X)$  are the **conditional entropies**, and  $\mathbb{H}(X,Y)$  is the **joint entropy** of  $X$  and  $Y$ , see Fig. 2.10<sup>13</sup>.

<sup>13</sup> [http://en.wikipedia.org/wiki/Mutual\\_information](http://en.wikipedia.org/wiki/Mutual_information)

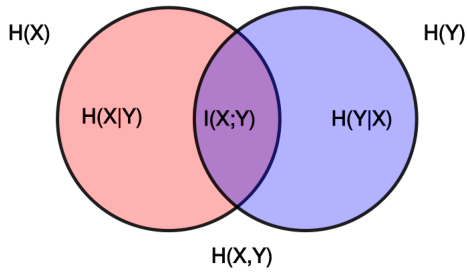


Fig. 2.10: Individual  $\mathbb{H}(X)$ ,  $\mathbb{H}(Y)$ , joint  $\mathbb{H}(X, Y)$ , and conditional entropies for a pair of correlated subsystems  $X, Y$  with mutual information  $\mathbb{I}(X; Y)$ .

Intuitively, we can interpret the MI between  $X$  and  $Y$  as the reduction in uncertainty about  $X$  after observing  $Y$ , or, by symmetry, the reduction in uncertainty about  $Y$  after observing  $X$ .

A quantity which is closely related to MI is the **point-wise mutual information** or **PMI**. For two events (not random variables)  $x$  and  $y$ , this is defined as

$$PMI(x, y) \triangleq \log \frac{p(x, y)}{p(x)p(y)} = \log \frac{p(x|y)}{p(x)} = \log \frac{p(y|x)}{p(y)} \quad (2.59)$$

This measures the discrepancy between these events occurring together compared to what would be expected by chance. Clearly the MI of  $X$  and  $Y$  is just the expected value of the PMI. Interestingly, we can rewrite the PMI as follows:

$$PMI(x, y) = \log \frac{p(x|y)}{p(x)} = \log \frac{p(y|x)}{p(y)} \quad (2.60)$$

This is the amount we learn from updating the prior  $p(x)$  into the posterior  $p(x|y)$ , or equivalently, updating the prior  $p(y)$  into the posterior  $p(y|x)$ .