# Mathematical Methods for Computer Science

R.J. Gibbens

Computer Laboratory
University of Cambridge

Lent Term 2007/8
(Last revised on 19 Feb 2008)

# Outline

- Part I: Fourier and related methods
  - Inner product spaces, Fourier series and transforms (3 lectures)
  - Discrete Fourier Transform and Fast Fourier Transform (2 lectures)
  - Wavelets (1 lecture)
- Part II: Probability methods
  - Review of elementary probability theory (1 lecture)
  - Probability generating functions (1 lecture)
  - Elementary stochastic processes (2 lectures)
  - Limits and inequalities (3 lectures)
  - Markov chains (3 lectures)

# Reference books

- (*) Pinkus, A. & Zafrany, S.
  *Fourier series and integral transforms*.
  Cambridge University Press, 1997
- Oppenheim, A.V. & Willsky, A.S.
  *Signals and systems*.
  Prentice-Hall, 1997
- (*) Ross, Sheldon M.
  *Probability Models for Computer Science*.
  Harcourt/Academic Press, 2002
- Mitzenmacher, Michael & Upfal, Eli.
  *Probability and Computing: Randomized Algorithms and Probabilistic Analysis*.
  Cambridge University Press, 2005

# Fourier and related methods

# Introduction

In this section we shall consider what it means to represent a function $f(x)$ in terms of other, perhaps simpler, functions. One example is Fourier series of the form

$$\frac{a_0}{2} + \sum_{n=1}^{\infty} \left[ a_n \cos(nx) + b_n \sin(nx) \right] .$$

How are the coefficients $a_n$ and $b_n$ related to the choice of $f(x)$ and what other representations can we use?
We shall take a quite general approach to these questions and derive the necessary framework that underpins a wide range of applications.

# Linear space

## Definition (Linear space)

A non-empty set $V$ of vectors is a linear space over a field $F$ of scalars if the following are satisfied.

1. Binary operation $+$ such that if $u, v \in V$ then $u + v \in V$
2. $+$ is associative: for all $u, v, w \in V$ then $(u + v) + w = u + (v + w)$
3. There exists a zero vector, written $\vec{0}$, such that $\vec{0} + v = v$ for all $v \in V$.
4. For all $v \in V$, there exists an inverse vector, written $-v$, such that $v + (-v) = \vec{0}$
5. $+$ is commutative: for all $u, v \in V$ then $u + v = v + u$
6. For all $v \in V$ and $a \in F$ then $av \in V$ is defined
7. For all $a \in F$ and $u, v \in V$ then $a(u + v) = au + av$
8. For all $a, b \in F$ and $v \in V$ then $(a + b)v = av + bv$ and $a(bu) = (ab)u$
9. For all $v \in V$ then $1v = v$, where $1 \in F$ is the unit scalar.

# Linear subspaces

Two common choices of scalar fields are the real numbers, $\mathbb{R}$, and the complex numbers, $\mathbb{C}$, giving rise to real and complex linear spaces, respectively.

## Definition (Linear subspace)

A subset $W \subset V$ is a linear subspace of $V$ if the W is again a linear space over the same field of scalars.

Thus $W$ is a linear subspace if $W \neq \emptyset$ and for all $u, v \in W$
and $a, b \in F$ we have that $au + bv \in W$.

# Linear combinations and spans

## Definition (Linear combinations)

If $V$ is a linear space and $v_1, v_2, \ldots, v_n \in V$ are vectors in $V$ then $u \in V$ is a linear combination of $v_1, v_2, \ldots, v_n$ if there exist scalars $a_1, a_2, \ldots, a_n$ such that

$$u = a_1 v_1 + a_2 v_2 + \cdots + a_n v_n\,.$$

We also define

$$\mathrm{span}\{v_1, v_2, \ldots, v_n\} = \{u \in V \,:\, u \text{ is a linear combination of } v_1, v_2, \ldots, v_n\}\,.$$

Thus, $W = \mathrm{span}\{v_1, v_2, \ldots, v_n\}$ is a linear subspace of $V$.

# Linear independence

## Definition (Linear independence)

Let $V$ be a linear space. The vectors $v_1, v_2, \ldots, v_n \in V$ are linearly independent if whenever

$$a_1 v_1 + a_2 v_2 + \cdots + a_n v_n = \vec{0} \qquad a_1, a_2, \ldots a_n \in F$$

then $a_1 = a_2 = \cdots = a_n = 0$

The vectors $v_1, v_2, \ldots, v_n$ are linearly dependent otherwise.

# Bases

## Definition (Basis)

A finite set of vectors $v_1, v_2, \ldots v_n \in V$ is a basis for the linear space $V$ if $v_1, v_2, \ldots, v_n$ are linearly independent and $V = \text{span}\{v_1, v_2, \ldots, v_n\}$. The number $n$ is called the dimension of $V$, written $n = \dim(V)$.

A result from linear algebra is that while there are infinitely many choices of basis vectors any two bases will always consist of the same number of element vectors. Thus, the dimension of a linear space is well-defined.

# Inner product

Suppose that $V$ is either a real or complex linear space (that is, $F = \mathbb{R}$ or $\mathbb{C}$).

## Definition (Inner product)

The inner product of two vectors $u, v \in V$, written $\langle u, v \rangle \in F$, is a scalar value satisfying

1. For each $v \in V$, $\langle v, v \rangle$ is a non-negative real number, so $\langle v, v \rangle \geq 0$
2. For each $v \in V$, $\langle v, v \rangle = 0$ if and only if $v = \vec{0}$
3. For all $u, v, w \in V$ and $a, b \in F$, $\langle au + bv, w \rangle = a\langle u, w \rangle + b\langle v, w \rangle$
4. For all $u, v \in V$ then $\langle u, v \rangle = \overline{\langle v, u \rangle}$.

A linear space together with an inner product is called an inner product space. Here, $\overline{\langle v, u \rangle}$ denotes the complex conjugate of the complex number $\langle v, u \rangle$. When the field of scalars, $F$, is the real numbers, $\mathbb{R}$, then $\overline{\langle v, u \rangle} = \langle v, u \rangle$.

# Useful properties of the inner product

Before looking at some examples of inner products there are several consequences of the definition of an inner product that are useful in calculations.

1. For all $v \in V$ and $a \in \mathbb{C}$ then $\langle av, av \rangle = |a|^2 \langle v, v \rangle$
2. For all $v \in V$, $\langle \vec{0}, v \rangle = 0$
3. For all $v \in V$ and finite sequences of vectors $u_1, u_2, \ldots, u_n \in V$ and scalars $a_1, a_2, \ldots, a_n$ then

$$\left\langle \sum_{i=1}^{n} a_i u_i, v \right\rangle = \sum_{i=1}^{n} a_i \langle u_i, v \rangle$$

$$\left\langle v, \sum_{i=1}^{n} a_i u_i \right\rangle = \sum_{i=1}^{n} \overline{a_i} \langle v, u_i \rangle$$

# Inner product: examples

### Example (Euclidean space, $\mathbb{R}^n$)

$V = \mathbb{R}^n$ with the usual operations of vector addition and multiplication by a scalar is a linear space over $\mathbb{R}$. Given two vectors $x = (x_1, x_2, \ldots, x_n)$ and $y = (y_1, y_2, \ldots, y_n)$ in $\mathbb{R}^n$ we can define an inner product

$$\langle x, y \rangle = \sum_{i=1}^{n} x_i y_i \, .$$

Often, this inner product is known as the dot product and is written $x \cdot y$.

### Example

Similarly, for $V = \mathbb{C}^n$, we can define an inner product by

$$\langle x, y \rangle = x \cdot y = \sum_{i=1}^{n} x_i \overline{y_i} \, .$$

## Example (Space of continuous functions)

$V = C[a,b]$, the space of continuous functions $f : [a,b] \to \mathbb{C}$ with the standard operations of the sum of two functions and multiplication by a scalar, is a linear space over $\mathbb{C}$ and we can define an inner product for $f, g \in C[a,b]$ by

$$\langle f, g \rangle = \int_a^b f(x)\overline{g(x)}dx\,.$$

# Norms

The concept of a norm is closely related to an inner product and we shall see that there is a natural way to define a norm given an inner product.

## Definition (Norm)

Let $V$ be a real or complex linear space. A norm on $V$ is a function from $V$ to $\mathbb{R}_+$, written $||v||$, that satisfies

1. For all $v \in V$, $||v|| \geq 0$
2. $||v|| = 0$ if and only if $v = \vec{0}$
3. For each $v \in V$ and $a \in \mathbb{C}$, $||av|| = |a|\,||v||$
4. For all $u, v \in V$, $||u + v|| \leq ||u|| + ||v||$ (the triangle inequality).

A norm can be thought of as a generalisation of the notion of distance, where for any two vectors $u, v \in V$ the number $||u - v||$ is the distance between $u$ and $v$.

# Norms: examples

## Example (Eucidean norm)

If $V = \mathbb{R}^n$ or $\mathbb{C}^n$ then for $x = (x_1, x_2, \ldots, x_n) \in V$ define

$$||x|| = \sqrt{\sum_{i=1}^{n} |x_i|^2} .$$

## Example (Uniform norm)

If $V = \mathbb{R}^n$ or $\mathbb{C}^n$ then for $x = (x_1, x_2, \ldots, x_n) \in V$ define

$$||x||_\infty = \max \{|x_i| \,:\, i = 1, 2, \ldots, n\} \ .$$

## Example (Uniform norm)

If $V = C[a, b]$ then for each function $f \in V$, define

$$||f||_\infty = \max \{|f(x)| \,:\, x \in [a, b]\} \ .$$

# Cauchy-Schwarz inequality

## Theorem (Cauchy-Schwarz inequality)

*Let $V$ be a real or complex inner product space then for all $u, v \in V$*

$$|\langle u, v\rangle|^2 \leq \langle u, u\rangle \langle v, v\rangle \, .$$

## Proof.

If $v = \vec{0}$ then the result holds trivially. Now assume $v \neq \vec{0}$ so that $\langle v, v\rangle \neq 0$ and let $\lambda \in \mathbb{C}$ then

$$0 \leq \langle u - \lambda v, u - \lambda v\rangle = \langle u, u\rangle - \overline{\lambda}\langle u, v\rangle - \lambda\langle v, u\rangle + |\lambda|^2\langle v, v\rangle$$

Now set $\lambda = \frac{\langle u,v\rangle}{\langle v,v\rangle}$ so that

$$0 \leq \langle u, u\rangle - \frac{|\langle u, v\rangle|^2}{\langle v, v\rangle}$$

and hence

$$|\langle u, v\rangle|^2 \leq \langle u, u\rangle\langle v, v\rangle \, .$$

□

# Inner products and norms

Given an inner product space, $V$, with inner product $\langle \cdot, \cdot \rangle$ there is a natural choice of norm, namely, for all $v \in V$

$$||v|| = \sqrt{\langle v, v \rangle}$$

that is, the non-negative square root.
Most of the properties that make this a norm follow simply from the properties of the inner product but we shall use the Cauchy-Schwarz inequality to establish the triangle inequality. We have,

$$\begin{aligned} ||u+v||^2 &= \langle u+v, u+v \rangle \\ &= ||u||^2 + \langle u, v \rangle + \langle v, u \rangle + ||v||^2 \\ &\leq ||u||^2 + 2|\langle u, v \rangle| + ||v||^2 \\ &\leq ||u||^2 + 2||u||\,||v|| + ||v||^2 \\ &= (||u|| + ||v||)^2 \,. \end{aligned}$$

Hence, the triangle inequality, $||u+v|| \leq ||u|| + ||v||$ holds.

# Orthogonal and orthonormal systems

Let $V$ be an inner product space.

### Definition (Orthogonality)

We say that $u, v \in V$ are orthogonal if $\langle u, v \rangle = 0$, written $u \perp v$.

### Definition (Orthogonal system)

A finite or infinite sequence of vectors $(u_i)$ in $V$ is an orthogonal system if

1. $u_i \neq \vec{0}$ for all such vectors $u_i$
2. $u_i \perp u_j$ for all $i \neq j$.

### Definition (Orthonormal system)

An orthogonal system is called an orthonormal system if, in addition, $||u_i|| = 1$ for all such vectors $u_i$.

A vector $v \in V$ such that $||v|| = 1$ is called a unit vector.

## Theorem

*Suppose that $\{e_1, e_2, \dots, e_n\}$ is an orthonormal system in the inner product space $V$. If $u = \sum_{i=1}^{n} a_i e_i$ then $a_i = \langle u, e_i \rangle$.*

## Proof.

$$\begin{aligned} \langle u, e_i \rangle &= \langle a_1 e_1 + a_2 e_2 + \cdots + a_n e_n, e_i \rangle \\ &= a_1 \langle e_1, e_i \rangle + a_2 \langle e_2, e_i \rangle + \cdots + a_n \langle e_n, e_i \rangle \\ &= a_i \,. \end{aligned}$$

□

Hence, if $\{e_1, e_2, \dots, e_n\}$ is an orthonormal system then for all $u \in \operatorname{span}\{e_1, e_2, \dots, e_n\}$ we have

$$u = \sum_{i=1}^{n} a_i e_i = \sum_{i=1}^{n} \langle u, e_i \rangle e_i \,.$$

# Fourier coefficients

Let $V$ be an inner product space and $e_1, e_2, \ldots, e_n$ an orthonormal system ($n$ being finite or infinite).

## Definition (Generalized Fourier coefficients)

Given a vector $u \in V$, the scalars $\langle u, e_i \rangle$ ($i = 1, 2, \ldots, n$) are called the Generalized Fourier coefficients of $u$ with respect to the given orthonormal system.

These coefficients are generalized in the sense that they refer to a general orthonormal system.

Let $V$ be an inner product space and $e_1, e_2, \dots, e_n$ an orthonormal system. If $a_1, a_2, \dots, a_n$ and $b_1, b_2, \dots, b_n$ are any sequences of scalars then

$$\left\langle \sum_{i=1}^{n} a_i e_i, \sum_{i=1}^{n} b_i e_i \right\rangle = \sum_{i=1}^{n} a_i \overline{b_i} \,.$$

Equivalently, for $u, v \in \operatorname{span}\{e_1, e_2, \dots, e_n\}$

$$\langle u, v \rangle = \sum_{i=1}^{n} \langle u, e_i \rangle \overline{\langle v, e_i \rangle} \,.$$

A consequence of these relations is the following theorem.

## Theorem (Generalized Pythagorean Theorem)

*Suppose that $\{u_1, u_2, \dots, u_n\}$ is an orthogonal system in $V$ and $a_1, a_2, \dots, a_n$ are scalars then*

$$|| \sum_{i=1}^{n} a_i u_i ||^2 = \sum_{i=1}^{n} |a_i|^2 \, ||u_i||^2 \,.$$

# Orthogonal projections

Suppose that $V$ is an inner product space and $e_1, e_2, \ldots, e_n$ an orthonormal system. Define $W = \operatorname{span}\{e_1, e_2, \ldots, e_n\}$ and let $u \in V$ be any vector. We have seen that for $u \in W$

$$u = \sum_{i=1}^{n} \langle u, e_i \rangle e_i$$

but if $u \notin W$ then certainly

$$u \neq \sum_{i=1}^{n} \langle u, e_i \rangle e_i$$

since $u$ is not a linear combination of the vectors $e_1, e_2, \ldots, e_n$. Nevertheless, there is a close connection between $u$ and the expression $\sum_{i=1}^{n} \langle u, e_i \rangle e_i$.

## Definition (Orthogonal projection)

For all $u \in V$ we define the orthogonal projection of $u$ in $W$, $\tilde{u}$, by

$$\tilde{u} = \sum_{i=1}^{n} \langle u, e_i \rangle e_i \, .$$

## Theorem

*For each $u \in V$ and for all $w \in W$*

1. $\langle u - \tilde{u}, w \rangle = 0$
2. $||u - w||^2 = ||u - \tilde{u}||^2 + ||\tilde{u} - w||^2$.

## Proof

First $\langle u - \tilde{u}, e_j \rangle = 0$ for all $j = 1, 2, \ldots, n$ since

$$\langle u - \tilde{u}, e_j \rangle = \langle u, e_j \rangle - \left\langle \sum_{i=1}^{n} \langle u, e_i \rangle e_i, e_j \right\rangle = \langle u, e_j \rangle - \sum_{i=1}^{n} \langle u, e_i \rangle \langle e_i, e_j \rangle$$
$$= \langle u, e_j \rangle - \langle u, e_j \rangle \langle e_j, e_j \rangle = \langle u, e_j \rangle - \langle u, e_j \rangle = 0\,.$$

So take any $w \in W$ with $w = \sum_{j=1}^{n} b_j e_j$ for some scalars $b_1, b_2, \ldots, b_n$ and

$$\langle u - \tilde{u}, w \rangle = \left\langle u - \tilde{u}, \sum_{j=1}^{n} b_j e_j \right\rangle = \sum_{j=1}^{n} \overline{b_j} \langle u - \tilde{u}, e_j \rangle = \sum_{j=1}^{n} \overline{b_j} \cdot 0 = 0 \,.$$

Now $(u - \tilde{u}) \perp w$ for all $w \in W$ and so since $\tilde{u} - w \in W$ $(u - \tilde{u}) \perp (\tilde{u} - w)$. Hence,

$$||u - w||^2 = ||u - \tilde{u} + \tilde{u} - w||^2 = ||u - \tilde{u}||^2 + ||\tilde{u} - w||^2 \,.$$

□

# Best approximation

## Theorem

*Let $V$ be an inner product space and $\{e_1, e_2, \ldots, e_n\}$ an orthonormal system. Let $W = span\{e_1, e_2, \ldots, e_n\}$ and $u \in V$ be any vector then $\tilde{u} = \sum_{i=1}^{n} \langle u, e_i \rangle e_i$ is the closest vector to $u$ in $W$. Moreover, $\tilde{u}$ is the unique such vector in $W$.*

## Proof.

For all $w \in W$,

$$||u - w||^2 = ||u - \tilde{u}||^2 + ||\tilde{u} - w||^2$$

and so $||u - \tilde{u}|| \leq ||u - w||$ for all $w \in W$.
To show uniqueness, suppose that $||u - \tilde{u}|| = ||u - w||$ for some $w \in W$ then $||\tilde{u} - w|| = 0$ and so $w = \tilde{u}$. □

# Infinite orthonormal systems

We now consider the situation of an inner product space, $V$, with $\dim(V) = \infty$ and consider orthonormal systems $\{e_1, e_2, \ldots\}$ consisting of infinitely many vectors.

## Definition (Convergence in norm)

Let $\{u_1, u_2, \ldots\}$ be an infinite sequence of vectors in the normed linear space $V$ and let $\{a_1, a_2, \ldots\}$ be a sequence of scalars. We say that the series

$$\sum_{n=1}^{\infty} a_n u_n$$

converges in norm to $w \in V$ if

$$\lim_{m\to\infty} ||w - \sum_{n=1}^{m} a_n u_n|| = 0\,.$$

# Closure and completeness

Two further properties are defined for an infinite orthonormal system $\{e_1, e_2, \dots\}$ in an inner product space $V$.

## Definition (Closed)

The system is called closed in $V$ if for all $u \in V$

$$\lim_{m \to \infty} ||u - \sum_{n=1}^{m} \langle u, e_n \rangle e_n|| = 0 \, .$$

## Definition (Complete)

The system is called complete in $V$ if the zero vector $u = \vec{0}$ is the only solution to the set of equations

$$\langle u, e_n \rangle = 0 \quad n = 1, 2, \dots \, .$$

# Remarks on closure and completeness

- It can be shown that a closed infinite orthonormal system $\{e_1, e_2, \ldots\}$ is necessarily complete (but not the converse).
- If a system is not closed then there must exist some $u \in V$ such that the linear combination

  $$\sum_{n=1}^{m} \langle u, e_n \rangle e_n$$

  cannot be made arbitrarily close to $u$, for all choices of $m$.
- If the system is closed it may still be that the required number of terms in the above linear combination for a "good" approximation is too great for practical purposes.
- Seeking alternative closed systems of orthonormal vectors may produce "better" approximations in the sense of requiring fewer terms for a given accuracy.

# Representing functions

In seeking to represent functions as linear combinations of simpler functions we shall need to consider spaces of functions with closed orthonormal systems.

## Definition (piecewise continuous)

A function is piecewise continuous if it is continuous, except at a finite number of points and at each such point of discontinuity, the right and left limits exists and are finite.

The space, $E$, of piecewise continuous functions $f : [-\pi, \pi] \to \mathbb{C}$ is seen to be a linear space, under the convention that we regard two functions in $E$ as identical if they are equal at all but a finite number of points.
For $f, g \in E$, then

$$\langle f, g \rangle = \frac{1}{\pi} \int_{-\pi}^{\pi} f(x)\overline{g(x)}dx$$

defines an inner product on $E$.

# A closed infinite orthonormal system for $E$

An important result is that

$$\left\{\frac{1}{\sqrt{2}}, \sin x, \cos x, \sin 2x, \cos 2x, \sin 3x, \cos 3x, \ldots\right\}$$

is a closed infinite orthonormal system in the space $E$.
Here we shall just demonstrate orthonormality and omit establishing that this system is closed.

Writing $||f|| = \sqrt{< f, f >}$ as the norm associated with our inner product, it can be establish that

$$||\frac{1}{\sqrt{2}}||^2 = 1$$

and similarly that for each $n = 1, 2, \ldots$

$$||\sin nx||^2 = ||\cos nx||^2 = 1$$

and that for $m, n \in \mathbb{N}$

- $\langle \frac{1}{\sqrt{2}}, \sin nx \rangle = 0$
- $\langle \frac{1}{\sqrt{2}}, \cos nx \rangle = 0$
- $\langle \sin mx, \cos nx \rangle = 0$
- $\langle \sin mx, \sin nx \rangle = 0$, $m \neq n$
- $\langle \cos mx, \cos nx \rangle = 0$, $m \neq n$.

# Fourier series

From our knowledge of closed orthonormal systems $\{e_1, e_2, \ldots\}$ we know that we can represent any function $f \in E$ by a linear combination

$$\sum_{n=1}^{\infty} \langle f, e_n \rangle e_n \, .$$

We now turn to consider the individual terms $\langle f, e_n \rangle e_n$ in the case of the closed orthonormal system

$$\left\{ \frac{1}{\sqrt{2}}, \sin x, \cos x, \sin 2x, \cos 2x, \sin 3x, \cos 3x, \ldots \right\} \, .$$

There are three cases, either $e_n = \frac{1}{\sqrt{2}}$ or $\sin nx$ or $\cos nx$. Recall that the vectors $e_n$ are actually functions
in $E = \{f : [-\pi, \pi] \to \mathbb{C} : f \text{ is piecewise continuous}\}$

If $e_n = 1/\sqrt{2}$ then

$$\langle f, e_n \rangle e_n = \frac{1}{\pi} \left( \int_{-\pi}^{\pi} f(t) \frac{1}{\sqrt{2}} dt \right) \frac{1}{\sqrt{2}} = \frac{1}{2\pi} \int_{-\pi}^{\pi} f(t) dt \, .$$

If $e_n = \sin nx$ then

$$\langle f, e_n \rangle e_n = \frac{1}{\pi} \left( \int_{-\pi}^{\pi} f(t) \sin nt \, dt \right) \sin nx \, .$$

If $e_n = \cos nx$ then

$$\langle f, e_n \rangle e_n = \frac{1}{\pi} \left( \int_{-\pi}^{\pi} f(t) \cos nt \, dt \right) \cos nx \, .$$

# Fourier coefficients

Thus the linear combination $\sum_{n=1}^{\infty} \langle f, e_n \rangle e_n$ becomes the familiar Fourier series for a function $f$, namely

$$\frac{a_0}{2} + \sum_{n=1}^{\infty} [a_n \cos nx + b_n \sin nx]$$

where

$$a_n = \frac{1}{\pi} \int_{-\pi}^{\pi} f(x) \cos nx \, dx, \quad n = 0, 1, 2, \ldots$$
$$b_n = \frac{1}{\pi} \int_{-\pi}^{\pi} f(x) \sin nx \, dx, \quad n = 1, 2, 3, \ldots .$$

Note how the constant term is written $a_0/2$ where $a_0 = \frac{1}{\pi} \int_{-\pi}^{\pi} f(x) dx$.

# Periodic functions

Our Fourier series

$$\frac{a_0}{2} + \sum_{n=1}^{\infty} [a_n \cos nx + b_n \sin nx]$$

defines a function, $g(x)$, say, that is $2\pi$-periodic in the sense that

$$g(x + 2\pi) = g(x), \qquad \text{for all } x \in \mathbb{R}\,.$$

Hence, it is convenient to extend $f \in E$ to a $2\pi$-periodic function defined on $\mathbb{R}$ instead of being restricted to $[-\pi, \pi]$.

# Even and odd functions

A particularly useful simplification occurs when the function $f \in E$ is either an even function, that is, for all $x$,

$$f(-x) = f(x)$$

or an odd function, that is, for all $x$,

$$f(-x) = -f(x)\,.$$

The following properties can be easily verified.

1. If $f, g$ are even then $fg$ is even
2. If $f, g$ are odd then $fg$ is even
3. If $f$ is even and $g$ is odd then $fg$ is odd
4. If $g$ is odd then for any $h > 0$ then $\int_{-h}^{h} g(x)dx = 0$
5. If $g$ is even then for any $h > 0$ then $\int_{-h}^{h} g(x)dx = 2\int_{0}^{h} g(x)dx$.

# Even functions and cosine series

Recall that the Fourier coefficients are given by

$$a_n = \frac{1}{\pi}\int_{-\pi}^{\pi} f(x)\cos nx\,dx, \quad n = 0, 1, 2, \ldots$$

$$b_n = \frac{1}{\pi}\int_{-\pi}^{\pi} f(x)\sin nx\,dx, \quad n = 1, 2, 3, \ldots$$

so if $f$ is even then they become

$$a_n = \frac{2}{\pi}\int_{0}^{\pi} f(x)\cos nx\,dx, \quad n = 0, 1, 2, \ldots$$

$$b_n = 0, \quad n = 1, 2, 3, \ldots .$$

# Odd functions and sine series

Similarly, the Fourier coefficients

$$a_n = \frac{1}{\pi} \int_{-\pi}^{\pi} f(x) \cos nx \, dx, \quad n = 0, 1, 2, \ldots$$
$$b_n = \frac{1}{\pi} \int_{-\pi}^{\pi} f(x) \sin nx \, dx, \quad n = 1, 23, \ldots,$$

for the case where $f$ is an odd function become

$$a_n = 0, \quad n = 0, 1, 2, \ldots$$
$$b_n = \frac{2}{\pi} \int_{0}^{\pi} f(x) \sin nx \, dx, \quad n = 1, 2, \ldots .$$

# Fourier series: examples I

Consider $f(x) = x$ for $x \in [-\pi, \pi]$ then $f$ is clearly odd and so we need to calculate a sine series with coefficients, $b_n$, $n = 1, 2, \ldots$ given by

$$
\begin{aligned}
b_n &= \frac{2}{\pi}\int_0^{\pi} x \sin nx \, dx = \frac{2}{\pi}\left\{\left[-x\frac{\cos nx}{n}\right]_0^{\pi} + \int_0^{\pi} \frac{\cos nx}{n} dx\right\} \\
&= \frac{2}{\pi}\left\{-\pi\frac{(-1)^n}{n} + \left[\frac{\sin nx}{n^2}\right]_0^{\pi}\right\} \\
&= \frac{2}{\pi}\left\{-\pi\frac{(-1)^n}{n} + 0\right\} = \frac{2(-1)^{n+1}}{n} \, .
\end{aligned}
$$

Hence the Fourier series of $f(x) = x$ is

$$\sum_{n=1}^{\infty} \frac{2(-1)^{n+1}}{n} \sin nx \, .$$

Observe that the series does not agree with $f(x)$ at $x = \pm\pi$ — a matter that we shall return to later.

# Fourier series: examples II

Now suppose $f(x) = |x|$ for $x \in [-\pi, \pi]$ which is clearly an even function so we need to construct a cosine series with coefficients

$$a_0 = \frac{2}{\pi}\int_0^\pi x dx = \frac{2}{\pi}\frac{\pi^2}{2} = \pi$$

and for $n = 1, 2, \ldots$

$$a_n = \frac{2}{\pi}\int_0^\pi x\cos nx\, dx = \frac{2}{\pi}\left\{\left[\frac{x\sin nx}{n}\right]_0^\pi - \int_0^\pi \frac{\sin nx}{n}dx\right\}$$
$$= \frac{2}{\pi}\left\{\left[\frac{\cos nx}{n^2}\right]_0^\pi\right\} = \frac{2}{\pi}\left\{\frac{(-1)^n - 1}{n^2}\right\} = \begin{cases} -\frac{4}{\pi n^2} & n \text{ is odd} \\ 0 & n \text{ is even} \end{cases} .$$

Hence, the Fourier series of $f(x) = |x|$ is

$$\frac{\pi}{2} - \sum_{k=1}^{\infty} \frac{4}{\pi(2k-1)^2}\cos(2k-1)x\, .$$

# Complex Fourier series I

We have used real-valued functions sin $nx$ and cos $nx$ as our orthonormal system for the linear space $E$ but we can also use complex-valued functions. In this case, we should amend our inner product to

$$\langle f, g\rangle = \frac{1}{2\pi}\int_{-\pi}^{\pi} f(x)\overline{g(x)}dx\,.$$

A suitable orthonormal system in this case is the collection of functions

$$\left\{1, e^{ix}, e^{-ix}, e^{i2x}, e^{-i2x}, \dots\right\}\,.$$

Then if $f \in E$ we have a representation, known as the complex Fourier series of $f \in E$, given by

$$\sum_{n=-\infty}^{\infty} c_n e^{inx}$$

where

$$c_n = \frac{1}{2\pi}\int_{-\pi}^{\pi} f(x)e^{-inx}dx, \quad , n = 0, \pm 1, \pm 2, \dots\,.$$

# Complex Fourier series II

Euler's formula ($e^{ix} = \cos x + i \sin x$) gives for $n = 1, 2, \ldots$ that

$$\begin{aligned} e^{inx} &= \cos nx + i \sin nx \\ e^{-inx} &= \cos nx - i \sin nx \end{aligned}$$

and $e^{i0x} = 1$. Using these relations it can be shown that for $n = 1, 2, \ldots$

$$c_n = \frac{a_n - ib_n}{2}, \qquad c_{-n} = \frac{a_n + ib_n}{2} .$$

Hence,

$$a_n = c_n + c_{-n}, \qquad b_n = i(c_n - c_{-n})$$

and

$$c_0 = \frac{1}{2\pi} \int_{-\pi}^{\pi} f(x) e^{-i0x} dx = \frac{1}{2\pi} \int_{-\pi}^{\pi} f(x) dx = \frac{a_0}{2} .$$

# Pointwise convergence and Dirichlet's conditions

The closure property of the trigonometric orthonormal system guarantees that the Fourier series for any function $f \in E$ converges in norm to $f$. That is,

$$\lim_{m\to\infty} ||f(x) - \left(\frac{a_0}{2} + \sum_{n=1}^{m}[a_n \cos nx + b_n \sin nx]\right)|| = 0$$

or, equivalently,

$$\lim_{m\to\infty} \int_{-\pi}^{\pi} \left| f(x) - \left(\frac{a_0}{2} + \sum_{n=1}^{m}[a_n \cos nx + b_n \sin nx]\right)\right|^2 dx = 0\,.$$

As we have already seen in the example of $f(x) = x$, this does not imply convergence to $f(x)$ at every point $x$.

# The Dirichlet conditions

We now consider conditions on the space of functions that allow us to determine how the Fourier series behaves at individual points $x$.

## Definition (Dirichlet conditions)

We define a subspace, $E'$, of $E$ by the Dirichlet conditions:

1. $f \in E$
2. For all $x \in [-\pi, \pi)$ both the left and right derivatives exist (and are finite).

Recall, that in the space $E$ each function has a left and right limit at every point. Let these values be $f(x-)$ and $f(x+)$, respectively.

## Theorem (Dirichlet's theorem)

*For all $x \in [-\pi, \pi]$ the Fourier series of a function $f \in E'$ converges to the value of the expression*

$$\frac{f(x-) + f(x+)}{2}.$$

- Here we should consider $f$ not just defined on $[-\pi, \pi]$ but also make it $2\pi$-periodic to handle the end points $\pm\pi$ correctly.
- Recall that functions $f \in E$ can have at most a finite number of points of discontinuity (that is, points where $f(x-)$ and $f(x+)$ differ).
- Hence, we can conclude that if a function $f$ satisfies the Dirichlet conditions it's Fourier series converges to $f$ at all points where $f$ is continuous and at points of discontinuity it converges to the average of the left and right hand limits. This was indeed the case in our earlier example where $f(x) = x$.

# General intervals

We have so far considered functions defined on the interval $[-\pi, \pi]$ but we may readily extend our approach to a general interval of the form $[a, b]$. If we define $E[a, b]$ to be the space of piecewise continuous functions $f : [a, b] \to \mathbb{C}$ then we may define the Fourier series of $f \in E[a, b]$ as

$$\frac{a_0}{2} + \sum_{n=1}^{\infty} \left[ a_n \cos \frac{2n\pi x}{(b-a)} + b_n \sin \frac{2n\pi x}{(b-a)} \right]$$

where

$$a_n = \frac{2}{(b-a)} \int_a^b f(x) \cos \frac{2n\pi x}{(b-a)} dx, \quad n = 0, 1, 2, \dots$$

$$b_n = \frac{2}{(b-a)} \int_a^b f(x) \sin \frac{2n\pi x}{(b-a)} dx, \quad n = 1, 2, 3, \dots .$$

This may be justified by showing, for example, that

$$\left\{ \frac{1}{\sqrt{2}}, \cos \frac{2n\pi x}{(b-a)}, \sin \frac{2n\pi x}{(b-a)} \quad \text{for } n = 1, 2, \ldots \right\}$$

is an infinite orthonormal system for functions in $E[a, b]$ with respect to the inner product

$$\langle f, g \rangle = \frac{2}{(b-a)} \int_a^b f(x)\overline{g(x)}dx\,.$$

# Fourier transforms

# Introduction

- We have seen how functions $f : [-\pi, \pi] \to \mathbb{C}, f \in E$ can be represented in alternative ways using closed orthonormal systems, such as

  $$\sum_{n=-\infty}^{\infty} c_n e^{inx}$$

  where

  $$c_n = \frac{1}{2\pi} \int_{-\pi}^{\pi} f(x) e^{-inx} dx \quad n = 0, \pm 1, \pm 2, \ldots .$$

  The domain $[-\pi, \pi]$ can be swapped for a general interval $[a, b]$ and the function can be regarded as $L$-periodic and defined for all $\mathbb{R}$, where $L = (b - a) < \infty$ is the length of the interval.
- We shall now consider the situation where $f : \mathbb{R} \to \mathbb{C}$ may be a non-periodic function.

# Fourier transform

## Definition (Fourier transform)

For $f : \mathbb{R} \to \mathbb{C}$ define the Fourier transform of $f$ to be the function $F : \mathbb{R} \to \mathbb{C}$ given by

$$F(\omega) = \mathcal{F}_{[f]}(\omega) = \frac{1}{2\pi} \int_{-\infty}^{\infty} f(x) e^{-i\omega x} dx$$

whenever the integral exists.

We shall use the notation $F(\omega)$ or $\mathcal{F}_{[f]}(\omega)$ as convenient. The notation $\hat{f}(\omega)$ is also seen widely in the literature.

For functions $f : \mathbb{R} \to \mathbb{C}$ define the two properties

1. piecewise continuous: if $f$ is piecewise continuous on every finite interval. Thus $f$ may have an infinite number of discontinuities but only a finite number in any subinterval.
2. absolutely integrable: if

$$\int_{-\infty}^{\infty} |f(x)| dx < \infty$$

Let $G(\mathbb{R})$ be the collection of all functions $f : \mathbb{R} \to \mathbb{C}$ that are piecewise continuous and absolutely integrable.

# Immediate properties

It may be shown that $G(\mathbb{R})$ is a linear space over the scalars $\mathbb{C}$ and that for $f \in G(\mathbb{R})$

1. $F(\omega)$ is defined for all $\omega \in \mathbb{R}$
2. $F$ is a continuous function
3. $\lim_{\omega \to \pm\infty} F(\omega) = 0$

# Examples

For $a > 0$, let $f(x) = e^{-a|x|}$ then

$$
\begin{aligned}
F(\omega) &= \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-a|x|} e^{-i\omega x} dx \\
&= \frac{1}{2\pi} \left\{ \int_{0}^{\infty} e^{-ax} e^{-i\omega x} dx + \int_{-\infty}^{0} e^{ax} e^{-i\omega x} dx \right\} \\
&= \frac{1}{2\pi} \left\{ - \left[ \frac{e^{-(a+i\omega)x}}{a+i\omega} \right]_{0}^{\infty} + \left[ \frac{e^{(a-i\omega)x}}{a-i\omega} \right]_{-\infty}^{0} \right\} \\
&= \frac{1}{2\pi} \left\{ \frac{1}{a+i\omega} + \frac{1}{a-i\omega} \right\} \\
&= \frac{a}{\pi(a^2+\omega^2)} \,.
\end{aligned}
$$

# Properties

Several properties of the Fourier transform are very helpful in calculations.
First, note that by the linearity of integrals we have that if $f, g \in G(\mathbb{R})$ and $a, b \in \mathbb{C}$ then

$$\mathcal{F}_{[af+bg]}(\omega) = a\mathcal{F}_{[f]}(\omega) + b\mathcal{F}_{[g]}(\omega)$$

and $af + bg \in G(\mathbb{R})$.
Secondly, if $f$ is real-valued only then

$$F(-\omega) = \overline{F(\omega)}\,.$$

# Even and odd real-valued functions

## Theorem

*If $f \in G(\mathbb{R})$ is an even real-valued function then $F$ is even and real-valued. If $f$ is an odd real-valued function then $F$ is odd and purely imaginary.*

## Proof.

Suppose that $f$ is even and real-valued then

$$\begin{aligned} F(\omega) &= \frac{1}{2\pi}\int_{-\infty}^{\infty} f(x)e^{-i\omega x}dx \\ &= \frac{1}{2\pi}\int_{-\infty}^{\infty} f(x)[\cos\omega x - i\sin\omega x]dx \\ &= \frac{1}{2\pi}\int_{-\infty}^{\infty} f(x)\cos\omega x dx\,. \end{aligned}$$

Hence, $F$ is real-valued and even (the imaginary part has vanished and both $f$ and $\cos\omega x$ are themselves even functions). The second part follows similarly. □

# Shift property

## Theorem

*Let $f \in G(\mathbb{R})$ and $a, b \in \mathbb{R}$ with $a \neq 0$ and define*

$$g(x) = f(ax + b)$$

*then $g \in G(\mathbb{R})$ and*

$$\mathcal{F}_{[g]}(\omega) = \frac{1}{|a|} e^{i\omega b/a} \mathcal{F}_{[f]}\left(\frac{\omega}{a}\right)$$

# Proof

Set $y = ax + b$ so for $a > 0$ then

$$\mathcal{F}_{[g]}(\omega) = \frac{1}{2\pi} \int_{-\infty}^{\infty} f(y) e^{-i\omega(\frac{y-b}{a})} \frac{dy}{a}$$

and for $a < 0$

$$\mathcal{F}_{[g]}(\omega) = -\frac{1}{2\pi} \int_{-\infty}^{\infty} f(y) e^{-i\omega(\frac{y-b}{a})} \frac{dy}{a} .$$

Hence,

$$\mathcal{F}_{[g]}(\omega) = \frac{1}{|a|} e^{i\omega b/a} \frac{1}{2\pi} \int_{-\infty}^{\infty} f(y) e^{-i\omega y/a} dy = \frac{1}{|a|} e^{i\omega b/a} \mathcal{F}_{[f]}\left(\frac{\omega}{a}\right) .$$

□

# Special cases

Two special cases of the shift property are worth highlighting.

1. $a \neq 0$ and $b = 0$ so $g(x) = f(ax)$ then

$$\mathcal{F}_{[g]}(\omega) = \frac{1}{|a|}\mathcal{F}_{[f]}\left(\frac{\omega}{a}\right) .$$

2. $a = 1$ so $g(x) = f(x + b)$ then

$$\mathcal{F}_{[g]}(\omega) = e^{i\omega b}\mathcal{F}_{[f]}(\omega) .$$

## Theorem

*For $f \in G(\mathbb{R})$ and $c \in \mathbb{R}$ then*

$$\mathcal{F}_{[e^{icx}f(x)]}(\omega) = \mathcal{F}_{[f]}(\omega - c)\,.$$

## Proof.

$$\begin{aligned}\mathcal{F}_{[e^{icx}f(x)]}(\omega) &= \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{icx} f(x) e^{-i\omega x} dx \\ &= \frac{1}{2\pi}\int_{-\infty}^{\infty} f(x) e^{-i(\omega - c)x} dx \\ &= \mathcal{F}_{[f]}(\omega - c)\,.\end{aligned}$$

□

# Modulation property

## Theorem

*For $f \in G(\mathbb{R})$ and $c \in \mathbb{R}$ then*

$$\mathcal{F}_{[f(x)\cos cx]}(\omega) = \frac{\mathcal{F}_{[f]}(\omega - c) + \mathcal{F}_{[f]}(\omega + c)}{2}$$

$$\mathcal{F}_{[f(x)\sin cx]}(\omega) = \frac{\mathcal{F}_{[f]}(\omega - c) - \mathcal{F}_{[f]}(\omega + c)}{2i}.$$

## Proof.

We have that

$$\begin{aligned}\mathcal{F}_{[f(x)\cos cx]}(\omega) &= \mathcal{F}_{\left[f(x)\frac{e^{icx}+e^{-icx}}{2}\right]}(\omega)\\ &= \frac{1}{2}\mathcal{F}_{[f(x)e^{icx}]}(\omega) + \frac{1}{2}\mathcal{F}_{[f(x)e^{-icx}]}(\omega)\\ &= \frac{\mathcal{F}_{[f]}(\omega - c) + \mathcal{F}_{[f]}(\omega + c)}{2}.\end{aligned}$$

Similarly, for $\mathcal{F}_{[f(x)\sin cx]}(\omega)$. □

# Derivatives

There are further properties relating to the Fourier transform of derivatives that we shall state here but omit further proofs.

## Theorem

*If $f$ is a continuous function and $f, f' \in G(\mathbb{R})$ then*

$$\mathcal{F}_{[f']}(\omega) = i\omega \mathcal{F}_{[f]}(\omega)\,.$$

# Inverse Fourier transform

We have studied the Fourier transform. There is also an inverse operation of recovering a function $f$ given the function $F(\omega) = \mathcal{F}_{[f]}(\omega)$ which takes the form

$$f(x) = \int_{-\infty}^{\infty} \mathcal{F}_{[f]}(\omega) e^{i\omega x} d\omega \,.$$

More precisely, and recalling Dirichlet's theorem for Fourier series, the following holds.

## Theorem (Inverse Fourier transform)

*If $f \in G(\mathbb{R})$ then for every point $x \in \mathbb{R}$ where the one-sided derivatives exist*

$$\frac{f(x-) + f(x+)}{2} = \lim_{M \to \infty} \int_{-M}^{M} \mathcal{F}_{[f]}(\omega) e^{i\omega x} d\omega \,.$$

# Convolution

An important operation between two functions in signal processing applications is convolution defined as follows.

## Definition (Convolution)

If $f$ and $g$ are two functions $\mathbb{R} \to \mathbb{C}$ then the convolution function, written $f * g$, is given by

$$(f * g)(x) = \int_{-\infty}^{\infty} f(x - y)g(y)dy$$

whenever the integral exists.

It may be shown that the convolution operation is commutative, that is $f * g = g * f$.

# Fourier transforms and convolutions

The importance of Fourier transform techniques in signal processing rests, in part, on the following result that leads to much simpler descriptions and mathematical formulae in the Fourier domain.

## Theorem (Convolution theorem)

*For* $f, g \in G(\mathbb{R})$ *then*

$$\mathcal{F}_{[f*g]}(\omega) = 2\pi \mathcal{F}_{[f]}(\omega) \cdot \mathcal{F}_{[g]}(\omega) \, .$$

# Proof

We have that

$$
\begin{aligned}
\mathcal{F}_{[f*g]}(\omega) &= \frac{1}{2\pi}\int_{-\infty}^{\infty}(f*g)(x)e^{-i\omega x}dx \\
&= \frac{1}{2\pi}\int_{-\infty}^{\infty}\left(\int_{-\infty}^{\infty} f(x-y)g(y)dy\right)e^{-i\omega x}dx \\
&= \frac{1}{2\pi}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x-y)e^{-i\omega(x-y)}g(y)e^{-i\omega y}dxdy \\
&= \int_{-\infty}^{\infty}\left(\frac{1}{2\pi}\int_{-\infty}^{\infty} f(x-y)e^{-i\omega(x-y)}dx\right)g(y)e^{-i\omega y}dy \\
&= \mathcal{F}_{[f]}(\omega)\int_{-\infty}^{\infty} g(y)e^{-i\omega y}dy \\
&= 2\pi\mathcal{F}_{[f]}(\omega)\cdot\mathcal{F}_{[g]}(\omega)\,.
\end{aligned}
$$

□

# Some signal processing applications

We first note two types of limitations on functions.

## Definition (Time-limited)

A function $f$ is time-limited if

$$f(x) = 0 \quad \text{for all } |x| \geq M$$

for some constant $M$.

## Definition (Band-limited)

A function $f \in G(\mathbb{R})$ is band-limited if

$$\mathcal{F}_{[f]}(\omega) = 0 \quad \text{for all } |\omega| \geq L$$

for some constant $L$.

Let us first calculate the Fourier transform of

$$f(x) = \begin{cases} 1 & a \leq x \leq b \\ 0 & \text{otherwise}\,. \end{cases}$$

We have that

$$\begin{aligned} F(\omega) &= \frac{1}{2\pi}\int_{-\infty}^{\infty} f(x)e^{-i\omega x}dx = \frac{1}{2\pi}\int_a^b e^{-i\omega x}dx \\ &= \left[\frac{1}{2\pi}\frac{e^{-i\omega x}}{-i\omega}\right]_a^b = \frac{e^{-i\omega a} - e^{-i\omega b}}{2\pi i\omega}\,. \end{aligned}$$

If we set $a = -b$ with $b > 0$ then

$$F(\omega) = \frac{e^{i\omega b} - e^{-i\omega b}}{2\pi i\omega} = \frac{\sin\omega b}{\omega\pi}\,.$$

# Low-pass filters

Suppose that $f \in G(\mathbb{R})$ with Fourier transform $F(\omega)$ and choose a constant $L > 0$. Define

$$F_L(\omega) = \begin{cases} F(\omega) & |\omega| \leq L \\ 0 & |\omega| > L\,. \end{cases}$$

We wish to find $f_L$ such that $\mathcal{F}_{[f_L]} = F_L$, that is, a function band-limited by $L$ whose Fourier transform equals $F$ in $[-L, L]$.
Rewrite $F_L(\omega) = F(\omega)G_L(\omega)$ where

$$G_L(\omega) = \begin{cases} 1 & |\omega| \leq L \\ 0 & |\omega| > L\,. \end{cases}$$

We will now use the convolution theorem to find $f_L$.

By the inverse transform theorem we have that for $|x| \neq L$

$$G_L(x) = \int_{-\infty}^{\infty} \frac{\sin \omega L}{\omega \pi} e^{i\omega x} d\omega$$

But $G_L$ is clearly an even function so

$$G_L(x) = \int_{-\infty}^{\infty} \frac{\sin \omega L}{\omega \pi} e^{-i\omega x} d\omega$$

and if we interchange the variables $x$ and $\omega$ we have

$$G_L(\omega) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \frac{2 \sin Lx}{x} e^{-i\omega x} dx \, .$$

This says that if $g_L(x) = \frac{2 \sin Lx}{x}$ then $\mathcal{F}_{[g_L]}(\omega) = G_L(\omega)$.

In terms of convolutions we have

$$\begin{aligned} f_L &= \frac{1}{2\pi}(f * g_L) \\ f_L(x) &= \frac{1}{2\pi}\int_{-\infty}^{\infty} f(y)\frac{2\sin(L(x-y))}{x-y}dy \\ &= \frac{1}{\pi}\int_{-\infty}^{\infty}\frac{f(y)\sin(L(x-y))}{x-y}dy \end{aligned}$$

In particular, if $f \in G(\mathbb{R})$ is such that $\mathcal{F}_{[f]}(\omega) = 0$ for $|\omega| \geq L$ then $f$ satisfies

$$f(x) = \frac{1}{\pi}\int_{-\infty}^{\infty}\frac{f(y)\sin(L(x-y)))}{x-y}dy\,.$$

# Shannon sampling theorem

## Theorem (Shannon sampling theorem)

*If $f \in G(\mathbb{R})$ is band-limited by the constant $L$ then*

$$f(x) = \sum_{n=-\infty}^{\infty} f\left(\frac{n\pi}{L}\right) \frac{\sin(Lx - n\pi)}{Lx - n\pi}\,.$$

## Proof

Set $F(\omega) = \mathcal{F}_{[f]}(\omega)$ and use the inverse Fourier transform theorem to give

$$f(x) = \int_{-\infty}^{\infty} F(\omega) e^{i\omega x} d\omega = \int_{-L}^{L} F(\omega) e^{i\omega x} d\omega\,.$$

So, taking $x = \frac{n\pi}{L}$ for $n \in \mathbb{Z}$ we get

$$f\left(\frac{n\pi}{L}\right) = \int_{-L}^{L} F(\omega) e^{i\omega n\pi/L} d\omega\,.$$

Consider the complex Fourier series of $F$ restricted to $[-L, L]$ given by

$$\sum_{n=-\infty}^{\infty} c_n e^{-in\pi\omega/L}$$

where the coefficients are

$$c_n = \frac{1}{2L}\int_{-L}^{L} F(\omega) e^{in\pi\omega/L} d\omega = \frac{1}{2L} f\left(\frac{n\pi}{L}\right)$$

Thus, since $f$ is band-limited by $L$

$$F(\omega) = \sum_{n=-\infty}^{\infty} c_n e^{-in\pi\omega/L} G_L(\omega)\,.$$

Hence,

$$F(\omega) = \frac{1}{2L}\sum_{n=-\infty}^{\infty} f\left(\frac{n\pi}{L}\right) e^{-in\pi\omega/L} G_L(\omega)\,.$$

But we have seen that $G_L(\omega) = \mathcal{F}_{[\frac{2\sin Lx}{x}]}(\omega)$ hence using the shift formula

$$e^{-in\pi\omega/L} G_L(\omega) = \mathcal{F}_{[g_{L,n}]}(\omega)$$

where

$$g_{L,n}(x) = \frac{2\sin(Lx - n\pi)}{x - \frac{n\pi}{L}} .$$

Putting this all together we have that

$$F(\omega) = \frac{1}{2L} \sum_{n=-\infty}^{\infty} f\left(\frac{n\pi}{L}\right) \mathcal{F}_{[g_{L,n}]}(\omega)$$

and taking inverse transforms

$$f(x) = \frac{1}{2L} \sum_{n=-\infty}^{\infty} f\left(\frac{n\pi}{L}\right) g_{L,n}(x) = \sum_{n=-\infty}^{\infty} f\left(\frac{n\pi}{L}\right) \frac{\sin(Lx - n\pi)}{Lx - n\pi} .$$

# Remarks on Shannon's sampling theorem

- The theorem says that band-limited functions by a constant $L$ (that is, $\mathcal{F}_{[f]}(\omega) = 0$ for $|\omega| > L$) are completely determined by their values at evenly spaced points a distance $\frac{\pi}{L}$ apart.
- Moreover, we may recover the function exactly given only it's values at this sequence of points.
- It may be shown that the functions

  $$\frac{\sin(Lx - n\pi)}{Lx - n\pi}$$

  for $n \in \mathbb{Z}$ form an orthonormal system with inner product

  $$\langle f, g \rangle = \frac{L}{\pi} \int_{-\infty}^{\infty} f(x) \overline{g(x)} dx \, .$$

# Discrete Fourier Transforms

We now shift attention from functions defined on intervals or on the whole of $\mathbb{R}$ to sequences of values $f[0], f[1], \ldots, f[N-1]$ and consider how we might represent them.

An important result in this area of discrete transforms is that the vectors $\{e_0, e_1, \ldots, e_{N-1}\}$ form an orthogonal system in the space $\mathbb{C}^N$ with the usual inner product where the $n^{th}$ component of $e_k$ is given by

$$(e_k)_n = e^{2\pi i n k/N} \quad n = 0, 1, 2, \ldots, N-1\,.$$

and $k = 0, 1, 2, \ldots, N-1$.

Applying the usual inner product

$$\langle u, v \rangle = \sum_{n=0}^{N-1} u[n]\overline{v[n]}$$

we find that

$$||e_k||^2 = \langle e_k, e_k \rangle = N\,.$$

In fact, using $\{e_0, e_1, \ldots, e_{N-1}\}$ we can represent any sequence $f = (f[0], f[1], \ldots, f[N-1]) \in \mathbb{C}^N$ by

$$f = \frac{1}{N} \sum_{k=0}^{N-1} \langle f, e_k \rangle e_k\,.$$

# Orthogonality

We shall show orthogonality of the vectors $e_k$ by considering the $N$ distinct complex roots of the equation $z^N = 1$. Put $w = e^{2\pi i/N}$ then the $N$ distinct roots $z_j$ ($j = 0, 1, \ldots, N-1$) of $z^N = 1$ are

$$z_j = e^{2\pi ij/N} = w^j \, .$$

Now for an arbitrary integer $n$

$$\begin{aligned} \frac{1}{N}\sum_{k=0}^{N-1} e^{2\pi ink/N} &= \frac{1}{N}\sum_{k=0}^{N-1} w^{nk} \\ &= \begin{cases} 1 & \text{if } n \text{ is an integer multiple of } N \\ \frac{1}{N}\frac{1-w^{nN}}{1-w^n} = 0 & \text{otherwise} \, . \end{cases} \end{aligned}$$

Thus,

$$\begin{aligned}
\langle e_a, e_b \rangle &= \sum_{k=0}^{N-1} e^{2\pi i k a/N} e^{-2\pi i k b/N} \\
&= \sum_{k=0}^{N-1} e^{2\pi i k (a-b)/N} \\
&= \begin{cases} N & \text{if } (a-b) \text{ is a multiple of } N \\ 0 & \text{otherwise} \, . \end{cases}
\end{aligned}$$

# Definition (Discrete Fourier Transform/DFT)

The sequence $F[k]$, $k \in \mathbb{Z}$, defined by

$$F[k] = \langle f, e_k \rangle = \sum_{n=0}^{N-1} f[n] e^{-2\pi i n k / N}$$

is called the $N$-point Discrete Fourier Transform of $f[n]$

Thus, for $n = 0, 1, 2, \ldots, N-1$,

$$f[n] = \frac{1}{N} \sum_{k=0}^{N-1} F[k] e^{2\pi i n k / N} .$$

# Periodicity

Note that the sequence $F[k]$ has period $N$ since

$$F[k+N] = \sum_{n=0}^{N-1} f[n] e^{-2\pi i n(k+N)/N} = \sum_{n=0}^{N-1} f[n] e^{-2\pi i n k/N} = F[k]$$

using the relation

$$e^{-2\pi i n(k+N)/N} = e^{-2\pi i n k/N} e^{-2\pi i n} = e^{-2\pi i n k/N}\,.$$

# Properties of the DFT

The DFT satisfies a range of similar properties to those of the FT relating to linearity, and shifts in either the $n$ or $k$ domain. However, the convolution operation is defined a little differently.

## Definition (Cyclical convolution)

The cyclical convolution of two periodic sequences $f[n]$ and $g[n]$ of period $N$ is defined as

$$(f * g)[n] = \sum_{m=0}^{N-1} f[m]g[n-m] \, .$$

It can then be shown that the DFT of $f * g$ is the product $F[k]G[k]$ where $F$ and $G$ are the DFTs of $f$ and $g$, respectively.

# Fast Fourier Transforms

# Fast Fourier Transform

The Fast Fourier Transform is not a new transform but a particular numerical algorithm for computing the DFT.
Since

$$
\begin{aligned}
F[k] &= \sum_{n=0}^{N-1} f[n] e^{-2\pi ink/N} \\
&= f[0] + f[1]e^{-2\pi ik/N} + \cdots + f[N-1]e^{-2\pi ik(N-1)/N}
\end{aligned}
$$

we can see that in order to compute $F[k]$ we need to do about $2N$ (complex) additions and multiplications. To compute $F[k]$ in this way for all $k = 0, 1, 2, \ldots, N-1$ would require about $2N^2$ such operations. In practice, where DFTs are computed for a large number of points $N$, faster algorithms have been developed. Most approaches are based on the factorization of $N$ into prime factors and are known collectively as Fast Fourier Transforms, or FFT. In most popular methods $N$ is supposed to be a power of 2.

# Fast algorithms for the DFT

In 1965, James W. Cooley and John W. Tukey published a new and substantially faster algorithm for computing the DFT than the direct $N^2$ approach.

They showed that when $N$ is a composite number with $N = P_1 P_2 \cdot P_m$ then it is possible to reduce the cost of computing the DFT of a vector of length $N$ from

$$N^2 = N(P_1 P_2 \cdots P_m) \quad \text{to} \quad N((P_1 - 1) + (P_2 - 1) + \cdots + (P_m - 1))$$

complex operations. In the case when $P_1 = P_2 = \cdots = P_m = 2$ then this reduces from $N^2 = 2^{2m}$ to $2^m \cdot m = N \log_2 N$.

For example, if $N = 1024 = 2^{10}$ then there is a 100 fold improvement from $N^2 = 1,048,576$ to $N \log_2 N = 10,240$.

See: J.W. Cooley and J.W. Tukey. (1965) An algorithm for the machine computation of complex Fourier series, *Math. Comp*, 19, 297–301.

We shall not derive any of the details here but instead give an impression of how the method operates.
First, the task of computing the DFT can be represented with matrices as

$$F = Af$$

but where the $N \times N$ matrix, $A$, has a great deal of internal structure. Cooley and Tukey exploited this structure in the case when $N = 2^m$ to rewrite $A$ as a product of matrices each of which is sparse

$$A = M_m M_{m-1} \cdots M_1 B\,.$$

Since each of these matrices contains only a small number of non-zero entries the effective number of complex operations is much reduced compared to working with $A$ itself.

# Wavelet Transforms

# Wavelets

Wavelets are a further method of representing functions that has received much interest in applied fields over the last several decades. The approach fits into the general scheme of expansion using orthonormal functions. Here we expand functions $f(x)$ in terms of a doubly-infinite series

$$f(x) = \sum_{j=-\infty}^{\infty} \sum_{k=-\infty}^{\infty} d_{jk} \Psi_{jk}(x)$$

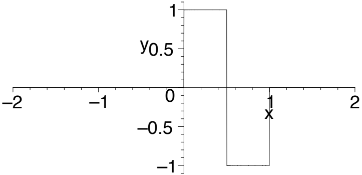
where $\Psi_{jk}(x)$ are the orthonormal functions.
The orthonormal functions arise from shifting and scaling operations applied to a single function, $\Psi(x)$, known as the mother wavelet.
The orthonormal functions are given for integers $j$ and $k$ by

$$\Psi_{jk}(x) = 2^{j/2}\Psi(2^j x - k)$$

# The Haar wavelet

A common example is the Haar wavelet whose mother function is both localised and oscillatory defined by

$$\Psi(x) = \begin{cases} 1 & \text{if } 0 \leq x < \frac{1}{2}\,, \\ -1 & \text{if } \frac{1}{2} \leq x < 1\,, \\ 0 & \text{otherwise}\,. \end{cases}$$



# Wavelet dilations and translations

The Haar mother wavelet oscillates and has a width (or scale) of one. The dyadic dilates of $\Psi(x)$, namely,

$$\ldots, \Psi(2^{-2}x), \Psi(2^{-1}x), \Psi(x), \Psi(2x), \Psi(2^{2}x), \ldots$$

have widths

$$\ldots, 2^{2}, 2^{1}, 1, 2^{-1}, 2^{-2}, \ldots$$

respectively. Since the dilate $\Psi(2^{j}x)$ has width $2^{-j}$, its translates

$$\Psi(2^{j}x - k) = \Psi(2^{j}(x - k2^{-j})), \quad k = 0, \pm 1, \pm 2, \ldots$$

will cover the whole $x$-axis. The collection of coefficients $d_{jk}$ are termed the Discrete wavelet transform, or DWT, of the function $f(x)$. Just as with Fourier transforms there are fast implementations that exploit structure.

# Interpretation of $d_{jk}$

How should we intrepret the values $d_{jk}$?
Since the Haar wavelet function $\Psi(2^j x - k)$ vanishes except when

$$0 \leq 2^j x - k < 1\,, \quad \text{that is} \quad k2^{-j} \leq x < (k+1)2^{-j}$$

we see that $d_{jk}$ gives us information about the behaviour of $f$ near the point $x = k2^{-j}$ measured on the scale of $2^{-j}$.
For example, the coefficients $d_{-10,k}$, $k = 0, \pm 1, \pm 2, \ldots$ correspond to variations of $f$ that take place over intervals of length $2^{10} = 1024$ while the coefficients $d_{10,k}$ $k = 0, \pm 1, \pm 2, \ldots$ correspond to fluctuations of $f$ over intervals of length $2^{-10}$.
These observations help explain how the discrete wavelet transform can be an exceptionally efficient scheme for representing functions.

# Comparison with Fourier analysis

Some of the practical motivations underlying the use of the orthonormal functions such as Fourier analysis or wavelet analysis are

- improved understanding,
- denoising signals, and
- data compression.

By representation of signals or functions in other forms these tasks become easier or more effective.
The approach taken with Fourier analysis represents signals in terms of trigonometric functions and as such is particularly suited to situations where the signal is relatively smooth and is not of limited extent.

# Properties of naturally arising data

Much naturally arising data has been found to be better represented using wavelets which are better able to cope with discontinuities and where the signal is of local extent. Generally, the efficiency of the representation depends on the types of signal involved. If your signal contains

- discontinuities (in both the signal and its derivatives), or
- varying frequency behaviour

then wavelets are likely to represent the signal more efficiently than is possible with Fourier analysis.

# Other classes of wavelets

- One of the most useful features of wavelets is the ease with which a scientist can select the wavelet functions adapted for the given problem.
- In fact, the Haar mother wavelet is perhaps the simplest of a very wide class of possible wavelet systems used in practice today.
- Many applied fields have started to make use of wavelets including astronomy, acoustics, signal and image processing, neurophysiology, music, magnetic resonance imaging, speach discrimination, optics, fractals, turbulence, earthquake prediction, radar, human vision, etc.

# Probability methods

# Outline

- Probability methods
  - Review of elementary probability theory (1 lecture)
  - Probability generating functions (1 lecture)
  - Elementary stochastic processes (2 lectures)
  - Limits and inequalities (3 lectures)
  - Markov chains (3 lectures)

# Review of elementary probability theory

# Random experiments

We will describe randomness by conducting experiments (or trials) with uncertain outcomes.

The set of all possible outcomes of an experiment is called the sample space and is denoted by $\Omega$.

We identify random events with particular subsets of $\Omega$ and write

$$\mathcal{F} = \{E \subseteq \Omega \,:\, E \text{ is an event}\}$$

for the collection of possible events.

For each such random event, $E \in \mathcal{F}$, we will attach a probability $\mathbb{P}(E) \in [0, 1]$.

# Event spaces

We formalize the notion of an event space, $\mathcal{F}$, by requiring the following to hold.

## Definition (Event space)

1. $\mathcal{F}$ is non-empty
2. $E \in \mathcal{F} \Rightarrow \Omega \setminus E \in \mathcal{F}$
3. $E_i \in \mathcal{F}$ for $i \in I \Rightarrow \cup_{i \in I} E_i \in \mathcal{F}$

## Example

$\Omega$ any set and $\mathcal{F} = \mathcal{P}(\Omega)$, the power set of $\Omega$.

## Example

$\Omega$ any set with some event $E \subset \Omega$ and $\mathcal{F} = \{\emptyset, E, \Omega \setminus E, \Omega\}$.

# Probability spaces

Given an experiment with outcomes $\Omega$ and an event space $\mathcal{F}$ we attach probabilities to events by defining a probability function $\mathbb{P} : \mathcal{F} \to \mathbb{R}$ as follows.

## Definition (Probability function)

1. $\mathbb{P}(E) \geq 0 \qquad \forall\, E \in \mathcal{F}$
2. $\mathbb{P}(\Omega) = 1$ and $\mathbb{P}(\emptyset) = 0$
3. $E_i \in \mathcal{F}$ for $i \in I$ disjoint (that is, $E_i \cap E_j = \emptyset$ for $i \neq j$) then

$$\mathbb{P}\left(\cup_{i \in I} E_i\right) = \sum_{i \in I} \mathbb{P}(E_i)\,.$$

We call the triple $(\Omega, \mathcal{F}, \mathbb{P})$ a probability space.

# Examples of probability spaces

1. $\Omega$ any set with event $E \subset \Omega$ ($E \neq \emptyset$, $E \neq \Omega$). Take $\mathcal{F} = \{\emptyset, E, \Omega \setminus E, \Omega\}$ as before and define the probability function by $\mathbb{P}(\emptyset) = 0$, $\mathbb{P}(\Omega) = 1$ and
$$\mathbb{P}(E) = p \quad , \quad \mathbb{P}(\Omega \setminus E) = 1 - p$$
for any $0 \leq p \leq 1$.
2. $\Omega = \{\omega_1, \omega_2, \ldots, \omega_n\}$ with $\mathcal{F} = \mathcal{P}(\Omega)$ and probabilities given for all $E \in \mathcal{F}$ by
$$\mathbb{P}(E) = \frac{|E|}{n} .$$
For a six-sided fair die $\Omega = \{1, 2, 3, 4, 5, 6\}$ and we take
$$\mathbb{P}(\{i\}) = \frac{1}{6} .$$
3. More generally, for each outcome $\omega_i \in \Omega$ assign a value $p_i$ where $p_i \geq 0$ and $\sum_{i=1}^{n} p_i = 1$. If $\mathcal{F} = \mathcal{P}(\Omega)$ then take
$$\mathbb{P}(E) = \sum_{i:\omega_i \in E} p_i \qquad \forall\, E \in \mathcal{F} .$$

# Conditional probabilities

Given a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and two events $E_1, E_2 \in \mathcal{F}$ how does knowledge that the random event $E_2$, say, has occurred influence the probability that $E_1$ has also ocurred?

## Definition (Conditional probability)

For $\mathbb{P}(E_2) > 0$, define the conditional probability, $\mathbb{P}(E_1|E_2)$, of $E_1$ given $E_2$ by

$$\mathbb{P}(E_1|E_2) = \frac{\mathbb{P}(E_1 \cap E_2)}{\mathbb{P}(E_2)} .$$

Note that $\mathbb{P}(E_2|E_2) = 1$. Moreover, for any $E' \in \mathcal{F}$ such that $\mathbb{P}(E') > 0$ then $(\Omega, \mathcal{F}, \mathbb{Q})$ is a probability space where $\mathbb{Q} : \mathcal{F} \to \mathbb{R}$ is defined by

$$\mathbb{Q}(E) = \mathbb{P}(E|E') \qquad \forall\, E \in \mathcal{F} .$$

# Independent events

Given a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ we can define independence between random events as follows.

## Definition (Independent events)

Two events, $E_1, E_2 \in \mathcal{F}$ are independent if

$$\mathbb{P}(E_1 \cap E_2) = \mathbb{P}(E_1)\mathbb{P}(E_2)$$

Otherwise, the events are dependent. Note that if $E_1$ and $E_2$ are independent events then

$$\begin{aligned}\mathbb{P}(E_1|E_2) &= \mathbb{P}(E_1)\\ \mathbb{P}(E_2|E_1) &= \mathbb{P}(E_2)\,.\end{aligned}$$

# Independence of multiple events

More generally, a collection of events $\{E_i : i \in I\}$ are independent events if for all subsets $J$ of $I$

$$\mathbb{P}(\cap_{j\in J} E_j) = \prod_{j\in J} \mathbb{P}(E_j)\,.$$

When this holds for $|J| = 2$ we have pairwise independence. This does not necessarily imply that the full independence holds.

## Example ($|I| = 3$ events)

$E_1, E_2, E_3$ are independent events if

$$\begin{aligned}
\mathbb{P}(E_1 \cap E_2) &= \mathbb{P}(E_1)\mathbb{P}(E_2) \\
\mathbb{P}(E_1 \cap E_3) &= \mathbb{P}(E_1)\mathbb{P}(E_3) \\
\mathbb{P}(E_2 \cap E_3) &= \mathbb{P}(E_2)\mathbb{P}(E_3) \\
\mathbb{P}(E_1 \cap E_2 \cap E_3) &= \mathbb{P}(E_1)\mathbb{P}(E_2)\mathbb{P}(E_3)
\end{aligned}$$

# Bayes' theorem

## Theorem (Bayes' theorem)

*If $E_1$ and $E_2$ are two events with $\mathbb{P}(E_1) > 0$ and $\mathbb{P}(E_2) > 0$ then*

$$\mathbb{P}(E_1|E_2) = \frac{\mathbb{P}(E_2|E_1)\mathbb{P}(E_1)}{\mathbb{P}(E_2)}\,.$$

## Proof.

We have that

$$\mathbb{P}(E_1|E_2)\mathbb{P}(E_2) = \mathbb{P}(E_1 \cap E_2) = \mathbb{P}(E_2 \cap E_1) = \mathbb{P}(E_2|E_1)\mathbb{P}(E_1)\,.$$

□

# Partitions

Given a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, we define a partition of $\Omega$ as follows.

## Definition (Partition)

A partition of $\Omega$ is a collection of disjoint events $\{E_i \in \mathcal{F} : i \in I\}$ with

$$\cup_{i \in I} E_i = \Omega\,.$$

We then have the following theorem.

## Theorem (Partition theorem)

*If $\{E_i \in \mathcal{F} : i \in I\}$ is a partition of $\Omega$ and $\mathbb{P}(E_i) > 0$ for all $i \in I$ then*

$$\mathbb{P}(E) = \sum_{i \in I} \mathbb{P}(E|E_i)\mathbb{P}(E_i) \qquad \forall\, E \in \mathcal{F}\,.$$

# Proof of partition theorem

We prove the partition theorem as follows.

Proof.

$$\begin{aligned}\mathbb{P}(E) &= \mathbb{P}\left(E \cap \left(\cup_{i\in I} E_i\right)\right)\\ &= \mathbb{P}\left(\cup_{i\in I}\left(E \cap E_i\right)\right)\\ &= \sum_{i\in I} \mathbb{P}\left(E \cap E_i\right)\\ &= \sum_{i\in I} \mathbb{P}(E|E_i)\mathbb{P}(E_i)\end{aligned}$$

□

# Bayes' theorem and partitions

A generalization of Bayes' theorem can be stated as follows combining Bayes' theorem with the partition theorem.

$$\mathbb{P}(E_i|E) = \frac{\mathbb{P}(E|E_i)\mathbb{P}(E_i)}{\sum_{j\in I}\mathbb{P}(E|E_j)\mathbb{P}(E_j)} \qquad \forall\, i \in I$$

where $\{E_i \in \mathcal{F} \,:\, i \in I\}$ forms a partition of $\Omega$.
As a special case consider the partition $\{E_1, \Omega \setminus E_1\}$ then we have

$$\mathbb{P}(E_1|E) = \frac{\mathbb{P}(E|E_1)\mathbb{P}(E_1)}{\mathbb{P}(E|E_1)\mathbb{P}(E_1) + \mathbb{P}(E|\Omega \setminus E_1)\mathbb{P}(\Omega \setminus E_1)}\,.$$

# Example of Bayes' theorem

Suppose that you have a good game of table football two times in three, otherwise a poor game. Your chance of scoring a goal is 3/4 in a good game and 1/4 in a poor game. What is your chance of scoring a goal in any given game? Conditional on having scored in a game, what is the chance that you had a good game?

So we know that $\mathbb{P}(\text{Good}) = 2/3$, $\mathbb{P}(\text{Poor}) = 1/3$, $\mathbb{P}(\text{Score}|\text{Good}) = 3/4$, $\mathbb{P}(\text{Score}|\text{Poor}) = 1/4$.

Thus, noting that $\{\text{Good}, \text{Poor}\}$ forms a partition,

$$\begin{aligned}\mathbb{P}(\text{Score}) &= \mathbb{P}(\text{Score}|\text{Good})\mathbb{P}(\text{Good}) + \mathbb{P}(\text{Score}|\text{Poor})\mathbb{P}(\text{Poor}) \\ &= (3/4) \times (2/3) + (1/4) \times (1/3) = 7/12\,.\end{aligned}$$

By Bayes' theorem

$$\mathbb{P}(\text{Good}|\text{Score}) = \frac{\mathbb{P}(\text{Score}|\text{Good})\mathbb{P}(\text{Good})}{\mathbb{P}(\text{Score})} = \frac{(3/4) \times (2/3)}{(7/12)} = 6/7.$$

# Random variables

Given a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ we may wish to work not with the outcomes $\omega \in \Omega$ directly but some real-valued function of them, say $X : \Omega \to \mathbb{R}$. This gives us the notion of a random variable (RV) measuring, for example, temperatures, profits, goals scored or minutes late.

We shall first consider the case of discrete random variables.

## Definition (Discrete random variable)

A function $X : \Omega \to \mathbb{R}$ is a discrete random variable on the probability space $(\Omega, \mathcal{F}, \mathbb{P})$ if

1. the image, $\text{Im}(X)$, is a countable subset of $\mathbb{R}$
2. $\{\omega \in \Omega : X(\omega) = x\} \in \mathcal{F} \qquad \forall\, x \in \mathbb{R}$

The first condition ensures discreteness of the values obtained. The second condition says that the set of outcomes, $\omega \in \Omega$, mapped to a common value by the function $X$ must be an event $E \in \mathcal{F}$ so that we can attach to it a probability $\mathbb{P}(E)$.

# Probability mass functions

Suppose that $X$ is a discrete RV. We shall write

$$\mathbb{P}(X = x) = \mathbb{P}(\{\omega \in \Omega \, : \, X(\omega) = x\}) \qquad \forall \, x \in \mathbb{R} \, .$$

So that

$$\sum_{x \in \mathrm{Im}(X)} \mathbb{P}(X = x) = \mathbb{P}(\cup_{x \in \mathrm{Im}(X)} \{\omega \in \Omega \, : \, X(\omega) = x\}) = \mathbb{P}(\Omega) = 1$$

and $\mathbb{P}(X = x) = 0$ if $x \notin \mathrm{Im}(X)$. It is usual to abbreviate this by writing

$$\sum_{x \in \mathbb{R}} \mathbb{P}(X = x) = 1 \, .$$

The RV $X$ is then said to have probability mass function $\mathbb{P}(X = x)$ thought of as a function of $x \in \mathbb{R}$. The probability mass function then describes the distribution of the RV $X$.

# Examples of discrete distributions

## Example (Bernoulli distribution)

Here $\operatorname{Im}(X) = \{0, 1\}$ and $\mathbb{P}(X = 1) = 1 - \mathbb{P}(X = 0) = p$ for $p \in [0, 1]$.

## Example (Binomial distribution, *Bin*$(n, p)$)

Here $\operatorname{Im}(X) = \{0, 1, \dots, n\}$ for some positive integer $n$ and $p \in [0, 1]$

$$\mathbb{P}(X = k) = \binom{n}{k} p^k (1 - p)^{n-k} \qquad \forall\, k = 0, 1, \dots, n\,.$$

## Example (Poisson distribution, Pois$(\lambda)$)

Here $\operatorname{Im}(X) = \{0, 1, \dots\}$ and $\lambda > 0$

$$\mathbb{P}(X = k) = \frac{\lambda^k e^{-\lambda}}{k!} \qquad \forall\, k = 0, 1, \dots\,.$$

## Example (Geometric distribution, Geo$(p)$)

Here $\operatorname{Im}(X) = \{1, 2, \dots\}$ and $0 < p \leq 1$

$$\mathbb{P}(X = k) = p(1 - p)^{k-1} \qquad \forall\, k = 1, 2, \dots\,.$$

# Expectation

One way to summarize the distribution of some RV, $X$, would be to construct a weighted average of the observed values, weighted by the probabilities of actually observing these values. This is the idea of expectation defined as follows.

## Definition (Expectation)

The expectation, $\mathbb{E}(X)$, of a discrete RV $X$ is defined as

$$\mathbb{E}(X) = \sum_{x \in \mathrm{Im}(X)} x\mathbb{P}(X = x)$$

so long as this sum is (absolutely) convergent (that is, $\sum_{x \in \mathrm{Im}(X)} |x\mathbb{P}(X = x)| < \infty$).

The expectation of a RV $X$ is also known as the expected value, the mean or simply the average.

# Expectations and transformations

Suppose that $X$ is a discrete RV and $g : \mathbb{R} \to \mathbb{R}$ is some transformation. We can check that $Y = g(X)$ is again a RV defined by $Y(\omega) = g(X)(\omega) = g(X(\omega))$.

## Theorem

*We have that*

$$\mathbb{E}(g(X)) = \sum_{x} g(x)\mathbb{P}(X = x)$$

*whenever the sum is absolutely convergent.*

## Proof.

$$\begin{aligned}\mathbb{E}(g(X)) = \mathbb{E}(Y) &= \sum_{y \in g(\mathrm{Im}(X))} y\mathbb{P}(Y = y) \\ &= \sum_{y \in g(\mathrm{Im}(X))} y \sum_{x \in \mathrm{Im}(X): g(x) = y} \mathbb{P}(X = x) \\ &= \sum_{x \in \mathrm{Im}(X)} g(x)\mathbb{P}(X = x)\end{aligned}$$

□

# Variance

For a discrete RV $X$ with expected value $\mathbb{E}(X)$ we can define the variance, written $\mathrm{Var}(X)$, as follows.

## Definition (Variance)

$$\mathrm{Var}(X) = \mathbb{E}\left((X - \mathbb{E}(X))^2\right)$$

Thus, writing $\mu = \mathbb{E}(X)$ and taking $g(x) = (x - \mu)^2$

$$\mathrm{Var}(X) = \mathbb{E}(g(X)) = \sum_x (x - \mu)^2 \mathbb{P}(X = x)\,.$$

Just as the expected value summarizes the location of outcomes taken by the RV $X$, the variance measures the dispersion of $X$ about its expected value.
The standard deviation of a RV $X$ is defined as $+\sqrt{\mathrm{Var}(X)}$.

# Calculating variances

Note that we can expand our expression for the variance as follows

$$\begin{aligned}\mathrm{Var}(X) &= \sum_x (x-\mu)^2 \mathbb{P}(X=x) \\ &= \sum_x (x^2 - 2\mu x + \mu^2)\mathbb{P}(X=x) \\ &= \sum_x x^2 \mathbb{P}(X=x) - 2\mu \sum_x x\mathbb{P}(X=x) + \mu^2 \sum_x \mathbb{P}(X=x) \\ &= \mathbb{E}(X^2) - 2\mu^2 + \mu^2 \\ &= \mathbb{E}(X^2) - \mu^2 \\ &= \mathbb{E}(X^2) - (\mathbb{E}(X))^2 \ .\end{aligned}$$

# Examples of means and variances

## Example (Bernoulli)

The expected value is given by

$$\begin{aligned}\mathbb{E}(X) &= \sum_x x\mathbb{P}(X = x) \\ &= 0 \times \mathbb{P}(X = 0) + 1 \times \mathbb{P}(X = 1) \\ &= 0 \times (1 - p) + 1 \times p = p\,.\end{aligned}$$

In order to calculate the variance we first calculate $\mathbb{E}(X^2)$

$$\begin{aligned}\mathbb{E}(X^2) &= \sum_x x^2\mathbb{P}(X = x) \\ &= 0^2 \times \mathbb{P}(X = 0) + 1^2 \times \mathbb{P}(X = 1) = p\,.\end{aligned}$$

Then the variance is given by

$$\mathrm{Var}(X) = \mathbb{E}(X^2) - (\mathbb{E}(X))^2 = p - p^2 = p(1 - p)\,.$$

# Bivariate random variables

Given a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, we may have two RVs, $X$ and $Y$, say. We can then use a joint probability mass function

$$\mathbb{P}(X = x, Y = y) = \mathbb{P}(\{\omega \in \Omega : X(\omega) = x\} \cap \{\omega \in \Omega : Y(\omega) = y\})$$

for all $x, y \in \mathbb{R}$.
If $g : \mathbb{R}^2 \to \mathbb{R}$ then we get a similar result to that obtained in the univariate case

$$\mathbb{E}(g(X, Y)) = \sum_{x \in \mathrm{Im}(X)} \sum_{y \in \mathrm{Im}(Y)} g(x, y)\mathbb{P}(X = x, Y = y)\,.$$

This idea can be extended to probability mass functions for the multivariate case with three or more RVs.

# Independence of random variables

We defined independence for events and can use the same idea for RVs $X$ and $Y$.

## Definition

Two RVs $X$ and $Y$ are independent if $\{\omega \in \Omega : X(\omega) = x\}$ and $\{\omega \in \Omega : Y(\omega) = y\}$ are independent for all $x, y \in \mathbb{R}$.

Thus, if $X$ and $Y$ are independent

$$\mathbb{P}(X = x, Y = y) = \mathbb{P}(X = x)\mathbb{P}(Y = y)\,.$$

If $X$ and $Y$ are independent discrete RV with expected values $\mathbb{E}(X)$ and $\mathbb{E}(Y)$ respectively then

$$\begin{aligned}
\mathbb{E}(XY) &= \sum_x \sum_y xy\mathbb{P}(X = x, Y = y) \\
&= \sum_x \sum_y xy\mathbb{P}(X = x)\mathbb{P}(Y = y) \\
&= \sum_x x\mathbb{P}(X = x) \sum_y y\mathbb{P}(Y = y) \\
&= \mathbb{E}(X)\mathbb{E}(Y)\,.
\end{aligned}$$

# Distribution functions

Given a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ we have so far considered discrete RVs that can take a countable number of values. More generally, we define $X : \Omega \to \mathbb{R}$ as a random variable if

$$\{\omega \in \Omega : X(\omega) \leq x\} \in \mathcal{F} \qquad \forall x \in \mathbb{R}\,.$$

Note that a discrete random variable, $X$, is a random variable since

$$\{\omega \in \Omega : X(\omega) \leq x\} = \cup_{y \in \mathrm{Im}(X): y \leq x} \{\omega \in \Omega : X(\omega) = y\} \in \mathcal{F}\,.$$

## Definition (Distribution function)

If $X$ is a RV then the distrbution function of $X$, written $F_X(x)$, is defined by

$$F_X(x) = \mathbb{P}(\{\omega \in \Omega : X(\omega) \leq x\}) = \mathbb{P}(X \leq x)\,.$$

# Properties of the distribution function

1. If $x \leq y$ then $F_X(x) \leq F_X(y)$.
2. If $x \to -\infty$ then $F_X(x) \to 0$.
3. If $x \to \infty$ then $F_X(x) \to 1$.
4. If $a < b$ then $\mathbb{P}(a < X \leq b) = F_X(b) - F_X(a)$.

# Continuous random variables

Random variables that take just a countable number of values are called discrete. More generally, we have that a RV can be defined by its distribution function, $F_X(x)$. A RV is said to be a continuous random variable when the distribution function has sufficient *smoothness* that

$$F_X(x) = \mathbb{P}(X \leq x) = \int_{-\infty}^{x} f_X(u)du$$

for some function $f_X(x)$. We can then take

$$f_X(x) = \begin{cases} \frac{dF_X(x)}{dx} & \text{if the derivative exists at } x \\ 0 & \text{otherwise} . \end{cases}$$

The function $f_X(x)$ is called the probability density function of the continuous RV $X$ or often just the density of $X$.

# Properties of the density function

1. $f_X(x) \geq 0$ for all $x \in \mathbb{R}$.
2. $$\int_{-\infty}^{\infty} f_X(x)dx = 1\,.$$
3. If $a \leq b$ then
$$\mathbb{P}(a \leq X \leq b) = \int_a^b f_X(x)dx\,.$$

# Examples of continuous random variables

We define these continuous RVs, $X$, by their density functions, $f_X(x)$.

## Example (Uniform distribution, $U(a, b)$)

If $a < b$ then

$$f_X(x) = \begin{cases} \frac{1}{(b-a)} & \text{if } a < x < b \\ 0 & \text{otherwise}\,. \end{cases}$$

## Example (Exponential distribution, Exp($\lambda$))

If $\lambda > 0$ then

$$f_X(x) = \begin{cases} \lambda e^{-\lambda x} & \text{if } x > 0 \\ 0 & \text{otherwise}\,. \end{cases}$$

## Example (Normal distribution, $N(\mu, \sigma^2)$)

$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-\mu)^2/(2\sigma^2)} \qquad -\infty < x < \infty$$

# Expectations of continuous random variables

Just as for discrete RVs we can define the expectation of a continuous RV with density function $f_X(x)$ by a weighted averaging.

## Definition (Expectation)

The expectation of $X$ is given by

$$\mathbb{E}(X) = \int_{-\infty}^{\infty} x f_X(x) dx$$

whenever the integral exists.

In a similar way to the discrete case we have that if $g : \mathbb{R} \to \mathbb{R}$ then

$$\mathbb{E}(g(X)) = \int_{-\infty}^{\infty} g(x) f_X(x) dx$$

whenever the integral exists.

# Variances of continuous random variables

Similarly, we can define the variance of a continuous RV, $X$, with density function $f_X(x)$ as follows.

## Definition (Variance)

$$\text{Var}(X) = \mathbb{E}\left((X - \mathbb{E}(X))^2\right) = \int_{-\infty}^{\infty} (x - \mu)^2 f_X(x) dx$$

whenever the integral exists and where $\mu = \mathbb{E}(X)$.

We again find that

$$\text{Var}(X) = \mathbb{E}(X^2) - (\mathbb{E}(X))^2 \ .$$

# Examples

Suppose that the RV $X$ has an exponential distribution with parameter $\lambda > 0$ then

$$\mathbb{E}(X) = \int_0^\infty x\lambda e^{-\lambda x} dx = \frac{1}{\lambda}$$

and

$$\mathbb{E}(X^2) = \int_0^\infty x^2\lambda e^{-\lambda x} dx = \frac{2}{\lambda^2}\,.$$

Hence,

$$\text{Var}(X) = \mathbb{E}(X^2) - (\mathbb{E}(X))^2 = \frac{2}{\lambda^2} - \frac{1}{\lambda^2} = \frac{1}{\lambda^2}\,.$$

# Bivariate continuous random variables

Given a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, we may have multiple continuous RVs, $X$ and $Y$, say.

## Definition (joint probability distribution function)

The joint probability distribution function is given by

$$\begin{aligned} F_{X,Y}(x,y) &= \mathbb{P}(\{\omega \in \Omega : X(\omega) \leq x\} \cap \{\omega \in \Omega : Y(\omega) \leq y\} \\ &= \mathbb{P}(X \leq x, Y \leq y) \end{aligned}$$

for all $x, y \in \mathbb{R}$.

Independence follows in a similar way to the discrete case and we say that two continuous RVs $X$ and $Y$ are independent if

$$F_{X,Y}(x,y) = F_X(x)F_Y(y)$$

for all $x, y \in \mathbb{R}$.

# Bivariate density functions

The bivariate density of two continuous RVs $X$ and $Y$ satifies

$$F_{X,Y}(x,y) = \int_{u=-\infty}^{x} \int_{v=-\infty}^{y} f_{X,Y}(u,v) du dv$$

and is given by

$$f_{X,Y}(x,y) = \begin{cases} \frac{\partial^2}{\partial x \partial y} F_{X,Y}(x,y) & \text{if the derivative exists at } (x,y) \\ 0 & \text{otherwise}\,. \end{cases}$$

We have that

$$f_{X,Y}(x,y) \geq 0 \qquad \forall\, x, y \in \mathbb{R}$$

and that

$$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f_{X,Y}(x,y) dx dy = 1\,.$$

# Marginal densities and independence

If $X$ and $Y$ have a joint density function $f_{X,Y}(x,y)$ then we have marginal densities

$$f_X(x) = \int_{v=-\infty}^{\infty} f_{X,Y}(x,v)dv$$

and

$$f_Y(y) = \int_{u=-\infty}^{\infty} f_{X,Y}(u,y)du\,.$$

In the case that $X$ and $Y$ are also independent then

$$f_{X,Y}(x,y) = f_X(x)f_Y(y)$$

for all $x, y \in \mathbb{R}$.

# Probability generating functions

# Probability generating functions

A very common situation is when a RV, $X$, can take only non-negative integer values, that is $\mathrm{Im}(X) \subset \{0, 1, 2, \ldots\}$. The probability mass function, $\mathbb{P}(X = k)$, is given by a sequence of values $p_0, p_1, p_2, \ldots$ where

$$p_k = \mathbb{P}(X = k) \qquad \forall\, k = 0, 1, 2, \ldots$$

and we have that

$$p_k \geq 0 \qquad \forall\, k = 0, 1, 2, \ldots \qquad \text{and} \qquad \sum_{k=0}^{\infty} p_k = 1\,.$$

The terms of this sequence can be wrapped together to define a function called the probability generating function (PGF).

## Definition (Probability generating function)

The probability generating function, $G_X(z)$, of a (non-negative integer-valued) RV $X$ is defined as

$$G_X(z) = \sum_{k=0}^{\infty} p_k z^k$$

for all values of $z$ such that the sum converges appropriately.

# Elementary properties of the PGF

1. $G_X(z) = \sum_{k=0}^{\infty} p_k z^k$ so
$$G_X(0) = p_0 \qquad \text{and} \qquad G_X(1) = 1\,.$$

2. If $g(t) = z^t$ then
$$G_X(z) = \sum_{k=0}^{\infty} p_k z^k = \sum_{k=0}^{\infty} g(k)\mathbb{P}(X = k) = \mathbb{E}(g(X)) = \mathbb{E}(z^X)\,.$$

3. The PGF is defined for all $|z| \le 1$ since
$$\sum_{k=0}^{\infty} |p_k z^k| \le \sum_{k=0}^{\infty} p_k = 1\,.$$

4. Importantly, the PGF characterizes the distribution of a RV in the sense that
$$G_X(z) = G_Y(z) \qquad \forall\, z$$
if and only if
$$\mathbb{P}(X = k) = \mathbb{P}(Y = k) \qquad \forall\, k = 0, 1, 2, \ldots\,.$$

# Examples of PGFs

## Example (Binomial distribution, *Bin*$(n, p)$)

$$G_X(z) = \sum_{k=0}^{n} \binom{n}{k} p^k (q)^{n-k} z^k = (q + pz)^n \qquad \text{where } q = 1 - p\,.$$

## Example (Poisson distribution, Pois$(\lambda)$)

$$G_X(z) = \sum_{k=0}^{\infty} \frac{\lambda^k e^{-\lambda}}{k!} z^k = e^{\lambda z} e^{-\lambda} = e^{\lambda(z-1)}\,.$$

## Example (Geometric distribution, Geo$(p)$)

$$G_X(z) = \sum_{k=1}^{\infty} p q^{k-1} z^k = pz \sum_{k=0}^{\infty} (qz)^k = \frac{pz}{1 - qz} \text{ if } |z| < q^{-1} \text{ and } q = 1 - p\,.$$

# Derivatives of the PGF

We can derive a very useful property of the PGF by considering the derivative, $G'_X(z)$, with respect to $z$ of the PGF $G_X(z)$. Assume we can interchange the order of differentiation and summation, so that

$$\begin{aligned} G'_X(z) &= \frac{d}{dz}\left(\sum_{k=0}^{\infty} z^k \mathbb{P}(X=k)\right) \\ &= \sum_{k=0}^{\infty} \frac{d}{dz}\left(z^k\right)\mathbb{P}(X=k) \\ &= \sum_{k=0}^{\infty} kz^{k-1}\mathbb{P}(X=k) \end{aligned}$$

then putting $z = 1$ we have that

$$G'_X(1) = \sum_{k=0}^{\infty} k\mathbb{P}(X=k) = \mathbb{E}(X)$$

the expectation of the RV $X$.

# Further derivatives of the PGF

Taking the second derivative gives

$$G_X''(z) = \sum_{k=0}^{\infty} k(k-1)z^{k-2}\mathbb{P}(X = k)\,.$$

So that,

$$G_X''(1) = \sum_{k=0}^{\infty} k(k-1)\mathbb{P}(X = k) = \mathbb{E}(X(X-1))$$

Generally, we have the following result.

## Theorem

*If the RV $X$ has PGF $G_X(z)$ then the $r$-th derivative of the PGF, written $G_X^{(r)}(z)$, evaluated at $z = 1$ is such that*

$$G_X^{(r)}(1) = \mathbb{E}\left(X(X-1)\cdots(X-r+1)\right)\,.$$

# Using the PGF to calculate the moments $\mathbb{E}(X)$ and $\mathrm{Var}(X)$

We have that

$$\mathbb{E}(X) = G'_X(1)$$

and

$$\begin{aligned}\mathrm{Var}(X) &= \mathbb{E}(X^2) - (\mathbb{E}(X))^2 \\ &= [\mathbb{E}(X(X-1)) + \mathbb{E}(X)] - (\mathbb{E}(X))^2 \\ &= G''_X(1) + G'_X(1) - G'_X(1)^2 \,.\end{aligned}$$

For example, if $X$ is a RV with the Pois($\lambda$) distribution
then $G_X(z) = e^{\lambda(z-1)}$.
Thus, $G'_X(z) = \lambda e^{\lambda(z-1)}$ and $G''_X(z) = \lambda^2 e^{\lambda(z-1)}$.
So, $G'_X(1) = \lambda$ and $G''_X(1) = \lambda^2$.
Finally,

$$\mathbb{E}(X) = \lambda \qquad \text{and} \qquad \mathrm{Var}(X) = \lambda^2 + \lambda - \lambda^2 = \lambda \,.$$

# Sums of independent random variables

The following theorem shows how PGFs can be used to find the PGF of the sum of independent RVs.

## Theorem

*If $X$ and $Y$ are independent RVs with PGFs $G_X(z)$ and $G_Y(z)$ then*

$$G_{X+Y}(z) = G_X(z)G_Y(z)\,.$$

## Proof.

Using the independence of $X$ and $Y$ we have that

$$\begin{aligned} G_{X+Y}(z) &= \mathbb{E}(z^{X+Y}) \\ &= \mathbb{E}(z^X z^Y) \\ &= \mathbb{E}(z^X)\mathbb{E}(z^Y) \\ &= G_X(z)G_Y(z) \end{aligned}$$

□

# Example

For example, suppose that $X$ and $Y$ are independent RVs with $\text{Pois}(\lambda_1)$ and $\text{Pois}(\lambda_2)$ distributions, respectively. Then

$$\begin{aligned} G_{X+Y}(z) &= G_X(z)G_Y(z) \\ &= e^{\lambda_1(z-1)}e^{\lambda_2(z-1)} \\ &= e^{(\lambda_1+\lambda_2)(z-1)} . \end{aligned}$$

Hence, $X + Y$ has a $\text{Pois}(\lambda_1 + \lambda_2)$ distribution.

# Elementary stochastic processes

# Random walks

Consider a sequence $Y_1, Y_2, \dots$ of idependent and identically distributed (IID) RVs with $\mathbb{P}(Y_i = 1) = p$ and $\mathbb{P}(Y_i = -1) = 1 - p$ with $p \in [0, 1]$.

## Definition (Simple random walk)

The simple random walk is a sequence of RVs $X_n$ for $n = 1, 2, \dots$ defined by

$$X_n = X_0 + Y_1 + Y_2 + \cdots + Y_n$$

where $X_0 \in \mathbb{R}$ is the starting value.

## Definition (Simple symmetric random walk)

A simple symmetric random walk is a simple random walk with the choice $p = 1/2$.

# Examples

Practical examples of random walks abound across the physical sciences (motion of atomic particles) and the non-physical sciences (epidemics, gambling, asset prices).

The following is a simple model for the operation of a casino. Suppose that a gambler enters with a capital of £$X_0$. At each stage the gambler places a stake of £1 and with probability $p$ wins the gamble otherwise the stake is lost. If the gambler wins the stake is returned together with an additional sum of £1.

Thus at each stage the gambler's capital increases by £1 with probability $p$ or decreases by £1 with probability $1 - p$.

The gambler's capital $X_n$ at stage $n$ thus follows a simple random walk.

In this case, the gambler is bankrupt if $X_n$ reaches £0 and can not continue further.

# Returning to the starting state for a simple random walk

Let $X_n$ be a simple random walk and

$$r_n = \mathbb{P}(X_n = X_0) \qquad \text{for } n = 1, 2, \ldots$$

the probability of returning to the starting state at time $n$.
We will show the following theorem.

## Theorem

*If $n$ is odd then $r_n = 0$ else if $n = 2m$ is even then*

$$r_{2m} = \binom{2m}{m} p^m (1-p)^m \,.$$

## Proof.

The position of the random walk will change by an amount

$$X_n - X_0 = Y_1 + Y_2 + \cdots + Y_n$$

between times 0 and $n$. Hence, for this change $X_n - X_0$ to be 0 there must be an equal number of up steps as down steps. This can never happen if $n$ is odd and so $r_n = 0$ in this case. If $n = 2m$ is even then note that the number of up steps in a total of $n$ steps is a binomial RV with parameters $2m$ and $p$. Thus,

$$r_{2m} = \mathbb{P}(X_n - X_0 = 0) = \binom{2m}{m} p^m (1-p)^m \,.$$

□

This result tells us about the probability of returning to the starting state at a given time $n$.

We will now look at the probability that we ever return to our starting state. For convenience, and without loss of generality, we shall take $X_0 = 0$ from now on.

# Recurrence and transience of simple random walks

Note first that $\mathbb{E}(Y_i) = p - (1-p) = 2p - 1$ for each $i = 1, 2, \ldots$. Thus there is a net drift upwards if $p > 1/2$ and a net drift downwards if $p < 1/2$. Only in the case $p = 1/2$ is there no net drift upwards or downwards.

We say that the simple random walk is recurrent if it is certain to revisit its starting state at some time in the future and transient otherwise.

We shall prove the following theorem.

### Theorem

*For a simple random walk with starting state $X_0 = 0$ the probability of revisiting the starting state is*

$$\mathbb{P}(X_n = 0 \textit{ for some } n = 1, 2, \ldots) = 1 - |2p - 1|\,.$$

Thus a simple random walk is recurrent only when $p = 1/2$.

# Proof

We have that $X_0 = 0$ and that the event $R_n = \{X_n = 0\}$ indicates that the simple random walk returns to its starting state at time $n$. Consider the event

$$F_n = \{X_n = 0, X_m \neq 0 \text{ for } m = 1, 2, \ldots, (n-1)\}$$

that the random walk first revisits its starting state at time $n$. If $R_n$ occurs then exactly one of $F_1, F_2, \ldots, F_n$ occurs. So,

$$\mathbb{P}(R_n) = \sum_{m=1}^{n} \mathbb{P}(R_n \cap F_m)$$

but

$$\mathbb{P}(R_n \cap F_m) = \mathbb{P}(F_m)\mathbb{P}(R_{n-m}) \qquad \text{for } m = 1, 2, \ldots, n$$

since we must first return at time $m$ and then return a time $n - m$ later which are independent events. So if we write $f_n = \mathbb{P}(F_n)$ and $r_n = \mathbb{P}(R_n)$ then

$$r_n = \sum_{m=1}^{n} f_m r_{n-m} \,.$$

We have already found an expression for $r_n$ and wish to solve these equations for the $f_m$.

# Proof, ctd

Define generating functions for the sequences $r_n$ and $f_n$ by

$$R(z) = \sum_{n=0}^{\infty} r_n z^n \qquad \text{and} \qquad F(z) = \sum_{n=0}^{\infty} f_n z^n$$

where $r_0 = 1$ and $f_0 = 0$ and take $|z| < 1$. We have that

$$\begin{aligned} \sum_{n=1}^{\infty} r_n z^n &= \sum_{n=1}^{\infty} \sum_{m=1}^{n} f_m r_{n-m} z^n \\ &= \sum_{m=1}^{\infty} \sum_{n=m}^{\infty} f_m z^m r_{n-m} z^{n-m} \\ &= \sum_{m=1}^{\infty} f_m z^m \sum_{k=0}^{\infty} r_k z^k \\ &= F(z) R(z)\,. \end{aligned}$$

The left hand side is $R(z) - r_0 z^0 = R(z) - 1$ thus we have that

$$R(z) = R(z)F(z) + 1 \qquad \text{if } |z| < 1.$$

# Proof, ctd

Now,

$$\begin{aligned} R(z) &= \sum_{n=0}^{\infty} r_n z^n \\ &= \sum_{m=0}^{\infty} r_{2m} z^{2m} \quad \text{as } r_n = 0 \text{ if } n \text{ is odd} \\ &= \sum_{m=0}^{\infty} \binom{2m}{m} (p(1-p)z^2)^m \\ &= (1 - 4p(1-p)z^2)^{-\frac{1}{2}} \,. \end{aligned}$$

The last step follows from the binomial series expansion of $(1-4\theta)^{-\frac{1}{2}}$ and the choice $\theta = p(1-p)z^2$.
Hence,

$$F(z) = 1 - (1 - 4p(1-p)z^2)^{\frac{1}{2}} \quad \text{for } |z| < 1 \,.$$

# Proof, ctd

But now

$$\begin{aligned}
\mathbb{P}(X_n = 0 \text{ for some } n = 1, 2, \ldots) &= \mathbb{P}(F_1 \cup F_2 \cup \cdots) \\
&= f_1 + f_2 + \cdots \\
&= \lim_{z \uparrow 1} \sum_{n=1}^{\infty} f_n z^n \\
&= F(1) \\
&= 1 - (1 - 4p(1-p))^{\frac{1}{2}} \\
&= 1 - ((p + (1-p))^2 - 4p(1-p))^{\frac{1}{2}} \\
&= 1 - ((2p-1)^2)^{\frac{1}{2}} \\
&= 1 - |2p - 1| .
\end{aligned}$$

So, finally, the simple random walk is certain to revisit its starting state just when $p = 1/2$.

# Mean return time

Consider the recurrent case when $p = 1/2$ and set

$$T = \min\{n \geq 1 \,:\, X_n = 0\} \qquad \text{so that} \qquad \mathbb{P}(T = n) = f_n$$

where $T$ is the time of the first return to the starting state. Then

$$\begin{aligned} \mathbb{E}(T) &= \sum_{n=1}^{\infty} n f_n \\ &= G'_T(1) \end{aligned}$$

where $G_T(z)$ is the PGF of the RV $T$ and for $p = 1/2$ we have that $4p(1-p) = 1$ so

$$G_T(z) = 1 - (1 - z^2)^{\frac{1}{2}}$$

so that

$$G'_T(z) = z(1 - z^2)^{-\frac{1}{2}} \to \infty \quad \text{as } z \uparrow 1\,.$$

Thus, the simple symmetric random walk ($p = 1/2$) is recurrent but the expected time to first return to the starting state is infinite.

# The Gambler’s ruin problem

We now consider a variant of the simple random walk. Consider two players A and B with a joint capital between them of £N. Suppose that initially A has £a ($0 \leq a \leq N$).
At each time step player B gives A £1 with probability $p$ and with probability $q = (1 - p)$ player A gives £1 to B instead. The outcomes at each time step are independent.
The game ends if either A or B reach £N.
We can think of this situation as a simple random walk on the states $\{0, 1, \ldots, N\}$ with absorbing barriers at 0 and $N$.
Define the probability of ruin for gambler A as

$$\rho_a = \mathbb{P}(\text{ruin}) = \mathbb{P}(\text{B wins}) \quad \text{for } 0 \leq a \leq N\,.$$

# Solution of the Gambler's ruin problem

## Theorem

*The probability of ruin when A starts with an initial capital of a is given by*

$$\rho_a = \begin{cases} \frac{\theta^a - \theta^N}{1-\theta^N} & \textit{if } p \neq q \\ 1 - \frac{a}{N} & \textit{if } p = q = 1/2 \end{cases}$$

*where* $\theta = q/p$.

# Proof

Consider what happens at the first time step

$$\begin{aligned}\rho_a &= \mathbb{P}(\text{ruin} \cap Y_1 = +1|S_0 = a) + \mathbb{P}(\text{ruin} \cap Y_1 = -1|S_0 = a)\\ &= p\mathbb{P}(\text{ruin}|S_0 = a+1) + q\mathbb{P}(\text{ruin}|S_0 = a-1)\\ &= p\rho_{a+1} + q\rho_{a-1}\end{aligned}$$

Now look for a solution to this difference equation of the form $\lambda^a$ with boundary conditions $\rho_0 = 1$ and $\rho_N = 0$.
Try a solution of the form $\rho_a = \lambda^a$ to give

$$\lambda^a = p\lambda^{a+1} + q\lambda^{a-1}$$

Hence,

$$p\lambda^2 - \lambda + q = 0$$

with solutions $\lambda = 1$ and $\lambda = q/p$.

# Proof, ctd

If $p \neq q$ there are two distinct solutions and the general solution of the difference equation is of the form $A + B(q/p)^a$.
Applying the boundary conditions

$$1 = \rho_0 = A + B \qquad \text{and} \qquad 0 = \rho_N = A + B(q/p)^N$$

we get

$$A = -B(q/p)^N$$

and

$$1 = B - B(q/p)^N$$

so

$$B = \frac{1}{1 - (q/p)^N)} \qquad \text{and} \qquad A = \frac{-(q/p)^N}{1 - (q/p)^N} \, .$$

Hence,

$$\rho_a = \frac{(q/p)^a - (q/p)^N}{1 - (q/p)^N} \, .$$

# Proof, ctd

If $p = q = 1/2$ then the general solution is $C + Da$.
So with the boundary conditions

$$1 = \rho_0 = C + D(0) \qquad \text{and} \qquad 0 = \rho_N = C + D(N)\,.$$

Therefore,

$$C = 1 \qquad \text{and} \qquad 0 = 1 + D(N)$$

so

$$D = -1/N$$

and

$$\rho_a = 1 - a/N\,.$$

# Mean time to ruin

Set $T_a$ as the time to be absorbed at either 0 or $N$ starting from the initial state $a$ and write $\mu_a = \mathbb{E}(T_a)$.
Then, conditioning on the first step as before

$$\mu_a = 1 + p\mu_{a+1} + q\mu_{a-1} \quad \text{for } 1 \le a \le N-1$$

and $\mu_0 = \mu_N = 0$.
It can be shown that $\mu_a$ is given by

$$\mu_a = \begin{cases} \frac{1}{p-q}\left(N\frac{(q/p)^a-1}{(q/p)^N-1} - a\right) & \text{if } p \neq q \\ a(N-a) & \text{if } p = q = 1/2\,. \end{cases}$$

# Notation

| | |
|---|---|
| RV | random variable |
| IID | independent, identically distributed |
| PGF | probability generating function $G_X(z)$ |
| $X \sim U(0,1)$ | RV X has the distribution $U(0,1)$, etc |
| $\mathbb{I}(A)$ | indicator function of the event $A$ |
| $\mathbb{P}(A)$ | probability that event $A$ occurs, e.g. $A = \{X = n\}$ |
| $\mathbb{E}(X)$ | expected value of RV $X$ |
| $\mathbb{E}(X^n)$ | $n^{th}$ moment of RV $X$, for $n = 1, 2, \ldots$ |
| $F_X(x)$ | distribution function, $F_X(x) = \mathbb{P}(X \leq x)$ |
| $f_X(x)$ | density of RV $X$ given, when it exists, by $F'_X(x)$ |

# Limits and inequalities

# Limits and inequalities

We are familiar with limits of real numbers. If $x_n = 1/n$ for $n = 1, 2, \ldots$ then $\lim_{n\to\infty} x_n = 0$ whereas if $x_n = (-1)^n$ no such limit exists.

Behaviour in the long-run or on average is an important characteristic of everyday life.

In this section we will be concerned with these notions of limiting behaviour when the real numbers $x_n$ are replaced by random variables $X_n$. As we shall see there are several distinct notions of convergence that can be considered.

To study these forms of convergence and the limiting theorems that emerge we shall on the way also gather a potent collection of concepts and tools for the probabilistic analysis of models and systems.

# Probabilistic inequalities

To help assess how close RVs are to each other it is useful to have methods that provide upper bounds on probabilities of the form

$$\mathbb{P}(X > a)$$

for fixed constants $a$, and where, for example, $X = |X_1 - X_2|$.
We shall consider several such bounds and related inequalities.

- Markov's inequality
- Chebychev's inequality
- Lyapunov's inequality

## Theorem (Markov's inequality)

*If* $\mathbb{E}(X) < \infty$ *then for any* $a > 0$*,*

$$\mathbb{P}(|X| \geq a) \leq \frac{\mathbb{E}(|X|)}{a} .$$

## Proof.

We have that

$$\mathbb{I}(|X| \geq a) = \begin{cases} 1 & |X| \geq a \\ 0 & \text{otherwise} . \end{cases}$$

Clearly,

$$|X| \geq a\mathbb{I}(|X| \geq a)$$

hence

$$\mathbb{E}(|X|) \geq \mathbb{E}(a\mathbb{I}(|X| \geq a)) = a\mathbb{P}(|X| \geq a)$$

which yields the result.

□

## Theorem (Chebychev's inequality)

*Let $X$ be a RV with mean $\mu$ and finite variance $\sigma^2$ then for all $a > 0$*

$$\mathbb{P}(|X - \mu| \geq a) \leq \frac{\sigma^2}{a^2}\,.$$

## Proof.

Consider, for example, the case of a continuous RV $X$ and put $Y = |X - \mu|$ then

$$\sigma^2 = \mathbb{E}(Y^2) = \int y^2 f_Y(y)dy = \int_{0 \leq y < a} y^2 f_Y(y)dy + \int_{y \geq a} y^2 f_Y(y)dy$$

so that

$$\sigma^2 \geq 0 + a^2\mathbb{P}(Y \geq a)\,.$$

□

## Theorem (Lyapunov's inequality)

*If* $r \geq s > 0$ *then* $\mathbb{E}(|X|^r)^{1/r} \geq \mathbb{E}(|X|^s)^{1/s}$*.*

## Proof.

Omitted. □

# Moment generating function

## Definition

The moment generating function (mgf) of a RV $X$ is given by

$$M_X(t) = \mathbb{E}(e^{tX})$$

and is defined for those values of $t$ for which this expectation exists.

Using the power series $e^x = 1 + x + x^2/2! + x^3/3! + \cdots$ we see that

$$M_X(t) = \mathbb{E}(e^{tX}) = 1 + \mathbb{E}(X)t + \mathbb{E}(X^2)t^2/2! + \mathbb{E}(X^3)t^3/3! + \cdots$$

and so the $n^{th}$ moment of $X$, $\mathbb{E}(X^n)$, is given by the coefficient of $t^n/n!$.

# Elementary properties of the mgf

1. If $X$ has mgf $M_X(t)$ then $Y = aX + b$ has mgf $M_Y(t) = e^{bt} M_X(at)$.
2. If $X$ and $Y$ are independent then $X + Y$ has mgf $M_{X+Y}(t) = M_X(t) M_Y(t)$.
3. $\mathbb{E}(X^n) = M_X^{(n)}(0)$ where $M_X^{(n)}$ is the $n^{th}$ derivative of $M_X$.
4. If $X$ is a discrete RV taking values $0, 1, 2, \ldots$ with probability generating function $G_X(z) = \mathbb{E}(z^X)$ then $M_X(t) = G_X(e^t)$.

# Fundamental properties of the mgf

1. Uniqueness: to each mgf there corresponds a unique distribution function having that mgf.
   In fact, if $X$ and $Y$ are RVs with the same mgf in some region $-a < t < a$ where $a > 0$ then $X$ and $Y$ have the same distribution.
2. Continuity: if distribution functions $F_n(x)$ converge pointwise to a distribution function $F(x)$, the corresponding mgf's (where they exist) converge to the mgf of $F(x)$. Conversely, if a sequence of mgf's $M_n(t)$ converge to $M(t)$ which is continuous at $t = 0$, then $M(t)$ is a mgf, and the corresponding distribution functions $F_n(x)$ converge to the distribution function determined by $M(t)$.

# Example: exponential distribution

If $X$ has an exponential distribution with parameter $\lambda > 0$ then $f_X(x) = \lambda e^{-\lambda x}$ for $0 < x < \infty$. Hence, for $t < \lambda$,

$$\begin{aligned} M_X(t) &= \int_0^\infty e^{tx} \lambda e^{-\lambda x} dx = \int_0^\infty \lambda e^{-(\lambda - t)x} dx \\ &= \left[ -\frac{\lambda}{(\lambda - t)} e^{-(\lambda - t)x} \right]_0^\infty = \frac{\lambda}{\lambda - t} \, . \end{aligned}$$

For $t < \lambda$

$$\frac{\lambda}{(\lambda - t)} = \left( 1 - \frac{t}{\lambda} \right)^{-1} = 1 + \frac{t}{\lambda} + \frac{t^2}{\lambda^2} + \cdots$$

and hence $\mathbb{E}(X) = 1/\lambda$ and $\mathbb{E}(X^2) = 2/\lambda^2$ so that

$$\operatorname{Var}(X) = \mathbb{E}(X^2) - (\mathbb{E}(X))^2 = 1/\lambda^2 \, .$$

# Example: normal distribution

Consider a normal RV $X \sim N(\mu, \sigma^2)$ then $f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2}$ so that

$$\begin{aligned} M_X(t) &= \int_{-\infty}^{\infty} e^{tx} \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2} dx \\ &= \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-(-2tx\sigma^2+(x-\mu)^2)/2\sigma^2} dx\,. \end{aligned}$$

So, by completing the square,

$$\begin{aligned} M_X(t) &= e^{\mu t+\sigma^2 t^2/2} \left\{ \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-(x-(\mu+t\sigma^2))^2/2\sigma^2} \right\} dx \\ &= e^{\mu t+\sigma^2 t^2/2}\,. \end{aligned}$$

# Example: uniform distribution

Consider a uniform RV $X \sim U(a, b)$. Then

$$f_X(x) = \begin{cases} \frac{1}{b-a} & a < x < b \\ 0 & \text{otherwise} \,. \end{cases}$$

Hence,

$$\begin{aligned} M_X(t) &= \int_a^b \frac{e^{tx}}{b-a} dx \\ &= \left[ \frac{e^{tx}}{(b-a)t} \right]_a^b \\ &= \frac{e^{bt} - e^{at}}{(b-a)t} \,. \end{aligned}$$

## Theorem (Chernoff's bound)

*Suppose that $X$ has mgf $M_X(t)$ and $a \in \mathbb{R}$ then for all $t \geq 0$*

$$\mathbb{P}(X \geq a) \leq e^{-ta} M_X(t) \, .$$

## Proof.

Using Markov's inequality, we have that

$$\begin{aligned} \mathbb{P}(X \geq a) &= \mathbb{P}(e^{tX} \geq e^{ta}) \\ &\leq \frac{\mathbb{E}(e^{tX})}{e^{ta}} \\ &= e^{-ta} M_X(t) \end{aligned}$$

□

Note that the above bound holds for all $t > 0$ so we can select the best such bound by choosing $t$ to minimize $e^{-ta} M_X(t)$.

# Notions of convergence: $X_n \to X$ as $n \to \infty$

For a sequence of RVs $(X_n)_{n\geq 1}$, we shall define several distinct notions of convergence to some RV $X$ as $n \to \infty$.

## Definition (Convergence in distribution)

$X_n \xrightarrow{D} X$ if $F_{X_n}(x) \to F_X(x)$ for all points $x$ at which $F_X$ is continuous.

## Definition (Convergence in probability)

$X_n \xrightarrow{P} X$ if $\mathbb{P}(|X_n - X| > \epsilon) \to 0$ for all $\epsilon > 0$.

## Definition (Convergence almost surely)

$X_n \xrightarrow{a.s.} X$ if $\mathbb{P}(X_n \to X) = 1$.

## Definition (Convergence in $r^{th}$ mean)

$X_n \xrightarrow{r} X$ if $\mathbb{E}(|X_n - X|^r) \to 0$.

# Convergence theorems

**Theorem**

*If* $X_n \xrightarrow{a.s.} X$ *then* $X_n \xrightarrow{P} X$.

**Theorem**

*If* $X_n \xrightarrow{P} X$ *then* $X_n \xrightarrow{D} X$.

**Theorem**

*If* $r > s \geq 1$ *and* $X_n \xrightarrow{r} X$ *then* $X_n \xrightarrow{s} X$.

**Theorem**

*If* $r \geq 1$ *and* $X_n \xrightarrow{r} X$ *then* $X_n \xrightarrow{P} X$.

## Theorem

*If* $X_n \xrightarrow{a.s.} X$ *then* $X_n \xrightarrow{P} X$.

## Proof.

Omitted. □

## Theorem

*If* $X_n \xrightarrow{P} X$ *then* $X_n \xrightarrow{D} X$*.*

## Proof

We prove this theorem as follows. Fix, $\epsilon > 0$ then

$$F_{X_n}(x) = \mathbb{P}(X_n \leq x \cap X > x + \epsilon) + \mathbb{P}(X_n \leq x \cap X \leq x + \epsilon)$$

since $X > x + \epsilon$ and $X \leq x + \epsilon$ form a partition. But if $X_n \leq x$
and $X > x + \epsilon$ then $|X_n - X| > \epsilon$
and $\{X_n \leq x \cap X \leq x + \epsilon\} \subset \{X \leq x + \epsilon\}$. Therefore,

$$F_{X_n}(x) \leq \mathbb{P}(|X_n - X| > \epsilon) + F_X(x + \epsilon)\,.$$

Similarly,

$$\begin{aligned} F_X(x - \epsilon) &= \mathbb{P}(X \leq x - \epsilon \cap X_n > x) + \mathbb{P}(X \leq x - \epsilon \cap X_n \leq x) \\ &\leq \mathbb{P}(|X_n - X| > \epsilon) + F_{X_n}(x)\,. \end{aligned}$$

The proof is completed by noting that together these inequalities show that

$$F_X(x-\epsilon) - \mathbb{P}(|X_n - X| > \epsilon) \leq F_{X_n}(x) \leq \mathbb{P}(|X_n - X| > \epsilon) + F_X(x+\epsilon)\,.$$

But $X_n \xrightarrow{P} X$ implies that $\mathbb{P}(|X_n - X| > \epsilon) \to 0$. So, as $n \to \infty$, $F_{X_n}(x)$ is squeezed between $F_X(x-\epsilon)$ and $F_X(x+\epsilon)$.

Hence, if $F_X$ is continuous at $x$, $F_{X_n}(x) \to F_X(x)$ and so $X_n \xrightarrow{D} X$. □

## Theorem

*If $r > s \geq 1$ and $X_n \xrightarrow{r} X$ then $X_n \xrightarrow{s} X$.*

## Proof.

Set $Y_n = |X_n - X| \geq 0$ then by Lyapunov's inequality

$$\mathbb{E}(Y_n^r)^{1/r} \geq \mathbb{E}(Y_n^s)^{1/s}.$$

Hence, if $\mathbb{E}(Y_n^r) \to 0$ then $\mathbb{E}(Y_n^s) \to 0$. □

## Theorem

*If $r \geq 1$ and $X_n \xrightarrow{r} X$ then $X_n \xrightarrow{P} X$.*

## Proof.

By Markov's inequality, for all $\epsilon > 0$

$$\mathbb{P}(|X_n - X| > \epsilon) \leq \frac{\mathbb{E}(|X_n - X|)}{\epsilon} .$$

But $X_n \xrightarrow{r} X$ implies $X_n \xrightarrow{1} X$ and so the right hand side tends to zero and as required $X_n \xrightarrow{P} X$. □

# Limit theorems

Given a sequence of RVs $(X_n)_{n\geq 1}$, let

$$S_n = X_1 + X_2 + \cdots + X_n \qquad \text{and} \qquad \overline{X}_n = S_n/n\,.$$

What happens to $\overline{X}_n$ for large $n$?

## Theorem (Weak Law of Large Numbers/WLLN)

*Suppose $(X_n)_{n\geq 1}$ are IID RVs with finite mean $\mu$ (and finite variance $\sigma^2$) then $\overline{X}_n \xrightarrow{P} \mu$.*

## Theorem (Strong Law of Large Numbers/SLLN)

*Suppose $(X_n)_{n\geq 1}$ are IID RVs with finite mean $\mu$ (and finite fourth moment) then $\overline{X}_n \xrightarrow{a.s.} \mu$.*

Note that convergence to $\mu$ in the WLLN and SLLN actually means convergence to a degenerate RV, $X$, with $\mathbb{P}(X = \mu) = 1$.

# WLLN

## Theorem (Weak Law of Large Numbers/WLLN)

*Suppose $(X_n)_{n\geq 1}$ are IID RVs with finite mean $\mu$ and finite variance $\sigma^2$ then $\overline{X}_n \xrightarrow{P} \mu$.*

## Proof.

Recall that $\mathbb{E}(\overline{X}_n) = \mu$ and $\mathrm{Var}(\overline{X}_n) = \sigma^2/n$. Hence, by Chebychev's inequality, for all $\epsilon > 0$

$$\mathbb{P}(|\overline{X}_n - \mu| > \epsilon) \leq \frac{\sigma^2/n}{\epsilon^2} = \frac{\sigma^2}{n\epsilon^2}$$

and so, letting $n \to \infty$,

$$\mathbb{P}(|\overline{X}_n - \mu| > \epsilon) \to 0$$

hence $\overline{X}_n \xrightarrow{P} \mu$ as required. □

# SLLN

### Theorem (Strong Law of Large Numbers/SLLN)

*Suppose* $(X_n)_{n\geq 1}$ *are IID RVs with finite mean* $\mu$ *(and finite fourth moment) then* $\overline{X}_n \xrightarrow{a.s.} \mu$.

### Proof.

Omitted. □

# Applications: estimating probabilities

Suppose we wish to estimate the probability, $p$, that we succeed when we play some game. For $i = 1, \ldots, n$, let

$$X_i = \mathbb{I}(\{i^{th}\text{game is success}\})\,.$$

So $\overline{X}_n = m/n$ if we succeed $m$ times in $n$ attempts.
We have that $\mu = \mathbb{E}(X_i) = \mathbb{P}(X_i = 1) = p$ so then

$$m/n \xrightarrow{a.s.} p$$

by the SLLN.
Thus we have shown the important result that the empirical estimate of the probability of some event by its observed sample frequency converges to the correct value as the number of samples grows.
This result forms the basis of all simulation methods.

# Applications: Shannon's entropy

## Theorem (Asymptotic Equipartition Property/AEP)

*If $X_n$ is a sequence of IID discrete RV with probability distribution given by $\mathbb{P}(X_i = x) = p(x)$ for each $x \in I$ then*

$$-\frac{1}{n}\log_2 p(X_1, X_2, \ldots, X_n) \xrightarrow{P} H(X)$$

*where Shannon's entropy is defined by*

$$H(X) = H(X_1) = \cdots = H(X_n) = -\sum_{x \in I} p(x)\log_2 p(x)$$

*and*

$$p(x_1, x_2, \ldots, x_n) = \prod_{i=1}^{n} p(x_i)$$

*is the joint probability distribution of the $n$ IID RVs $X_1, X_2, \ldots, X_n$.*

## Proof.

Observe that $p(X_i)$ is a RV taking the value $p(x)$ with probabilty $p(x)$ and similarly $p(X_1, X_2, \ldots, X_n)$ is a RV taking a value $p(x_1, x_2, \ldots, x_n)$ with probability $p(x_1, x_2, \ldots, x_n)$. Therefore,

$$\begin{aligned}
-\frac{1}{n}\log_2 p(X_1, X_2, \ldots, X_n) &= -\frac{1}{n}\log_2 \prod_{i=1}^{n} p(X_i) \\
&= -\frac{1}{n}\sum_{i=1}^{n} \log_2 p(X_i) \\
&= \frac{1}{n}\sum_{i=1}^{n} (-\log_2 p(X_i)) \\
&\xrightarrow{P} \mathbb{E}(-\log_2 p(X_i)) \qquad \text{by WLLN} \\
&= -\sum_{x \in I}^{n} p(x)\log_2 p(x) \\
&= H(X)
\end{aligned}$$

□

# AEP implications

By the AEP, for all $\epsilon > 0$,

$$\lim_{n\to\infty} \mathbb{P}(|-\frac{1}{n}\log_2 p(X_1, X_2, \ldots, X_n) - H(X)| \leq \epsilon) = 1$$

$$\lim_{n\to\infty} \mathbb{P}(H(X) - \epsilon \leq -\frac{1}{n}\log_2 p(X_1, X_2, \ldots, X_n) \leq H(X) + \epsilon) = 1$$

$$\lim_{n\to\infty} \mathbb{P}(-n(H(X) - \epsilon) \geq \log_2 p(X_1, X_2, \ldots, X_n) \geq -n(H(X) + \epsilon)) = 1$$

$$\lim_{n\to\infty} \mathbb{P}(2^{-n(H(X)+\epsilon)} \leq p(X_1, X_2, \ldots, X_n) \leq 2^{-n(H(X)-\epsilon)}) = 1$$

Thus, the sequences of outcomes $(x_1, x_2, \ldots, x_n)$ for which

$$2^{-n(H(X)+\epsilon)} \leq p(x_1, x_2, \ldots, x_n) \leq 2^{-n(H(X)-\epsilon)}$$

have a high probability and are refered to as typical sequences. An efficient (optimal) coding is to assign short codewords to such sequences leaving longer codewords for any non-typical sequence. Such long codewords must arise only rarely in the limit.

# Central limit theorem

## Theorem (Central limit theorem/CLT)

*Let $(X_n)_{n\geq 1}$ be a sequence of IID RVs with mean $\mu$, variance $\sigma^2$ and whose moment generating function converges in some interval $-a < t < a$ with $a > 0$. Then*

$$Z_n = \frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}} \xrightarrow{D} Z \sim N(0,1)\,.$$

# Proof of CLT

Set $Y_i = (X_i - \mu)/\sigma$ then $\mathbb{E}(Y_i) = 0$ and $\mathbb{E}(Y_i^2) = \mathrm{Var}(Y_i) = 1$ so

$$M_{Y_i}(t) = 1 + \frac{t^2}{2} + o(t^2)$$

where $o(t^2)$ refers to terms of higher order than $t^2$ which will therefore tend to 0 as $t \to 0$. Also,

$$Z_n = \frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}} = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} Y_i \,.$$

Hence,

$$\begin{aligned} M_{Z_n}(t) &= \left( M_{Y_i}\left( \frac{t}{\sqrt{n}} \right) \right)^n \\ &= \left( 1 + \frac{t^2}{2n} + o\left( \frac{t^2}{n} \right) \right)^n \\ &\to e^{t^2/2} \qquad \text{as} \qquad n \to \infty \,. \end{aligned}$$

But $e^{t^2/2}$ is the mgf of the $N(0,1)$ distribution so, together with the continuity property, the CLT now follows as required.

# CLT example

Suppose $X_1, X_2, \ldots, X_n$ are the IID RVs showing the $n$ sample outcomes of a 6-sided die with common distribution

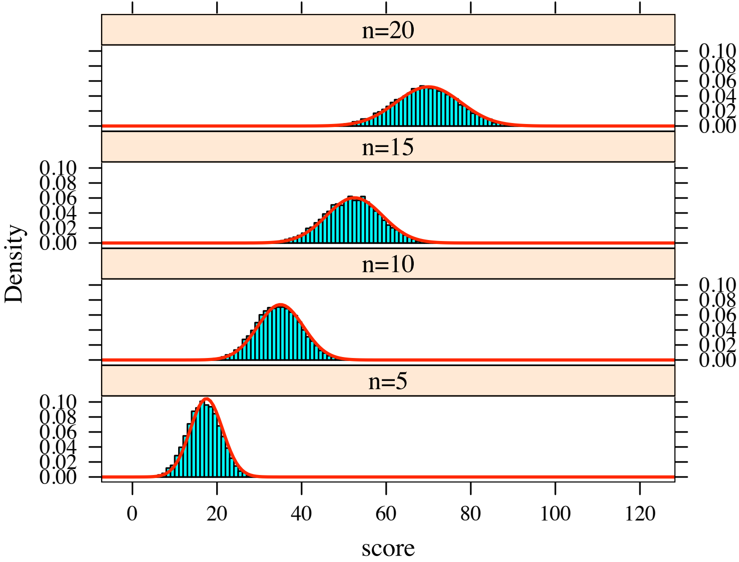
$$\mathbb{P}(X_i = j) = p_j\,, \qquad j = 1, 2, \ldots, 6$$

Set $S_n = X_1 + X_2 + \cdots + X_n$, the total score obtained, and consider the two cases

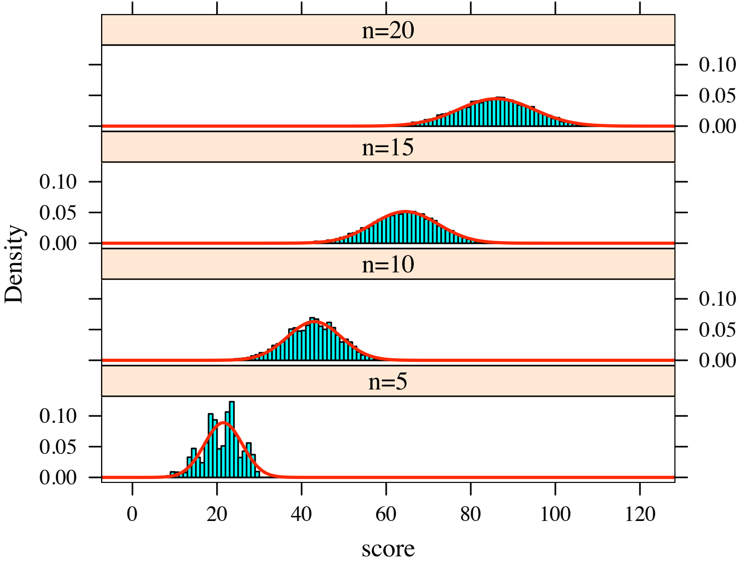
- symmetric: $(p_j) = (1/6, 1/6, 1/6, 1/6, 1/6, 1/6)$ so that $\mu = \mathbb{E}(X_i) = 3.5$ and $\sigma^2 = \mathrm{Var}(X_i) \approx 2.9$
- asymmetric: $(p_j) = (0.2, 0.1, 0.0, 0.0, 0.3, 0.4)$ so that $\mu = \mathbb{E}(X_i) = 4.3$ and $\sigma^2 = \mathrm{Var}(X_i) \approx 4.0$

for varying sample sizes $n = 5, 10, 15$ and 20.
The CLT tells us that for large $n$, $S_n$ is approximately distributed as $N(n\mu, n\sigma^2)$ where $\mu$ and $\sigma^2$ are the mean and variance, respectively, of $X_i$.

# CLT example: symmetric



# CLT example: asymmetric



# Confidence intervals I

One of the major statistical applications of the CLT is to the construction of confidence intervals. The CLT shows that

$$Z_n = \frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}}$$

is asymptotically distributed as $N(0, 1)$. If, the true value of $\sigma^2$ is unknown we may estimate it by the sample variance given by

$$S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \overline{X}_n)^2 \, .$$

For instance, it can be shown that $\mathbb{E}(S^2) = \sigma^2$ and then

$$\frac{\overline{X}_n - \mu}{S/\sqrt{n}}$$

is approximately distributed as $N(0, 1)$ for large $n$.

# Confidence intervals II

Define $z_\alpha$ so that $\mathbb{P}(Z > z_\alpha) = \alpha$ where $Z \sim N(0,1)$ and so

$$\mathbb{P}(-z_{\alpha/2} < Z < z_{\alpha/2}) = 1 - \alpha\,.$$

Hence,

$$\mathbb{P}\left(-z_{\alpha/2} < \frac{\overline{X}_n - \mu}{S/\sqrt{n}} < z_{\alpha/2}\right) \approx 1 - \alpha$$

$$\mathbb{P}\left(\overline{X}_n - z_{\alpha/2}\frac{S}{\sqrt{n}} < \mu < \overline{X}_n + z_{\alpha/2}\frac{S}{\sqrt{n}}\right) \approx 1 - \alpha\,.$$

The interval $\overline{X}_n \pm z_{\alpha/2} S/\sqrt{n}$ is thus an (approximate) $100(1-\alpha)$ percent confidence interval for the unknown parameter $\mu$.

# Confidence intervals: example

Consider a collection of $n$ IID RVs, $X_i$, with common distribution $X_i \sim \text{Pois}(\lambda)$. Hence,

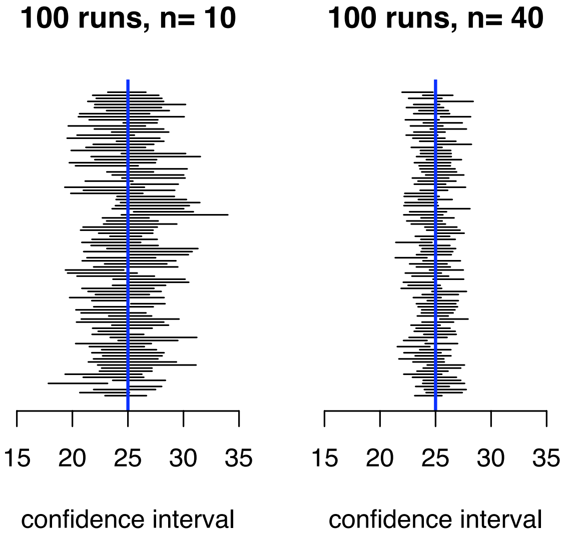
$$\mathbb{P}(X_i = j) = \frac{\lambda^j e^{-\lambda}}{j!} \qquad j = 0, 1, \ldots$$

with mean $\mathbb{E}(X_i) = \lambda$.
Then a 95% confidence interval for the mean value $\lambda$ is given by

$$\overline{X}_n \pm 1.96 S/\sqrt{n}$$

where $z_{0.025} = 1.96$.
Alternatively, to obtain 99% confidence intervals replace 1.96 by $z_{0.005} = 2.58$.

# Confidence intervals: illustration



# Monte Carlo simulation

Suppose we wish to estimate the value of $\pi$. One way to proceed is to perform the following experiment. Select a point $(X, Y) \in [-1, 1] \times [-1, 1]$, the square of side 2 and area 4 units, with $X$ and $Y$ chosen independently and uniformly in $[-1, 1]$. Now consider those points within unit distance of the origin then

$$\mathbb{P}((X, Y) \text{ lies in unit circle}) = \mathbb{P}(X^2 + Y^2 \leq 1) = \frac{\text{area of circle}}{\text{area of square}} = \frac{\pi}{4}\,.$$

Suppose we have access to a stream of random variables $U_i \sim U(0, 1)$ then $2U_i - 1 \sim U(-1, 1)$. Now set $X_i = 2U_{2i-1} - 1$, $Y_i = 2U_{2i} - 1$ and $H_i = \mathbb{I}(\{X_i^2 + Y_i^2 \leq 1\})$ so that

$$\mathbb{E}(H_i) = \mathbb{P}(X_i^2 + Y_i^2 \leq 1) = \frac{\pi}{4}\,.$$

Then by the SLLN the proportion of points $(X_i, Y_i)$ falling within the unit circle converges almost surely to $\pi/4$.

# Markov Chains

# Markov chains

## Definition (Markov chain)

Suppose that $(X_n)$ $n \geq 0$ is a sequence of discrete random variables taking values in some countable state space $S$. The sequence $(X_n)$ is a Markov chain if

$$\mathbb{P}(X_n = x_n | X_0 = x_0, X_1 = x_1, \ldots, X_{n-1} = x_{n-1}) = \mathbb{P}(X_n = x_n | X_{n-1} = x_{n-1})$$

for all $n \geq 1$ and for all $x_0, x_1, \ldots, x_n \in S$.

Since, $S$ is countable we can always choose to label the possible values of $X_n$ by integers and say that when $X_n = i$ the Markov chain is in the "$i^{th}$ state at the $n^{th}$ step" or "visits $i$ at time $n$".

# Transition probabilities

The dynamics of the Markov chain are governed by the transition probabilites $\mathbb{P}(X_n = j | X_{n-1} = i)$.

## Definition (time-homogeneous MC)

A Markov chain $(X_n)$ is time-homogeneous if

$$\mathbb{P}(X_n = j | X_{n-1} = i) = \mathbb{P}(X_1 = j | X_0 = i)$$

for all $n \geq 1$ and states $i, j \in S$.

- We shall assume that our MCs are time-homogeneous unless explicitly stated otherwise.

# Transition matrix

## Definition (Transition matrix)

The transition matrix, $P$, of a MC $(X_n)$ is given by $P = (p_{ij})$ where for all $i, j \in S$

$$p_{ij} = \mathbb{P}(X_n = j | X_{n-1} = i)\,.$$

- Note that $P$ is a stochastic matrix, that is, it has non-negative entries ($p_{ij} \geq 0$) and the row sums all equal one ($\sum_j p_{ij} = 1$).
- The transition matrix completely characterizes the dynamics of the MC.

# Example

Suppose the states of the MC are $S = \{1, 2, 3\}$ and that the transition matrix is given by

$$P = \begin{pmatrix} 1/3 & 1/3 & 1/3 \\ 1/2 & 0 & 1/2 \\ 2/3 & 0 & 1/3 \end{pmatrix} .$$

- Thus, in state 1 we are equally likely to be in any of the three states at the next step.
- In state 2, we can move with equal probabilities to 1 or 3 at the next step.
- Finally in state 3, we either move to state 1 with probability $2/3$ or remain in state 3 at the next step.

# *n*-step transition matrix

## Definition (*n*-step transition matrix)

The *n*-step transition matrix is $P^{(n)} = (p_{ij}^{(n)})$ where

$$p_{ij}^{(n)} = \mathbb{P}(X_n = j | X_0 = i)\,.$$

Thus $P^{(1)} = P$ and we also set $P^{(0)} = I$, the identity matrix.

# Chapman-Kolmogorov equations

## Theorem (Chapman-Kolmogorov)

*For all states $i, j$ and for all steps $m, n$*

$$p_{ij}^{(m+n)} = \sum_k p_{ik}^{(m)} p_{kj}^{(n)} .$$

*Hence, $P^{(m+n)} = P^{(m)} P^{(n)}$ and $P^{(n)} = P^n$, the $n^{th}$ power of $P$.*

## Proof.

$$\begin{aligned} p_{ij}^{(m+n)} &= \mathbb{P}(X_{m+n} = j | X_0 = i) = \sum_k \mathbb{P}(X_{m+n} = j, X_m = k | X_0 = i) \\ &= \sum_k \mathbb{P}(X_{m+n} = j | X_m = k, X_0 = i) \mathbb{P}(X_m = k | X_0 = i) \\ &= \sum_k \mathbb{P}(X_{m+n} = j | X_m = k) \mathbb{P}(X_m = k | X_0 = i) \\ &= \sum_k p_{kj}^{(n)} p_{ik}^{(m)} \end{aligned}$$

The Chapman-Kolmorgorov equations tell us how the long-term evolution of the MC depends on the short-term evolution specified by the transition matrix.
If we let $\lambda_i^{(n)} = \mathbb{P}(X_n = i)$ be the elements of a row vector $\lambda^{(n)}$ specifying the distribution of the MC at the $n^{th}$ time step then the follow holds.

## Lemma

$$\lambda^{(m+n)} = \lambda^{(m)} P^{(n)}$$

*and so,*

$$\lambda^{(n)} = \lambda^{(0)} P^{(n)}$$

*where $\lambda^{(0)}$ is the initial distribution $\lambda_i^{(0)} = \mathbb{P}(X_0 = i)$.*

## Proof.

$$\lambda_j^{(m+n)} = \mathbb{P}(X_{m+n} = j) = \sum_i \mathbb{P}(X_{m+n} = j | X_m = i)\mathbb{P}(X_m = i)$$

$$= \sum_i \lambda_i^{(m)} p_{ij}^{(n)} = \left(\lambda^{(m)} P^{(n)}\right)_j$$

# Classification of states

## Definition (Accessibility)

If, for some $n \geq 0$, $p_{ij}^{(n)} > 0$ then we say that state $j$ is accessible from state $i$, written $i \rightsquigarrow j$.
If $i \rightsquigarrow j$ and $j \rightsquigarrow i$ then we say that $i$ and $j$ communicate, written $i \leftrightsquigarrow j$.

Observe that the relation communicates $\leftrightsquigarrow$ is

- reflexive
- symmetric
- transitive

and hence is an equivalence relation. The corresponding equivalence classes partition the state space into subsets of states, called communicating classes, that communicate with each other.

# Irreducibility

- A communicating class, $C$, that once entered can not be left is called closed, that is $p_{ij} = 0$ for all $i \in C, j \notin C$.
- A closed communicating class consisting of a single state is called absorbing.
- When the state space forms a single communicating class, the MC is called irreducible and is called reducible otherwise.

# Recurrence and transience

Write for $n \geq 1$

$$f_{ij}^{(n)} = \mathbb{P}(X_1 \neq j, \ldots, X_{n-1} \neq j, X_n = j | X_0 = i)$$

so that $f_{ij}^{(n)}$ is the probability starting in state $i$ that we visit state $j$ for the first time at time $n$. Also, let

$$f_{ij} = \sum_{n \geq 1} f_{ij}^{(n)}$$

the probability that we ever visit state $j$, starting in state $i$.

## Definition

- If $f_{ii} < 1$ then state $i$ is transient
- If $f_{ii} = 1$ then state $i$ is recurrent.

# Recurrence and transience, ctd

- Observe that if we return to a state $i$ at some time $n$ then the evolution of the MC is independent of the path before time $n$. Hence, the probability that we will return at least $N$ times is $f_{ii}^N$.
- Now, if $i$ is recurrent $f_{ii}^N = 1$ for all $N$ and we are sure to return to state $i$ infinitely often.
- Conversely, if state $i$ is transient then $f_{ii}^N \to 0$ as $N \to \infty$ and so there is zero probability of returning infinitely often.

## Theorem

- *$i$ is transient* $\Leftrightarrow \sum_{n\geq 1} p_{ii}^{(n)}$ *converges*
- *$i$ is recurrent* $\Leftrightarrow \sum_{n\geq 1} p_{ii}^{(n)}$ *diverges*

*If $i$ and $j$ belong to the same communicating class then they are either both recurrent or both transient — the solidarity property.*

## Proof

First, define generating functions

$$P_{ii}(z) = \sum_{n=0}^{\infty} p_{ii}^{(n)} z^n \qquad \text{and} \qquad F_{ii}(z) = \sum_{n=0}^{\infty} f_{ii}^{(n)} z^n$$

where we take $p_{ii}^{(0)} = 1$ and $f_{ii}^{(0)} = 0$.

By examining the first time, $r$, that we return to $i$, we have for $m = 1, 2, \ldots$ that

$$p_{ii}^{(m)} = \sum_{r=1}^{m} f_{ii}^{(r)} p_{ii}^{(m-r)} \,.$$

Now multiply by $z^m$ and summing over $m$ we get

$$\begin{aligned} P_{ii}(z) &= 1 + \sum_{m=1}^{\infty} z^m p_{ii}^{(m)} \\ &= 1 + \sum_{m=1}^{\infty} z^m \sum_{r=1}^{m} f_{ii}^{(r)} p_{ii}^{(m-r)} \\ &= 1 + \sum_{r=1}^{\infty} f_{ii}^{(r)} z^r \sum_{m=r}^{\infty} p_{ii}^{(m-r)} z^{m-r} \\ &= 1 + F_{ii}(z) P_{ii}(z) \end{aligned}$$

Thus, $P_{ii}(z) = 1/(1 - F_{ii}(z))$. Now let $z \nearrow 1$ then $F_{ii}(z) \to F_{ii}(1) = f_{ii}$ and $P_{ii}(z) \to \sum_n p_{ii}^{(n)}$.

If $i$ is transient then $f_{ii} < 1$ so $\sum_n p_{ii}^{(n)}$ converges. Conversely, if $i$ is recurrent then $f_{ii} = 1$ and $\sum_n p_{ii}^{(n)}$ diverges.

Furthermore, if $i$ and $j$ are in the same class then there exist $m$ and $n$ so that $p_{ij}^{(m)} > 0$ and $p_{ji}^{(n)} > 0$. Now, for all $r \geq 0$

$$p_{ii}^{(m+r+n)} \geq p_{ij}^{(m)} p_{jj}^{(r)} p_{ji}^{(n)}$$

so that $\sum_r p_{jj}^{(r)}$ and $\sum_k p_{ii}^{(k)}$ diverge or converge together. □

# Mean recurrence time

First, let

$$T_j = \min\{n \geq 1 : X_n = j\}$$

be the time of the first visit to state $j$ and set $T_j = \infty$ if no such visit ever occurs.
Thus, $\mathbb{P}(T_i = \infty | X_0 = i) > 0$ if and only if $i$ is transient in which case $\mathbb{E}(T_i | X_0 = i) = \infty$.

## Definition (Mean recurrence time)

The mean recurrent time, $\mu_i$, of a state $i$ is defined as

$$\mu_i = \mathbb{E}(T_i | X_0 = i) = \begin{cases} \sum_n n f_{ii}^{(n)} & \text{if } i \text{ is recurrent} \\ \infty & \text{if } i \text{ is transient}\,. \end{cases}$$

- Note that $\mu_i$ may still be infinite when $i$ is recurrent.

# Positive and null recurrence

## Definition

A recurrent state *i* is

- positive recurrent if $\mu_i < \infty$ and
- null recurrent if $\mu_i = \infty$.

# Example: simple random walk

Recall the simple random walk where $X_n = \sum_{i=1}^{n} Y_i$ where $(Y_n)$ are IID RVs with $\mathbb{P}(Y_i = 1) = p = 1 - \mathbb{P}(Y_i = -1)$. Thus $X_n$ is the position after $n$ steps where we take unit steps up or down with probabilities $p$ and $1 - p$, respectively.

It is clear that return to the origin is only possible after an even number of steps. Thus the sequence $(p_{00}^{(n)})$ alternates between zero and a positive value.

# Periodicity

Let $d_i$ be the greatest common divisor of $\{n : p_{ii}^{(n)} > 0\}$.

## Definition

- If $d_i = 1$ then $i$ is aperiodic.
- If $d_i > 1$ then $i$ is periodic with period $d_i$.

- It may be shown that the period is a class property, that is, if $i, j \in C$ then $d_i = d_j$.

We will now concentrate on irreducible and aperiodic Markov chains.

# Stationary distributions

## Definition

The vector $\pi = (\pi_j; j \in S)$ is a stationary distribution for the MC with transition matrix $P$ if

1. $\pi_j \geq 0$ for all $j \in S$ and $\sum_{j \in S} \pi_j = 1$
2. $\pi = \pi P$, or equivalently, $\pi_j = \sum_{i \in S} \pi_i p_{ij}$.

Such a distribution is stationary in the sense that $\pi P^2 = (\pi P)P = \pi P = \pi$ and for all $n \geq 0$

$$\pi P^n = \pi \, .$$

Thus if $X_0$ has distribution $\pi$ then $X_n$ has distribution $\pi$ for all $n$. Moreover, $\pi$ is the limiting distribution of $X_n$ as $n \to \infty$.

# Markov's example

Markov was lead to the notion of a Markov chain by study the patterns of vowels and consonants in text. In his original example, he found a transition matrix for the states $\{\text{vowel}, \text{consonant})$ as

$$P = \begin{pmatrix} 0.128 & 0.872 \\ 0.663 & 0.337 \end{pmatrix} .$$

Taking successive powers of $P$ we find

$$P^2 = \begin{pmatrix} 0.595 & 0.405 \\ 0.308 & 0.692 \end{pmatrix} \qquad P^3 = \begin{pmatrix} 0.345 & 0.655 \\ 0.498 & 0.502 \end{pmatrix} \qquad P^4 = \begin{pmatrix} 0.478 & 0.522 \\ 0.397 & 0.603 \end{pmatrix} .$$

As $n \to \infty$,

$$P^n \to \begin{pmatrix} 0.432 & 0.568 \\ 0.432 & 0.568 \end{pmatrix} .$$

Check that $\pi = (0.432, 0.568)$ is a stationary distribution, that is $\pi P = \pi$.

# Limiting behaviour as $n \to \infty$

## Theorem (Erdös-Feller-Pollard)

*For all states $i$ and $j$ in an irreducible, aperiodic MC,*

1. *if the chain is transient, $p_{ij}^{(n)} \to 0$*
2. *if the chain is recurrent, $p_{ij}^{(n)} \to \pi_j$, where*
   2.1 *(null recurrent) either, every $\pi_j = 0$*
   2.2 *(positive recurrent) or, every $\pi_j > 0$, $\sum_j \pi_j = 1$ and $\pi$ is the unique probability distribution solving $\pi P = \pi$.*
3. *In case (2), let $T_i$ be the time to return to $i$ then $\mu_i = \mathbb{E}(T_i) = 1/\pi_i$ with $\mu_i = \infty$ if $\pi_i = 0$.*

## Proof.

Omitted. □

# Remarks

- The limiting distribution, $\pi$, is seen to be a stationary one. Suppose the current distribution is given by $\pi$ and consider the evolution of the MC for a further period of $T$ steps. Since $\pi$ is stationary, the probability of being in any state $i$ remains $\pi_i$, so we will make around $T\pi_i$ visits to $i$. Consequently, the mean time between visits to $i$ would be $T/(T\pi_i) = 1/\pi_i$.
- Using $\lambda_j^{(n)} = \mathbb{P}(X_n = j)$ and since $\lambda^{(n)} = \lambda^{(0)} P^n$
  1. for transient or null recurrent states $\lambda^{(n)} \to 0$, that is, $\mathbb{P}(X_n = j) \to 0$ for all states $j$
  2. for a positive recurrent state, $p^{(n)} \to \pi > 0$, that is, $\mathbb{P}(X_n = j) \to \pi_j > 0$ for all $j$, where $\pi$ is the unique probability vector solving $\pi P = \pi$.
- Note the distinction between a transient and a null recurrent chain is that in a transient chain we might never make a return visit to some state $i$ and there is zero probability that we will return infinitely often. However, in a null recurrent chain we are sure to make infinitely many return visits but the mean time between consecutive visits is infinite.

# Time-reversibility

Suppose now that $(X_n : -\infty < n < \infty)$ is an irreducible, postive recurrent MC with transition matrix $P$ and unique stationary distribution $\pi$. Suppose also that $X_n$ has the distribution $\pi$ for all $-\infty < n < \infty$. Now define the reversed chain by

$$Y_n = X_{-n} \qquad \text{for } -\infty < n < \infty$$

Then $(Y_n)$ is also a MC and where $Y_n$ has the distribution $\pi$.

## Definition (Reversibility)

A MC $(X_n)$ is reversible if the transition matrices of $(X_n)$ and $(Y_n)$ are equal.

## Theorem

*A MC $(X_n)$ is reversible if and only if*

$$\pi_i p_{ij} = \pi_j p_{ji} \qquad \textit{for all } i, j \in S\,.$$

## Proof.

Consider the transition probabilities $q_{ij}$ ofthe MC $(Y_n)$ then

$$\begin{aligned} q_{ij} &= \mathbb{P}(Y_{n+1} = j | Y_n = i) \\ &= \mathbb{P}(X_{-n-1} = j | X_{-n} = i) \\ &= \mathbb{P}(X_m = i | X_{m-1} = j)\mathbb{P}(X_{m-1} = j)/\mathbb{P}(X_m = i) \qquad \text{where } m = -n \\ &= p_{ji}\pi_j/\pi_i\,. \end{aligned}$$

Hence, $p_{ij} = q_{ij}$ if and only if $\pi_i p_{ij} = \pi_j p_{ji}$. □

## Theorem

*For an irreducible chain, if there exists a vector $\pi$ such that*

1. $0 \leq \pi_i \leq 1$ *and* $\sum_i \pi = 1$
2. $\pi_i p_{ij} = \pi_j p_{ji}$ *for all* $i, j \in S$

*then the chain is reversible and positive recurrent, with stationary distribution* $\pi$*.*

## Proof.

Suppose that $\pi$ satisfies the conditions of the theorem then

$$\sum_i \pi_i p_{ij} = \sum_i \pi_j p_{ji} = \pi_j \sum_i p_{ji} = \pi_j$$

and so $\pi = \pi P$ and the distribution is stationary. □

The conditions $\pi_i p_{ij} = \pi_j p_{ji}$ for all $i, j \in S$ are known as the local balance conditions.

# Ehrenfest model

Suppose we have two containers $A$ and $B$ containing a total of $m$ balls. At each time step a ball is chosen uniformly at random and switched between containers. Let $X_n$ be the number of balls in container $A$ after $n$ units of time. Thus, $(X_n)$ is a MC with transition matrix given by

$$p_{i,i+1} = 1 - \frac{i}{m}\,, \qquad p_{i,i-1} = \frac{i}{m}\,.$$

Instead of solving the equations $\pi = \pi P$ we look for solutions to

$$\pi_i p_{ij} = \pi_j p_{ji}$$

which yields $\pi_i = \binom{m}{i}(\frac{1}{2})^m$, a binomial distribution with parameters $m$ and $\frac{1}{2}$.

# Random walk on a graph

Consider a graph $G$ consisting of a countable collection of vertices $i \in N$ and a finite collection of edges $(i, j) \in E$ joining (unordered) pairs of vertices. Assume also that $G$ is connected. A natural way to construct a MC on $G$ uses a random walk through the vertices. Let $v_i$ be the number of edges incident at vertex $i$. The random walk then moves from vertex $i$ by selecting one of the $v_i$ edges with equal probability $1/v_i$. So the transition matrix, $P$, is

$$p_{ij} = \begin{cases} \frac{1}{v_i} & \text{if } (i, j) \text{ is an edge} \\ 0 & \text{otherwise}\,. \end{cases}$$

Since $G$ is connected, $P$ is irreducible. The local balance conditions for $(i,j) \in E$ are

$$\pi_i p_{ij} = \pi_j p_{ji}$$
$$\pi_i \frac{1}{v_i} = \pi_j \frac{1}{v_j}$$
$$\frac{\pi_i}{\pi_j} = \frac{v_i}{v_j} \,.$$

Hence,

$$\pi_i \propto v_i$$

and the normalization condition $\sum_{i \in N} \pi_i = 1$ gives

$$\pi_i = \frac{v_i}{\sum_{j \in N} v_j}$$

and $P$ is reversible.

# Ergodic results

Ergodic results tell us about the limiting behaviour of averages taken over time. In the case of Markov Chains we shall consider the long-run proportion of time spent in a given state.

Let $V_i(n)$ be the number of visits to $i$ before time $n$ then

$$V_i(n) = \sum_{k=0}^{n-1} \mathbb{I}(\{X_k = i\}) .$$

Thus, $V_i(n)/n$ is the proportion of time spent in state $i$ before time $n$.

## Theorem (Ergodic theorem)

*Let $(X_n)$ be a MC with irreducible transition matrix $P$ then*

$$\mathbb{P}\left( \frac{V_i(n)}{n} \to \frac{1}{\mu_i} \quad as \quad n \to \infty \right) = 1$$

*where $\mu_i = \mathbb{E}(T_i | X_0 = i)$ is the expected return time to state $i$.*

# Proof

If $P$ is transient then the total number of visits, $V_i$, to $i$ is finite with probability one, so

$$\frac{V_i(n)}{n} \leq \frac{V_i}{n} \to 0 = \frac{1}{\mu_i} \quad n \to \infty\,.$$

Alternatively, if $P$ is recurrent let $Y_i^{(r)}$ be the $r^{th}$ duration between visits to any given state $i$. Then $Y_i^{(1)}, Y_i^{(2)}, \ldots$ are non-negative IID RVs with $\mathbb{E}(Y_i^{(r)}) = \mu_i$.

But

$$Y_i^{(1)} + \cdots + Y_i^{(V_i(n)-1)} \leq n - 1$$

since the time of the last visit to $i$ before time $n$ occurs no later than time $n - 1$ and

$$Y_i^{(1)} + \cdots + Y_i^{(V_i(n))} \geq n$$

since the time of the first visit to $i$ after time $n - 1$ occurs no earlier than time $n$.

Hence,

$$\frac{Y_i^{(1)} + \cdots + Y_i^{(V_i(n)-1)}}{V_i(n)} \leq \frac{n}{V_i(n)} \leq \frac{Y_i^{(1)} + \cdots + Y_i^{(V_i(n))}}{V_i(n)} .$$

However, by the SLLN,

$$\mathbb{P}\left( \frac{Y_i^{(1)} + \cdots + Y_i^{(n)}}{n} \to \mu_i \quad \text{as} \quad n \to \infty \right) = 1$$

and for $P$ recurrent we know that $\mathbb{P}(V_i(n) \to \infty \quad \text{as} \quad n \to \infty) = 1$. So,

$$\mathbb{P}\left( \frac{n}{V_i(n)} \to \mu_i \quad \text{as} \quad n \to \infty \right) = 1$$

which implies

$$\mathbb{P}\left( \frac{V_i(n)}{n} \to \frac{1}{\mu_i} \quad \text{as} \quad n \to \infty \right) = 1 .$$

□

# Example: random surfing on web graphs

Consider a web graph, $G = (V, E)$, with vertices given by a finite collection of web pages $i \in V$ and (directed) edges given by $(i, j)$ whenever there is a hyperlink from page $i$ to page $j$.

Random walks through the web graph have received much attention in the last few years.

Consider the following model, let $X_n \in V$ be the location (that is, web page visited) by the surfer at time $n$ and suppose we choose $X_{n+1}$ uniformly from the, $L(i)$, outgoing links from $i$, in the case where $L(i) > 0$ and uniformly among all pages in $V$ if $L(i) = 0$ (the dangling page case).

Hence, the transition matrix, $\hat{P}_{ij}$, say, is given by

$$\hat{p}_{ij} = \begin{cases} \frac{1}{L(i)} & \text{if } (i,j) \in E \\ \frac{1}{|V|} & \text{if } L(i) = 0 \\ 0 & \text{otherwise} \end{cases}$$

where $|V|$ is the number of pages (that is, vertices) in the web graph. A potential problem remains in that $\hat{P}$ may not be irreducible or may be periodic.

We make a further adjustment to ensure irreducibility and aperiodicity as follows.
For $0 \leq \alpha \leq 1$ set

$$p_{ij} = (1 - \alpha)\hat{p}_{ij} + \alpha e e^T / |V|$$

where $e^T$ is a row vector of ones with length $|V|$.
We can interpret this as an “easily bored web surfer” model and see that the transitions take the form of a mixture of two distributions. With probability $\alpha$ we jump to a random page selected uniformly from the entire set of pages while with probability $1 - \alpha$ we follow the randomly chosen outgoing link (unless the page is dangling in which case we move to a randomly chosen page).

# PageRank

Brin *et al* (1999) used this approach to define PageRank through the limiting distribution of this Markov Chain, that is $\pi_i$ where the vector $\pi$ satisfies

$$\pi = \pi P$$

They report typical values for $\alpha$ of between 0.1 and 0.2.

The ergodic theorem now tells us that the random surfer in this model spends a proportion $\pi_i$ of the time visiting page *i* — a notion in some sense of the importance of page *i*.

Thus, two pages *i* and *j* can be ranked according to the total order defined by

$$i \geq j \quad \text{if and only if} \quad \pi_i \geq \pi_j \, .$$

See, “The PageRank Citation Ranking: Bring Order to the Web” Sergey Brin, Lawrence Page, Rajeev Motwani and Terry Winograd (1999) Technical Report, Computer Science Department, Stanford University.

`http://dbpubs.stanford.edu:8090/pub/1999-66`

# Computing PageRank: the power method

We seek a solution to the system of equations

$$\pi = \pi P$$

that is, we are looking for an eigenvector of $P$ (with corresponding eigenvalue of one). Google's computation of PageRank is one of the world's largest matrix computations.
The power method starts from an initial distribution $\pi^{(0)}$, updating $\pi^{(k-1)}$ by the iteration

$$\pi^{(k)} = \pi^{(k-1)} P = \cdots = \pi^{(0)} P^k$$

Advanced methods from linear algebra can be used to speed up convergence of the power method and there has been much study of related MCs, to include web browser back buttons and many other properties and alternative notions of the "importance" of a web page.

# Hidden Markov Models

An extension of Markov Chains is provided by Hidden Markov Models (HMM) where a statistical model of observed data is constructed from an underlying but usually hidden Markov Chain.
Such models have proved very popular in a wide variety of fields including

- speech and optical character recognition
- natural language processing
- bioinformatics and genomics.

We shall not consider these applications in any detail but simply introduce the basic ideas and questions that Hidden Markov Models address.

# A Markov model with hidden states

Suppose we have a MC with transition matrix $P$ but that the states $i$ of the chain are not directly observable. Instead, we suppose that on visiting any state $i$ at time $n$ there is a randomly chosen output value or token, $Y_n$, that is observable.

The probability of observing the output token $t$ when in state $i$ is given by some distribution $b_i$, depending on the state $i$ that is visited.

Thus,

$$\mathbb{P}(Y_n = t | X_n = i) = (b_i)_t$$

where $(b_i)_t$ is the $t^{th}$ component of the distribution $b_i$.

For an excellent introduction to HMM, see “A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition” Lawence R. Rabiner. Proceedings of the IEEE, Vol 77, No 2, February 1988.

# Three central questions

There are many variants of this basic setup but three central problems are usually addressed.

## Definition (Evaluation problem)

Given a sequence $y_1, y_2, \ldots, y_n$ of observed output tokens and the parameters of the HMM (namely, $P$, $b_i$ and the distribution for the initial state $X_0$) how do we compute

$$\mathbb{P}(Y_1 = y_1, Y_2 = y_2, \ldots, Y_n = y_n | \text{HMM parameters})$$

that is, the probability of the observed sequence given the model?

Such problems are solved in practice by the forward algorithm.

A second problem that may occur in an application is the decoding problem.

### Definition (Decoding problem)

Given an observed sequence of output tokens $y_1, y_2, \ldots, y_n$ and the full description of the HMM parameters, how do we find the best fitting corresponding sequence of (hidden) states $i_1, i_2, \ldots, i_n$ of the MC?

Such problems are solved in practice by a dynamic programming approach called the Viterbi algorithm.

The third important problem is the learning problem.

## Definition (Learning problem)

Given an observed sequence of output tokens $y_1, y_2, \ldots, y_n$, how do we adjust the parameters of the HMM to maximize

$$\mathbb{P}(Y_1 = y_1, Y_2 = y_2, \ldots, Y_n = y_n | \text{HMM parameters})$$

The observed sequence used to adjust the model parameters is called a training sequence. Learning problems are crucial in most applications since they allow us to create the “best” models in real observed processes.

Iterative procedures, known as the Baum-Welch method, are used to solve this problem in practice.

# Applications of Markov Chains

These and other applications of Markov Chains are important topics in a variety of Part II courses, including

- Artificial Intelligence II
- Bioinformatics
- Computer Systems Modelling