Notes in progress, October 19, 2020.

balthasar@rhizomeworks.com

# Contents

# 1 Sets and Measure Theory

**Contents of this chapter**

Sets are a term to describe collections of things. The things could be countable objects, such as the integers between $1$ and $10$, or contain a continuum, for example all the real numbers between $1$ and $10$. Unless otherwise specified, the elements of a set are assumed to be distinct and as not having an internal order. That is, the set of letters in "Mississippi" is $\{M, i, s, p\} = \{i, p, M, s\} = ...$ and so on. A concise resource for notation is Bradley (n/a).

## 1.1 Representation

| | |
|---|---|
| Statement Form | $\{\text{integers between 1 and 5}\}$ |
| Roster Form | $\{1, 2, 3, 4, 5\}$ |
| Setbuilder Form | $\{x \| x \in \mathbb{N}, 1 \leq x \leq 5\}$ |

## 1.2 Set Properties and Types of Sets

### 1.2.1 Cardinality

The cardinality of a set is a measure of the size of a set. For finite sets, the cardinality is simply the number of elements. For example, the set $A = \{a, b, c, d\}$ has size $|A| = 4$.

For infinite sets, this intuition breaks down, though it is still possible to make meaningful comparisons.

### 1.2.2 $\emptyset$ Empy Set, Null Set

The empty set is the set with no elements. It is denoted $\{\}$ or $\emptyset$.

### 1.2.3 Singleton Set

A singleton set is a set with one element. I.e., $\{a\}, \{b\}$ are singleton subsets of $\{a, b\}$. A singleton has cardinality 1.

### 1.2.4 Countable Sets

A set is countable if it contains finitely many elements or if its elements can be brought into a one-to-one correspondence with the set of integers. The set of all even integers is countable. The set of real numbers between $0$ and $1$ is not countable.

### 1.2.5 Infinite Sets

### 1.2.6 Multisets

A multiset is a set that may contain an element more than once. I.e. it makes sense to write $A = \{a, a, a, b, b\}$. The number of times an element is included is the *multiplicity*. The multiset $A$ may be defined in terms of a *multiplicity function* $m_A(x)$, which describes the multiplicity of a type of element $x \in U$ where $U$ may be referred to as the *universe* (en lieu of saying "univseral"!). The multiplicity function allows the extension of set characteristics and operations to multisets.

#### 1.2.6.1 Support

The support of a multiset is:

$$Supp(A) = \{x \in U | m_A(x) > 0\} \tag{1.1}$$

Which is the set of distinct elements in $A$. I.e., if $A = \{a, a, a, b, b, c\}$ then $Supp(A) = \{a, b, c\}$.

#### 1.2.6.2 Cardinality

The cardinality is given by:

$$|A| = \sum_{x \in U} m_A(x) \tag{1.2}$$

#### 1.2.6.3 $\subseteq$ Inclusion

The concept of subsets can be extended to multisets as:

$$A \subseteq B \tag{1.3}$$

if

$$\forall x \in U, \;\; m_A(x) \leq m_B(x) \tag{1.4}$$

#### 1.2.6.4 $\bigcap$ Intersection

The intersection of multisets is sometimes called the *infimum* or *greatest common divisor*. If $A \cap B = C$, then $C$ has multiplicity function:

$$m_C(x) = min\left(m_A(x), m_B(x)\right) \tag{1.5}$$

That is, it is like an elementwise minimum.

#### 1.2.6.5 $\bigcup$ Union

In the context of multisets, the term *union* sometimes refers to multiset addition. Otherwise, it should refer to the overlap of two multisets, i.e.:

$$m_C(x) = max\left(m_A(x), m_B(x)\right) \tag{1.6}$$

That is, it is like an elementwise maximum.

#### 1.2.6.6 $\uplus$ Multiset Addition

In contrast to sets in which distinct elements are only contained once, multiset addition makes sense and uses a special symbol "$\uplus$". If $A \uplus B = C$, then $C$ has multiplicity function:

$$m_C(x) = m_A(x) + m_B(x) \tag{1.7}$$

#### 1.2.6.7 Multiset Subtraction

Multisets can be subtracted, with the condition that the multiplicity can not be less than zero. If $A - B = C$, then $C$ has multiplicity function:

$$m_C(x) = max\left(m_A(x) - m_B(x), 0\right) \tag{1.8}$$

### 1.2.7 Powersets

The powerset $P(S)$ of a set $S$ is the set of all subsets of $S$, including the empty set and the set itself. That is:

$$P(S) = \{A : A \subseteq S\} \tag{1.9}$$

My guess is it's called powerset because the number of subsets of S, $|P(S)| = 2^{|S|}$. I have also seen the notation $2^S$ to refer to the space of subsets of $S$.

### 1.2.8 Measurable Sets

Measurable sets have a way of measuring volume on them in a non-trivial way. That is, for some subset of a measurable set $A_i \subseteq A$, it is possible to define a metric $\mu$ so that $\mu(A_i) \neq 0)$. Measurable sets are the elements of a $\sigma$-algebra.

### 1.2.9 Universal Set

The *universal set* is understood to refer to something that is not allowed in the context of Russell's Paradox. It is nevertheless useful to define a set $S$ so that $S^c = \emptyset$ and I've seen this type of set referred to as *universal set* a couple of times. In the context of probability theory, $S$ may be the whole event space.

### 1.2.10 Open Sets, Closed Sets

Open sets and closed sets are generalizations of open and closed intervals on the real line. Open sets a mentioned plenty in the context of measure theory. A set is open if and only if its complement is closed, and a set is closed if and only if its complement is open. Counterintuitively, a set can be both open and closed at the same time, or it can be neither open nor closed at the same time.

#### 1.2.10.1 Interior Points

An interior point of a set $A$ is any point $X$ so that for some $\epsilon > 0$, the open interval $X - \epsilon, X + \epsilon \subseteq A$ is contained within $A$. One might imagine an interior point as a point that has a neighborhood (a "ball") that is included in $A$.

#### 1.2.10.2 Accumulation Points

In contrast, accumulation points do not have a neighborhood contained in $A$. Formally, for some $\epsilon > 0$, $(X - \epsilon, X + \epsilon) \cap (A \setminus \{X\}) \neq \emptyset$, which is only possible for a point that is on the very boundary of $A$.

#### 1.2.10.3 Open Sets

Unsurprisingly, an open set is a set that only has interior points, i.e. $\{X : X \in A, \exists \epsilon > 0 \text{ s. th. } X - \epsilon, X + \epsilon \subseteq A\}$.

#### 1.2.10.4 Closed Sets

Closed sets contain all of their accumulation points. That is, the boundary is included. Points might also be standing alone, for example:

$$[2, 4] \cup \{1\} \tag{1.10}$$

Is a closed set.

#### 1.2.10.5 Both Open and Closed, Neither Open Nor Closed

The set of real numbers $\mathbb{R}$ is both open and closed. Any point on the real line has a neighborhood in $\mathbb{R}$, so it is open. On the other hand, the accumulation points of $\mathbb{R}$ are $\mathbb{R}$, so it is closed. An interval that has some but not all accumulation points is neither open nor closed. For example $[1, 2)$.

### 1.2.11 Image

The image are the elements of the codomain that a given subset of the domain is mapped to. For example, given a function $f : \mathbb{R} \to \mathbb{R} : f(x) = x^2$, the image of $\{-3, -2, 2, 3\}$ in the domain are the elements $\{4, 9\}$ of the codomain.

### 1.2.12 Preimage, Inverse Image

The preimage are elements of the domain that are mapped to a given subset of the codomain. For example, given a function $f : \mathbb{R} \to \mathbb{R} : f(x) = x^2$, the preimage of the elements $\{4, 9\}$ of the codomain are the elements $\{-3, -2, 2, 3\}$ of the domain..

### 1.2.13 Convex Sets

Convex sets are subsets of vector spaces in which any point along the line connecting to points within the set is contained within the set. A disk in $\mathbb{R}^2$ is a convex set. A crescent in $\mathbb{R}^2$ is not a convex set. The boundary of a convex set is a convex function. For example, a parabola $f(x) = x^2$ with $x \in \mathbb{R}$ is a convex function, and the area above the parabola, called the epigraph, is a convex subset of $\mathbb{R}$. Convex optimization deals with the optimization of convex functions over convex sets. The intersection of convex sets is always convex, but the union of convex sets is only convex under certain conditions.

More generally, if $S$ is a convex set, then affine combinations of the elements $\mathbf{x}_i \in S$ of the form:

$$\sum_i \mathbf{x}_i \lambda_i \tag{1.11}$$

With $\sum_i \lambda_i = 1$ and $\lambda_i \geq 0$ are also contained in $S$. Rather than just the points along the line between two points, these are the essentially the weighted averages of multiple points in the set.

#### 1.2.13.1 Example: Discrete Probability Distributions

The set of discrete probability distributions $P = \{\mathbf{p} = (p_1, p_2, ...) : ||\mathbf{p}||_1 = 1, p_i \geq 0\}$ is a convex set. Property 1.11 means that a weighted average over members of $P$ is also a member of $P$, so long as the weights satisfy $\sum_i \lambda_i = 1$. In general, this guarantees that marginalization results in a probability. I.e. if $p(X|Y) = \mathbf{p}(X|Y) \in P$ and $p(Y) = \mathbf{p}(Y) \in P$, then $\mathbf{p}(X) = \sum_i p_i(X|Y)p_i(Y) \in P$.

### 1.2.14 Choice Sets, Transversal Sets, Cross-Sections

Choice sets, transversal sets or cross-sections are sets that are assembled by picking exactly one element from each member of a family of disjoint sets. The axiom of choice (cf. section 1.7) states that such a set can always be formed.

#### 1.2.14.1 Example: Integers

Consider the partition of the integers between $1$ and $40$ into a family of disjoint sets that each contain $10$ elements:

$$S = \{\{1, ..., 10\}, \{11, ..., 20\}, \{21, ..., 30\}, \{31, ..., 40\}\} \tag{1.12}$$

Then a transversal set $A$ could for example be formed by picking the smallest number in each of the subsets.

$$A = \{1, 11, 21, 31\} \tag{1.13}$$

Though we might have picked any other choice function.

## 1.3 Set Operations

### 1.3.1 $A^c$ Complement

Given a subset $A \subseteq X$, the complement of the subset $A^c$ refers to everything that is not contained in $A$, i.e. $A^c = X \setminus A$.

### 1.3.2 $\bigcup$ Union

The union of sets is the set that contains all of their elements, counting all elements only once. The cardinality of the union is calculated using the important inclusion exclusion principle (1.6). In terms of logic, think $A \cup B$ is "A or B (or both)".

### 1.3.3 $\bigcap$ Intersection

The intersection of sets is the elements shared by all sets. $A \cap B$ is "A and B".

### 1.3.4 $\bigsqcup$ Disjoint Union, Discriminated Union

The disjoint union, or discriminated union, of sets is the union formed in a way in which the information about which subset an element belonged to is preserved. One way to write this is to include a subset index with each element, i.e.:

$$\bigsqcup_{i \in \{i\}} A_i = \bigcup_{i \in \{i\}} \{(x, i) : x \in A_i\} \tag{1.14}$$

For example, for the two sets $A = \{1, 2, 3, 4\}$ and $B = \{1, 2, 3\}$, the disjoint union is:

$$A \sqcup B = \{(1, A), (2, A), (3, A), (4, A), (1, B), (2, B), (3, B)\} \tag{1.15}$$

The cardinality of the disjoint union is simply the sum:

$$\left| \bigsqcup_{i \in \{i\}} A_i \right| = \sum_{i \in \{i\}} |A_i \tag{1.16}$$

I have also seen the disjoint union be used in the context of forming the union of disjoint sets, perhaps to stress that the sets are disjoint and the cardinality can be calculated through simple summation as above. For example, in dividing up the interval: $[0, 2] = [0, 1) \sqcup [1, 2]$.

## 1.4 Bijection Principle

The bijection principle states that when a bijection exists between the elements of two sets, then those sets have the same size. This principle is very useful in combinatorics, and it allows for meaningful comparisons of infinite sets.

## 1.5 DeMorgan's Rules

DeMorgan's Rules relate the complement of the union to the intersection of the complements, and the complement of the intersection to the union of the complements.

$$\left( \bigcup_{i \in \{i\}} A_i \right)^c = \bigcap_{i \in \{i\}} A_i^c \tag{1.17}$$

$$\left( \bigcap_{i \in \{i\}} A_i \right)^c = \bigcup_{i \in \{i\}} A_i^c \tag{1.18}$$

## 1.6 Inclusion - Exclusion Principle

The inclusion-exclusion principle is used to calculate the size of the union of sets. This requires counting each region of some complicated overlapping Venn diagram exactly once, which, in turn requires accounting for overcounting wherever sets overlap. Let $\{A_i | i \in \{i\}_n\}$ be a collection of $n$ overlapping sets indexed by $i \in \{i\}_n$, then the inclusion-exclusion principle is given by:

$$\left|\bigcup_{i\in\{i\}_n} A_i\right| = \sum_{k=1}^{n}(-1)^{k-1} \sum_{\{j\}_k\subseteq\{i\}_n} \left|\bigcap_{j\in\{j\}_k} A_j\right| \tag{1.19}$$

Where the sum over $\{j\}_k \subseteq \{i\}_n$ is over all $k$-element subsets of $\{i\}_n$.

### 1.6.1 Example: n=2 Sets and n=3 Sets

**n=2**

$$\begin{aligned} \left|\bigcup_{i\in\{1,2\}} A_i\right| &= \sum_{k=1}^{2}(-1)^{k-1} \sum_{\{j\}_k\subseteq\{1,2\}} \left|\bigcap_{j\in\{j\}_k} A_j\right| \\ &= (-1)^0\left(|A_1| + |A_2|\right) \\ &\quad +(-1)^1\left(|A_1 \cap A_2|\right) \end{aligned} \tag{1.20}$$

**n=3**

$$\begin{aligned} \left|\bigcup_{i\in\{1,2,3\}} A_i\right| &= \sum_{k=1}^{3}(-1)^{k-1} \sum_{\{j\}_k\subseteq\{1,2,3\}} \left|\bigcap_{j\in\{j\}_k} A_j\right| \\ &= (-1)^0\left(|A_1| + |A_2| + |A_3|\right) \\ &\quad +(-1)^1\left(|A_1 \cap A_2| + |A_1 \cap A_3| + |A_2 \cap A_3|\right) \\ &\quad +(-1)^2\left(|A_1 \cap A_2 \cap A_3|\right) \end{aligned} \tag{1.21}$$

#### 1.6.1.1 Example: Counting Integers

How many integers are there between 1 and 100 that are neither divisible by 3,5 nor 7?

Let $S$ be the set of all integers between 1 and 100. The size of the set is $|S| = 100$. The subset of $S$ that is numbers divisible by 3 is $A_3 \subseteq S$ with $|A_3| = 33$ because $100/3 = 33.\overline{333}$. Similarly, $|A_5| = 20$ and $|A_7| = 14$. The set of integers that is not divisible by 3, 5 or 7 is:

$$S \setminus \bigcup_{i\in\{3,5,7\}} A_i \tag{1.22}$$

So that the sought after quantity is :

$$\begin{aligned} \left|S \setminus \bigcup_{i\in\{3,5,7\}} A_i\right| &= |S| - [|A_3| + |A_5| + |A_7| \\ &\quad -|A_3 \cap A_5| - |A_3 \cap A_7| - |A_5 \cap A_7| + |A_3 \cap A_5 \cap A_7|] \end{aligned} \tag{1.23}$$

The size of the intersection $|A_3 \cap A_5| = 6$ because 100 is 6 times divisible by $3 \times 5 = 15$. Similarly, $|A_3 \cap A_7| = 4$, $|A_5 \cap A_7| = 2$ and $|A_3 \cap A_5 \cap A_7| = 0|$. Hence:

$$\left|S \setminus \bigcup_{i\in\{3,5,7\}} A_i\right| = 100 - 33 - 20 - 14 + 6 + 4 + 2 - 0 = 45 \tag{1.24}$$

There are 45 integers between 1 and 100 that are not divisible by 3, 5 or 7.

## 1.7 Axiom of Choice

The axiom of choice simply states that, given a collection of nonempty, mutually disjoint sets, it is possible to assemble a *transversal* or *choice* set that consists of exactly one element from each of the sets in the collection. For example, consider the students in 1st, 2nd, 3rd.. etc grades at a school to be a collection of nonempty, mutually disjoint sets. Then the axiom of choice says that it is possible to assemble a subset of students with exactly one student from each grade.

In terms of functions, one might think of defining a *choice function* on the family of mutually disjoint sets, which selects members from the collection of sets and adds them to the *choice* set. According to the axiom of choice, a choice function can be defined for any collection of nonempty, mutually disjoint subsets. This implies that all surjective functions have a right inverse. That is, for any surjective function $f : X \to Y$, there exists a function $g : Y \to X$ so that $f(g(y)) = y$.

The axiom of choice turns out to be associated with famous names and deep consequences (Bell 2015).

### 1.7.1 Example: Pairs of Real Numbers, Right Inverse

The collection of rank-2 sets of pairs of real numbers $A = \{A_x = \{x, -x\} : x \in \mathbb{R}^+\}$ are a collection of mutually disjoint subsets. A transversal set $B$ may be assembled by choosing the largest element of the tuple: $B = \{y : ymax(A_x), A_x \in A\}$. An equivalent surjection $f : A \to B$ is $f(A_x) = x$. There is a right-inverse $g : B \to A$ which is $g(x) = (x, -x)$ so that $f(g(x)) = x$.

## 1.8 $\sigma$-Algebras, $\sigma$-Fields

### 1.8.1 Definition

Given a set $X$, a $\sigma$-Algebra $\mathscr{A}$ is a collection of subsets of a given set $X$, which has to satisfy the conditions:

- $\emptyset, X \in \mathscr{A}$
- If $A \in \mathscr{A}$, then $A^c := X \setminus A \in \mathscr{A}$
- $\mathscr{A}$ has (possibly infinitely many) countable subsets $A_i \in \mathscr{A},\ i \in \mathbb{N}$. Then $\bigcup_{i=1}^{\infty} A_i \in \mathscr{A}$. That is, the $\sigma$-algebra is closed under *countable unions* of its subsets.

An element of the $\sigma$-Algebra are $\mathscr{A}$-measurable sets, or measurable with respect to the $\sigma$-Algebra $\mathscr{A}$. A $\sigma$-algebra is a subset of the power set of $S$, i.e. $\mathscr{A} \subseteq P(X)$.

The difference to between an algebra and a $\sigma$-algebra is that an algebra is closed under *finite* unions of subsets, i.e. $A, B \in \mathscr{A}$ then $A \cup B \in \mathscr{A}$, while a $\sigma$-algebra is closed under *countable* unions, i.e. for $\{A_n : A_n \in \mathscr{A}, n \in \mathbb{N}\}$, $\bigcup_{n=1}^{\infty} A_n \in \mathscr{A}$. Sometimes the terms *field* and *$\sigma$-field* are used instead of *algebra* or *$\sigma$-algebra*. Sometimes measure and integration theory are developed from $\sigma$-rings instead o f $\sigma$-algebras, which are slightly different animals.

#### 1.8.1.1 Example: The Smallest Possible $\sigma$-Algebra

$$\mathscr{A} = \{\emptyset, X\} \tag{1.25}$$

#### 1.8.1.2 Example: The Largest Possible $\sigma$-Algebra

The largest possible $\sigma$-Algebra is the power set:

$$\mathscr{A} = P(X) \tag{1.26}$$

However, there are important examples where it is not possible to define a $\sigma$-algebra on the full power set.

### 1.8.2 Intersection Property

The intersection of $\sigma$ algebras is also a $\sigma$-algebra. That way, a $\sigma$-algebra with the desired properties can be constructed by creating individual $\sigma$-algebras with the properties in question and forming their intersection.

$$\bigcap_i \mathscr{A} \text{ is also a } \sigma - \text{algebra} \tag{1.27}$$

### 1.8.3 Generated $\sigma$-Algebras

For a subset of the powerset $\mathscr{M} \subseteq P(X)$, that does not necessarily have to satisfy the properties of the $\sigma$-Algebra, the *smallest* $\sigma$-algebra that contains $\mathscr{M}$ is the $\sigma$-algebra *generated* by $\mathscr{M}$. It can be constructed through the intersection of all $\sigma$-algebras on $X$ that contain $\mathscr{M}$.

$$\sigma(\mathscr{M}) = \bigcap_{\mathscr{A} \supseteq \mathscr{M}} \mathscr{A} \tag{1.28}$$

$\sigma(\mathscr{M})$ is the $\sigma$-algebra *generated* by $\mathscr{M}$.

#### 1.8.3.1 Example: Generated $\sigma$-Algebra

Take $X = \{a, b, c, d\}$, and $\mathscr{M} = \{\{a\}, \{b\}\}$. Note that $\mathscr{M}$ is not a $\sigma$-algebra. To find the smallest $\sigma$-algebra that contains $\mathscr{M}$, one adds the elements necessary to fulfill the conditions on a $\sigma$-algebra. These are the empty set $\emptyset$ and the full set $X$, the union $\{a, b\}$, and the complements.

$$\sigma(\mathscr{M}) = \{\emptyset, X, \{a\}, \{b\}, \{a, b\}, \{b, c, d\}, \{a, c, d\}, \{c, d\}\} \tag{1.29}$$

### 1.8.4 Borel $\sigma$-Algebras ($\mathscr{B}$)

Borel $\sigma$-Algebras is the $\sigma$-Algebra generated by *open sets*, for example $\sigma(\mathbb{R}^n)$ or $\sigma(\mathscr{M})$ with $\mathscr{M} = (0, 1)$.

## 1.9 Measures, Measurable Spaces and Measure Spaces

A *measure* is a function that gives a volume measure for subsets of a $\sigma$-algebra, where a $\sigma$-algebra is a special sort of collection of subsets of some set $X$. The combination $(X, \mathscr{A})$ of a set $X$ and a $\sigma$-Algebra that is defined on $X$ is called a *measurable space*. A measure is a function defined on the $\sigma$-Algebra and maps to the positive real line (including $\infty$!):

$$\mu : \mathscr{A} \to [0, \infty) \cup \{\infty\} \tag{1.30}$$

It has to satisfy the conditions:

- $\mu(\emptyset) = 0$
- $\mu(\bigcup_i A_i) = \sum_i \mu(A_i)$ if $A_i \cap A_j = \emptyset$ when $i \neq j$ (additive)
- $\mu(\bigcup_i^\infty A_i) = \sum_i^\infty \mu(A_i)$ if $A_i \cap A_j = \emptyset$ when $i \neq j$ ($\sigma$-additive)

Where the infinite sum corresponds to gradually approximating the full volume by taking the union of countably infinitely many subsets. The collection $(X, \mathscr{A}, \mu)$ is a *measure space*.

The reason why the $\sigma$-algebra is included in the definition of a measurable space is because unless a $\sigma$-algebra is defined on a set $X$, there is no guarantee that a measure exists on $X$.

### 1.9.1 Example: Counting Measure

$$\mu(A) = \begin{cases} |A| \text{ if } A \text{ has finite elements} \\ \infty \text{ else} \end{cases} \tag{1.31}$$

### 1.9.2 Example: Dirac Measure

$$\delta_p(A) = \begin{cases} 1 \text{ if } p \in A \\ 0 \text{ else} \end{cases} \tag{1.32}$$

### 1.9.3 Example: Volume Measure

For $X = \mathbb{R}^n$, the conventional volume measure satisfies the properties of:

- $\mu([0,1]^n) = 1$
- $\mu(x + A) = \mu(A)$ for some $x \in \mathbb{R}^n$ (translation invariance)

## 1.10 Measure Problem on $\mathbb{R}$

This is an important example of where it is not possible to define a measure on the whole powerset. This is why measure theory is founded on $\sigma$-algebras.

Take the following measure problem:

We search a measure $\mu$ on $P(\mathbb{R})$ with the properties:

1. $\mu([a,b]) = b - a,\ b > a$
2. $\mu(x + A) = \mu(A),\ A \in \mathbb{R},\ x \in \mathbb{R}$

The only solution to this problem is the trivial map, $\mu(\mathbb{R}) = 0$. A proof goes as follows:

**Claim** Let $\mu$ be a measure on $P(\mathbb{R})$ with $\mu((0,1]) < \infty$ and (2). The only measure that satisfies this condition is the zero measure, $\mu = 0$, which violates (1).

Take the interval $I := (0,1]$ with equivalence relation $x \sim y \equiv x - y \in \mathbb{Q}$. That is, $x$ and $y$ are considered equivalent if they differ by a rational number. That is, we define sets of equivalent numbers, equivalence classes $[x] := \{x + r : r \in \mathbb{Q}, x + r \in I\}$. The equivalence classes are a disjoint decomposition (a partition) of the unit interval $I$ in terms of possibly infinite, countable number of elements.

Take a choice set $A \subseteq I$ that consists of one number from each of the equivalence classes $[x]$ that make up $I$. (cf. sections 1.2.14, 1.7). It has the property:

- For each $[x]$ there is an $a \in A$ with $a \in [x]$
- For all $a, b \in A$, if $a, b \in [x] \implies a = b$

Take the translations of a set $A$, $A_n := r_n + A$ where $(r_n)_{n\in\mathbb{N}}$ are an enumeration of $\mathbb{Q} \cap (1,1]$.

**Proof** The axiom of choice guarantees that it is possible to form a choice set $A$ and the definition of equivalence classes guarantees that the translations of $A_n$ are still choice sets.

**Claim** $A_n \cap A_m =$ if $n \neq m$.

**Proof** Take $x \in A_n \cap A_m \implies \begin{array}{ll} x = r_n + a, & a_n \in A \\ x = r_m + a_m, & a_m \in A \end{array}$

Then, $r_n + a_n = r_m + a + m \implies a_n - a_m = r_m - r_n \in \mathbb{Q} \implies a_n \sim a_m$ according to the definition of the equivalence classes. Also, $a_m, a_n \in [a_m] \implies a_n = a_m \implies r_n = r_m \implies n = m$.

**Claim** $(0,1] \subseteq \bigcup_{n\in\mathbb{N}} A_n \subseteq (-1,2]$

**Proof** Given $r_n \in \mathbb{Q} \cap (-1,1]$, $-1 < r_n \leq 1$ and $a_n \in A,\ 0 < a_n \leq 1$ given that $A \subset (0,1]$, $-1 < r_n + a_n \leq 2$. Therefore $A_n \subseteq (-1,2]\ \forall n \in \mathbb{N}$. Therefore the union $(0,1] \subseteq \bigcup_{n\in\mathbb{N}} A_n \subseteq (-1,2]$.

**Assume** $\mu$ a measure on $P(\mathbb{R})$ with $\mu((0,1]) < \infty$ and (2).

By (2): $\mu(r_n + A) = A$ for all $n \in \mathbb{N}$.

By the property $(0,1] \subseteq \bigcup_{n\in\mathbb{N}} A_n \subseteq (-1,2]$, $\mu((0,1]) \leq \mu(\bigcup_{n\in\mathbb{N}} A_n) \leq \mu((-1,2])$.

Further, $\mu((0,1]) = c < \infty$.

So we can use the disjoint union to express $\mu((-1,2]) = \mu((-1,0] \sqcup (0,1] \sqcup (1,2]) = 3c$.

That implies, $c \leq \sum_{n=1}^{\infty} \mu(A_n) \leq 3c \implies c \leq \sum_{n=1}^{\infty} \mu(A) \leq 3c$. Given that the series is infinite, the only way for this to be true is if $\mu(A) = 0$.

**Conclusion** The measure $\mu(A) = 0$, which implies $\mu((0,1]) = 0$. Because of translation invariance and $\sigma$-additivity, $\mu(\mathbb{R}) = \mu(\bigcup_{m\in\mathbb{Z}^+}(m, m+1]) = 0$. That means that the only measure that satisfies the conditions stated in the measure problem assigns 0 to the length of the whole real line.

## 1.11 Measurable Maps

**Definition** Given two measurable spaces $(\Omega_1, \mathscr{A}_1), (\Omega_2, \mathscr{A}_2)$, a measurable map with respect to $\mathscr{A}_1, \mathscr{A}_2$ is $f : \Omega_1 \to \Omega_2$ if $f^{-1}(A_2) \in \mathscr{A}_1$ for all $A_2 \in \mathscr{A}_2$. That is, $f$ connects the two $\sigma$-algebras $\mathscr{A}_1$ and $\mathscr{A}_2$ in that the *preimage* (cf. section 1.2.12) of an element $A_2 \in \mathscr{A}_2$, $f^{-1}(A_2)$ is an element of $\mathscr{A}_1$.

### 1.11.1 Example: Characteristic Function, Indicator Function

Take the measurable spaces $(\Omega, \mathscr{A})$ and $(\mathbb{R}, \mathscr{B}(\mathbb{R}))$.

$$\chi_A : \omega \to \mathbb{R}, \;\; \chi_A(\omega) = \begin{cases} 1, \; \omega \in A \\ 0, \; \omega \notin A \end{cases} \tag{1.33}$$

For all measurable $A \in \mathscr{A}$, $\chi_A$ is a measurable map. The four possible preimages are:

$$\begin{aligned} &\chi_A^{-1}() =, \; \chi_A^{-1}(\mathbb{R}) = \Omega \\ &\chi_A(\{1\}) = A \; \chi_A(\{1\}) = A^c \end{aligned} \tag{1.34}$$

Where all of the preimages are contained in $\mathscr{A}_1$.

### 1.11.2 Example: Composition of Measurable Maps

Take measurable spaces $(\Omega_1, \mathscr{A}_1), (\Omega_2, \mathscr{A}_2), (\Omega_3, \mathscr{A}_3)$, connected through measurable maps:

$$\begin{aligned} f &: \Omega_1 \to \Omega_2 \\ g &: \Omega_2 \to \Omega_3 \end{aligned} \tag{1.35}$$

Then $f \circ g : \Omega_1 \to \Omega_3$ is also measurable, because $(g \circ f)^{-1}(A_3) = \underbrace{f^{-1}(\underbrace{g^{-1}(A_3)}_{\in \mathscr{A}_2})}_{\in \mathscr{A}_1}$.

### 1.11.3 Example: Sums and Products of Measurable Maps

Given $(\Omega, \mathscr{A}), (\mathbb{R}, \mathscr{B}(\mathbb{R}))$, if $f, g : \Omega \to \mathbb{R}$ are measurable maps, then $f + g, f - g, f \times g$ and $|f|$ are also measurable maps. This follows from the property that compositions of measurable maps are also measurable.

## 1.12 Lebesgue Integrals

### 1.12.1 Lebesgue Integrals for Step Functions

Take a measure space $(X, \mathscr{A}, \mu)$ and the measurable space $(\mathbb{R}, \mathscr{B})$, where $X$ is any set, $\mathscr{A}$ is a special collection of subsets of $X$, $\mu$ is a map $\mu : \mathscr{A} \to [0, \infty]$, and $\mathscr{B}$ is a Borel $\sigma$-algebra on $\mathbb{R}$. Then let $f : X \to \mathbb{R}$ be a measurable map, so that $f^{-1}(E) \in \mathscr{A}$ for all $E \subseteq \mathscr{B}(\mathbb{R})$.

**Characteristic Function** The integral of a characteristic function $\chi_A$ is $I(\chi_A) = \mu(A)$.

**Simple Functions** Simple functions are, for example, step functions, staircase functions, etc., that can be expressed in terms of a sum of characteristic functions. That is, for $A_1, A_2, ..., A_n \in \mathscr{A}$, $c_1, c_2, ..., c_n \in \mathbb{R}$:

$$f(x) = \sum_{i=1}^{n} c_i \chi_{A_i}(x) \tag{1.36}$$

Since characteristic functions are measurable and sums of characteristic functions are measurable, simple functions are measurable. Then the integral would be:

$$I(f) = \sum_{i=1}^{n} c_i \mu(A_i) \tag{1.37}$$

However, there is a problem with this definition, because of the possibility of having to subtract infinitely large intervals. The options are to restrict simple functions to either:

- restrict $A_i$ to be finite size sets.
- restrict $c_i$ to be positive.

**S+** The set of positive functions $S^+ := \{f : X \to \mathbb{R} : f \text{ simple function}, \ f \geq 0\}$ where $f$ is measurable and has finitely many values (picture a staircase, rather than a smooth curve). For $f \in S^+$, choose representation $f(x) = \sum_{i=1}^{n} c_i \chi_{A_i}(x)$, $c_i \geq 0$.

**Lebesgue Integral for Simple Functions** The Lebesgue integral of $f \in S^+$ with respect to the measure $\mu$:

$$\int_X f(x) d\mu(x) = \int_X f d\mu = I(f) = \sum_{i=1}^{n} c_i \mu(A_i) \ \in [0, \infty] \tag{1.38}$$

This is a well-defined object that is independent of the specific representation of $f(x)$. It has the properties:

- $I(\alpha f + \beta g) = \alpha I(f) + \beta I(g)$ for $\alpha, \beta \geq 0$
- $f \leq g \implies I(f) \leq I(g)$ (monotonicity)

The Lebesgue integral for simple functions enables defining the integral for more complex functions by approximating them.

**Definition** Given a non-negative function $f : X \to [0, \infty)$, there are positive simple functions $h \in S^+$ that approximate it from below $\{h : h \in S^+, h \leq f\}$ with $h = \sum_{i=1}^{n} c_i \chi_{A_i}$. Then the integral of $f$ is given by the largest possible function within that set.

The *Lebesgue Integral* of a function $f$ with respect to a measure $\mu$ is

$$\int_X f d\mu := \sup\{I(h) : h \in S^+, h \leq f\} \tag{1.39}$$

$f$ is called $\mu$-integrable if $\int_X f d\mu < \infty$.

The only thing that was needed to define the integral was a measure space $(X, \mathscr{A}, \mu)$

## 1.13 Monotone Convergence Theorem

A convergence theorem should show circumstances under which a limit can be pulled into the integral, and monotone implies that the theorem will deal with a monotonically changing series.

**Preliminaries** Take the measurable positive function $f : X \to [0, \infty)$ which has Lebesgue integral $\int_X f d\mu \in [0, \infty]$.

**Equality** If $f = g$, $\mu$ almost everywhere, then $\int_X f d\mu = \int_X g d\mu$.

That is, if $f = g$ "almost everywhere" with respect to the measure $\mu$, then the integrals are identical. This is more general than to simply say $f = g$. Rather, it requires $\mu(\{x \in X : f(x) \neq g(x)\}) = 0$. The Lebesgue integral "cannot see" things that happen on 0-measure sets. For example, if the measure is a length measure $\mu([a, b]) = b - a$, then, if $f$ and $g$ differ at a single point, then $\int_X f$ and $\int_X g$ are still the same.

**Monotonicity** Similartly, if $f \leq g$, $\mu$ almost everywhere, then $\int_X f d\mu \leq \int_X g d\mu$.

Given that a positive simple function $h : X \to [0, infty)$ assumes a finite amount of different values, it is permissible to represent $h(x)$ in terms of the values $t$ that it assumes:

$$h(x) = \sum_{i=1}^{n} c_i \chi_{A_i}(x) = \sum_{t \in h(X)} t \chi_{\{x \in X : h(x) = t\}} \tag{1.40}$$

Using this representation, Then the Lebesgue integral:

$$I(h) = \sum_{t \in h(X) \setminus \{0\}} t\mu(\{x \in X : h(x) = t\}) \tag{1.41}$$

Where omitting the $t \in \{0\}$ makes no difference to the integral. Divide the set $X$ into $X = \tilde{X} + \tilde{X}^c$ with $\mu(\tilde{X}^c) = 0$ and $\tilde{X} \in \mathscr{A}$. Then let:

$$\tilde{h}(X) := \begin{cases} h(x), x \in \tilde{X} \\ a, x \in \tilde{X}^c \end{cases} \tag{1.42}$$

Where $a \in [0, infty)$. Then:

$$\tilde{h}(x) = \int_{t \in h(X)} t \chi_{x \in \tilde{X} : h(x) = t} + a \chi_{\tilde{X}^c} \tag{1.43}$$

Then the integral:

$$I(\tilde{h}) = \sum_{t \in h(X)} t\mu(\{x \in \tilde{X} : h(x) = t\}) + a \underbrace{\mu(\tilde{X}^c)}_{0} \tag{1.44}$$

So that $I(h) = I(\tilde{h})$. This means that we can modify a simple function $h$ on a set with measure 0 however we like without affecting the integral. This is enough to prove that if $f \leq g$ then $\int_X f d\mu \leq \int_X g d\mu$ in the "almost everywhere" sense. Divide $X$ into $\tilde{X} := \{x \in X : f(x) \leq g(x)\}$ and $\tilde{X}^c$ with measure zero, where $f(x) \leq g(x)$ is not true. Since $I(h) = I(\tilde{h})$:

$$\begin{aligned} \textstyle\int_X f d\mu &= \sup\{I(h) : h \in S^+, h \leq f\} = \sup\{I(\tilde{h}) : h \in S^+, \tilde{h} \leq f \leq g \text{ on } \tilde{X}\} \\ &\leq \sup\{I(\tilde{h}) : h \in S^+, \tilde{h} \leq g \text{ on } \tilde{X}\} = \textstyle\int_X g d\mu \end{aligned} \tag{1.45}$$

**Zero Integral** $f = 0$, $\mu$ almost everywhere, then $\int_X d\mu = 0$. (Note that only positive functions $f$ are considered.

**Definition: Monotone Convergence Theorem** Take a measure space $(X, \mathscr{A}, \mu)$ and non-negative measurable functions $f_n : X \to [0, \infty), f : X \to [0, \infty)$, with $f_1 \leq f_2 \leq f_3 \leq ...$ with $\mu - a.e.$ (almost everywhere) and $\lim_{n\to\infty} f_n(x) = f(x).\mu - a.e.(x \in X)$, where the limit is taken in a point-wise sense. That is, it applies for a fixed point $x \in X$.

The monotone convergence theorem states that, given a monotonic series of functions, the limit can be pulled into the integral:

$$\lim_{n\to\infty} \int_X f_n d\mu = \int_X \lim_{n\to\infty} f_n d\mu = \int_X f d\mu \tag{1.46}$$

### 1.13.1 Application: Series

Take a series of measurable non-negative functions $(g_n)_{n\in\mathbb{N}}$, $g_n : X \to [0, \infty]$ measurable for all $n$, which does not necessarily have monotonic behavior. Then:

$$\sum_{n=1}^{\infty} g_n : X \to [0, \infty] \text{ is measurable} \tag{1.47}$$

Now, while the series $(g_n)_{n\in\mathbb{N}}$ was not necessarily monotonically increasing, the sum is. therefore:

$$\int_X \sum_{n=1}^{\infty} g_n d\mu = \sum_{n=1}^{\infty} \int_X g_n d\mu \tag{1.48}$$

## 1.14 Fatou's Lemma

The starting point is, again, a measure space $(X, \mathscr{A}, \mu)$ and a series of non-negative functions $f_n : X \to [0, \infty]$ that are measurable for all $n \in \mathbb{N}$. Then:

$$\int_X \liminf_{n\to\infty} f_n d\mu \leq \liminf_{n\to\infty} \int_X f_n d\mu \tag{1.49}$$

Where the limit inferior $\liminf_{n\to\infty} f_n(x)$ of a sequence of numbers $f_n(x)$ is $\lim_{n\to\infty}(\inf_{k\geq n} f_k(x))$. That is, that is, $\liminf_{n\to\infty} f_n$ is actually a function $(\liminf_{n\to\infty} f_n)(x)$, that, for a series of functions $f_n$, picks out the lower limit of the series of functions evaluated at the point $x$ as $n \to \infty$. Fatou's Lemma shows that the integral over the infimum of a series of functions $f_n$ as $n \to \infty$ is smaller than the infimum of the series of integrals of those functions as $n \to \infty$. That seems intuitively correct – the integral over the minimum value of a collection of positive functions should be smaller or equal than the minimum of the integrals. Fatou's Lemma follows from the monotone convergence theorem by recognizing the infimum of a series of functions $\inf_{k\geq n} f_k$ as a function $g_n = (\inf_{k\geq n} f_k)(x)$ and recognizing that $g_n$ forms a monotonically increasing series $g_1 \leq g_2 \leq ...$, because as $n$ get's larger, the infimum $\inf_{k\geq n} f_k$ must get either larger or stay the same.

## 1.15 Lebesgue's Dominated Convergence Theorem

Take a measure space $(X, \mathscr{A}, \mu)$, define a set of Lebesgue integrable functions $\mathscr{L}(X, \mathscr{A}, \mu)$, which is simply written $\mathscr{L}(\mu)$ because the measurable space is assumed fixed from the context. The set is given by: $\mathscr{L}(\mu) := \{f : X \to \mathbb{R},\ \text{measurable} : \int_X |f| d\mu\}$, where the exponent matters. So, the set of "L1-integrable functions", $\mathscr{L}^1(\mu) := \{f : X \to \mathbb{R},\ \text{measurable} : \int_X |f|^1 d\mu\}$. Since the Lebesgue integral was so far defined only for positive functions $f$, for $f \in \mathscr{L}(\mu)$ we can write $f = f^+ - f^-$ where $f^+$ and $f^-$ are both positive. $f$ can then be integrated by subtracting the integrals of $f^+$ and $f^-$.

**Theorem** Given a sequence of complex valued functions $f_n : X \to \mathbb{R}$ that are measurable for all $n \in \mathbb{N}$, which pointwise approaches the limit function $\lim_{n\to\infty} f_n(x) = f(x)$ for $x \in X$ almost everywhere with respect to $\mu$ (almost everywhere meaning that the limit expression holds for the $f_n$ evaluated at points $x$ everywhere except possibly some set of of points with measure zero).

If, pointwise, it is true $|f_n| \leq g$ with $g \in \mathscr{L}^1(\mu)$ for all $n \in \mathbb{N}$, then $g$ is an *integrable majorant*. The implication is: $f_1, f_2, f_3, ... \in \mathscr{L}^1(\mu), f \in \mathscr{L}^1(\mu)$.

Further, convergence gives:

$$\lim_{n\to\infty} \int_X f_n d\mu = \int_X f d\mu \tag{1.50}$$

Which looks exactly like the monotone convergence theorem, except that the requirement on the sequence $f_n$ is only that there is an integrable majorant $g_n$ in $\mathscr{L}^1(\mu)$. The theorem gives a sufficient condition under which the almost everywhere convergence of a sequence of functions guarantees convergence in the $\mathscr{L}^1(\mu)$ norm. The theorem can be proved, in particular, via Fatou's Lemma, which guarantees an upper bound on the integral of the limit of the infimum of a sequence of positive functions. The integrable majorant $g \geq |f|$ allows for the construction of a series of non-negative functions $h_n := 2g - |f_n - f| \geq 0$, recognizing that $|f_n - f| \leq |f_n| + |f| \leq 2g$.

### 1.15.1 Triangle Inequalities

$$|f + g| \leq |f| + |g| \tag{1.51}$$

$$|\int_X f| \leq \int_X |f| \tag{1.52}$$

## 1.16 Carathéodory's Extension Theorem

Carathéodory's extension theorem states that, given a set $X$, for a semiring of sets $\mathscr{A} \subseteq P(X)$, with a pre-measure $\mu : \mathscr{A} \to [0, \infty]$, there exists a unique extension to the semiring and the pre-measure that is the unique $\sigma$-algebra $\sigma(\mathscr{A})$ generated by the semiring and a measure $\tilde{\mu} : \sigma(\mathscr{A}) \to [0, \infty]$. The theorem is of particular importance because it guarantees the existence and uniqueness of the Lebesgue measure.

### 1.16.1 Semirings of Sets

For a set $X$, semirings of sets are a collection of sets $\mathscr{A} \subseteq P(X)$ that weaker criteria than $\sigma$-algebras.

The criteria are:

- $\in \mathscr{A}$
- For $A, B \in \mathscr{A}$, $A \cap B \in \mathscr{A}$
- For $A, B \in \mathscr{A}$, there is a difference operation $A \setminus B$. It is not required that $A \setminus B$ is an element of $\mathscr{A}$, but it is required that there is a union of sets $\bigcup_i A_i = A \setminus B$ with $A_i \in \mathscr{A}$.

#### 1.16.1.1 Example: Intervals on $\mathbb{R}$

$\mathscr{A} := \{[a, b) : a, b \in \mathbb{R}, a \leq b\}$ is not a $\sigma$-algebra because $\mathbb{R} \notin \mathscr{A}$. However, the generated $\sigma$-algebra $\sigma(\mathscr{A}) = \mathscr{B}(\mathbb{R})$ is unique and is the Borel $\sigma$-algebra. $\mathscr{A}$ fulfills the criteria of a semiring.

### 1.16.2 Pre-measure

A pre-measure $\mu : \mathscr{A} \to [0, \infty]$ in a semiring of sets $\mathscr{A}$ fulfills:

1. $\mu() = 0$

2. $\mu(\bigcup_{j=1}^{\infty} = \sum_{j=1}^{\infty} \mu(A_j)$ for $A_j \in \mathscr{A}$ if $A_i \cap A_j =$ if $i \neq j$ and also $\bigcup_{j=1}^{\infty} A_j \in \mathscr{A}$, where this condition is not necessarily satisfied for semirings, because, in contrast to $\sigma$-algebras, the union of elements of semirings is not necessarily also in the semiring.

### 1.16.3 Lebesgue Measure

Take the semiring $\mathscr{A} := \{[a,b) : a, b \in \mathbb{R}, a \leq b\}$ and the premeasure $\mu : \mathscr{A} \to [0, \infty]$, $\mu([a,b)) = b - a$, then Carathéodory's theorem guarantees that there is a unique extension, which is $\mathscr{B}(\mathbb{R})$ and the Lebesgue measure.

## 1.17 Lebesgue-Stieltjes Measures

Lebesgue-Stieltjes measures are measures that are constructed for monotonically increasing (non-decreasing) functions $F$. Take non-decreasing functions $F : \mathbb{R} \to \mathbb{R}$. Such functions could be discontinuous or constant, as long as they do not decrease. Then, on the semiring $\mathscr{A} := \{[a,b) : a, b \in \mathbb{R}, a \leq b\}$, a pre-measure $\mu_F : \mathscr{A} \to [0, \infty]$ so that $\mu_F([a,b)) = F(b^-) - F(a^-)$ where the superscript $(-)$ clarifies which side of a discontinuity has to be included, should $F$ be discontinuous at $a$ or $b$. This is consistent with the boundaries of the interval $[a,b)$. (It is equally well possible to approach this using intervals $(a,b]$.)

Now, Carathéodory's Theorem ensures that this can be extended to one unique measure and $\sigma$-algebra, $\mu_F : \mathscr{B}(\mathbb{R}) \to [0, \infty]$, which is called the Lebesgue-Stieltjes measure for the function $F$.

### 1.17.1 Example: Lebesgue Measure

An easy example is $F(x) = x$. Then $\mu_F([a,b)) = b - a$ which is the Lebesgue measure.

### 1.17.2 Example: Zero Measure

Take $F(x) = 1$, then $\mu_F([a,b)) = 0$, which is the zero measure.

### 1.17.3 Example: Dirac Measure

Take the discontinuous step function $F(x) = \begin{cases} 0, x < 0 \\ 1x \geq 0 \end{cases}$. Then $\mu_F([-\epsilon, \epsilon)) = 1$ for arbitrarily small $\epsilon$. That is, the Dirac Measure $\delta_0$ is the Lebesgue-Stieltjes Measure with respect to a step function on $\mathbb{R}$.

### 1.17.4 Example: Density Functions

Take a function $F : \mathbb{R} \to \mathbb{R}$ that is monotonically increasing and continuously differentiable, i.e. $F' : \mathbb{R} \to [0, \infty)$. Then the measure is simply the interval $\mu_I([a,b)) = F(b) - F(a) = \int_a^b F'(x)dx$. That means that the Lebesgue-Stieltjes measure $\mu_F : \mathscr{A} \to \int_{\mathscr{A}} F'(x)dx$ with respect to the monotonically increasing function $F$ is a measure from a Borel set to an integral of the derivative of $F'(x)$. In this context, $F'(x)$ is a *density function.*

# 2 Combinatorics

**Contents of this chapter**

## 2.1 Combinatorial Identities and Expansions

### 2.1.1 Binomial Coefficients and Binomial Expansions

For two positive integers $n$ and $k$, the binomial coefficient "$n$ choose $k$" is:

$$\binom{n}{k} = \begin{cases} \frac{n!}{k!(n-k)!} \text{ for } n \geq k \\ 0 \text{ otherwise} \end{cases} \tag{2.1}$$

The term can be defined for negative arguments, which is comes up often when working with generating functions.

$$\binom{-n}{k} = \begin{cases} (-1)^k \binom{n+k-1}{k} \text{ for } n \geq k \\ 0 \text{ otherwise} \end{cases} \tag{2.2}$$

$$\binom{-n}{-k} = \begin{cases} (-1)^{k-n} \binom{k-1}{k-n} \text{ for } n \geq k \\ 0 \text{ otherwise} \end{cases} \tag{2.3}$$

The generalizations can, for example, be derived using symmetry arguments and the Gamma function, which is the generalization of the factorial to non-integers (cf. Kronenburg (2011)).

The binomial expansion can be proven either by expanding the polynomial or by creating the Taylor series for the polynomial.

$$(x+y)^n = \sum_{k=0}^{n} \binom{n}{k} x^k y^{n-k} \tag{2.4}$$

This holds also for negative integer exponents $n$, in which case:

$$\frac{1}{(y+x)^n} = \sum_{k=0}^{\infty} \binom{-n}{k} x^k y^{-n-k} = (-1)^k \binom{n+k-1}{k} x^k y^{-(n+k)} \tag{2.5}$$

#### 2.1.1.1 Derivation of the Binomial Theorem for a Negative Exponent

Let $f(x) = \frac{1}{(y+x)^n} = (y+x)^{-n}$. The Taylor expansion about the point $x = 0$ is:

$$f(x) = \sum_{k=0}^{\infty} \frac{f^{(k)}(x=0)}{k!} x^k \tag{2.6}$$

The derivatives of $f$ are:

$$\begin{aligned}
f^{(0)}(x) &= (y+x)^{-n} \\
f^{(1)}(x) &= (y+x)^{-(n+1)}(-1)n \\
f^{(2)}(x) &= (y+x)^{-(n+2)}(-1)^2 n(n+1) \\
&\vdots \\
f^{(k)}(x) &= (y+x)^{-(n+k)}(-1)^k n(n+1)\ldots(n+k-1) \\
&= (y+x)^{-(n+k)}(-1)^k \tfrac{(n+k-1)!}{(n-1)!}
\end{aligned} \tag{2.7}$$

Combining Eqns. 2.6 and 2.7 gives:

$$f(x) = \sum_{k=0}^{\infty} (-1)^k \binom{n+k-1}{k} x^k y^{-(n+k)} \tag{2.8}$$

#### 2.1.1.2 Derivation of the Binomial Theorem for a Fractional Exponent

Following much the same logic as for a negative exponent, for a fractional exponent the Taylor series is also infinite.

Let $f(x) = (y+x)^m$ with $m \notin \mathbb{Z}$.

The derivatives of $f$ are:

$$\begin{aligned}
f^{(0)}(x) &= (y+x)^m \\
f^{(1)}(x) &= (y+x)^{m-1} m \\
f^{(2)}(x) &= (y+x)^{m-2} m(m-1) \\
&\vdots \\
f^{(k)}(x) &= (y+x)^{n-k} m(m-1)\ldots(m-k+1)
\end{aligned} \tag{2.9}$$

Combining Eqns. 2.6 and 2.9 gives:

$$f(x) = \sum_{k=0}^{\infty} \binom{m}{k} x^k y^{m-k} \tag{2.10}$$

Where, boldly, I defined $\binom{m}{k}$ to mean:

$$\binom{m}{k} = \frac{m(m-1)(m-2)\ldots(m-k+2)(m-k+1)}{k!} \tag{2.11}$$

### 2.1.2 Multinomial Expansion

$$(x_1 + x_2 + x_3 + \ldots + x_k)^n = \sum_{\substack{i_1, i_2, i_3, \ldots, i_k \\ \sum_j i_j = n}} \binom{n}{i_1, i_2, \ldots i_k} x^{i_1} x^{i_2} x^{i_3} \ldots x^{i_k} \tag{2.12}$$

with:

$$\binom{n}{i_1, i_2, \ldots i_k} = \frac{n!}{i_1! i_2!, i_3! \ldots i_k!} \tag{2.13}$$

Where the sum over all possible exponents $i_j$ so that $\sum_j i_j = n$ has $\binom{n+k-1}{n}$ terms.

### 2.1.3 Unnamed Polynomial Identity

I don't know what this is called, but it's useful.

$$\prod_i^n (1 - x_i) = \sum_{s=0}^n (-1)^s \sum_{0 \le \underbrace{i_1, i_2, \ldots, i_s}_{\{i\}_s} \le n} \prod_{i \in \{i\}_s} x_i \tag{2.14}$$

Where $\{i\}_s$ is a set of $s$ indices that range between $0$ and $n$, and the sum is over all possible such sets, of which there are $\binom{n}{s}$.

### 2.1.4 Factorial Expansion

$$x^{\underline{n}} = \frac{x!}{(x-n)!} = \sum_{k=0}^n s(n,k) x^k \tag{2.15}$$

where

$$s(n,k) = (-1)^{n-k} \left[ \begin{array}{c} n \\ k \end{array} \right] \tag{2.16}$$

are the stirling numbers of the first kind.

### 2.1.5 Stirling Numbers of the Second Kind

Stirling numbers of the second kind $S(k,n)$ measure the amount of ways in which $k$ objects can be divided into $n$ non-empty groups. They give the number of onto functions from a set of $k$ distinct objects to $n$ indistinct recipients. For example: how many ways can a set of $k$ pool balls be put into $n$ bags, so that there is at least one ball in each bag. All the pool balls have numbers on them and have different colors, so that $n = 2$ bags containing $[(1,2,3),(4)]$ and $[(1,2,4),(3)]$ count as different. This sort of problem is discussed at length in section 2.2

They are given by an explicit formula:

$$S(k,n) = \frac{1}{n!} \sum_{i=1}^n (-1)^{n-j} \binom{n}{j} j^k \tag{2.17}$$

They can also be generated via the recurrence relation:

$$\left\{ \begin{array}{c} k+1 \\ n \end{array} \right\} = n \left\{ \begin{array}{c} k \\ n \end{array} \right\} + \left\{ \begin{array}{c} k \\ n-1 \end{array} \right\} \tag{2.18}$$

The recurrence relation is explained by adding the combinations corresponding to two cases. If the $k+1$st object is added to one of the $n$ existing subsets with $k$ objects, then that corresponds to:

$$n\left\{ \begin{array}{c} k \\ n \end{array} \right\} = 1 \tag{2.19}$$

Possbilities. If the $k + 1$st object is in a set by itself (a singleton), then the remaining objects are distributed over $n - 1$ set. The combinations arising from this are:

$$\left\{ \begin{array}{c} k \\ n-1 \end{array} \right\} = 1 \tag{2.20}$$

Furthermore, the following holds:

$$\left\{ \begin{array}{c} 0 \\ 0 \end{array} \right\} = 1 \tag{2.21}$$

$$\left\{ \begin{array}{c} k \\ 0 \end{array} \right\} = \left\{ \begin{array}{c} 0 \\ n \end{array} \right\} = 0 \tag{2.22}$$

And $S(k, n) = 0$ if $n > k$.

## 2.2 Distributions: The Twentyfold Way

These notes (in particular) need review. There is a deeper perspective on distributions that is constructed in terms of functions and equivalence classes, which is poorly developed here so far, and I think there are some errors too. The different "interpretations" are also not well-developed here.

The twentyfold way is a taxonomy of distribution problems developed by Kenneth Bogard in his book *Combinatorics through Guided Discovery* (Bogart 2004). It divides up the way in which $k$ objects may be assigned to $n$ individuals, subject to whether the objects are distinct or identical, and subject to conditions on how the objects are received.

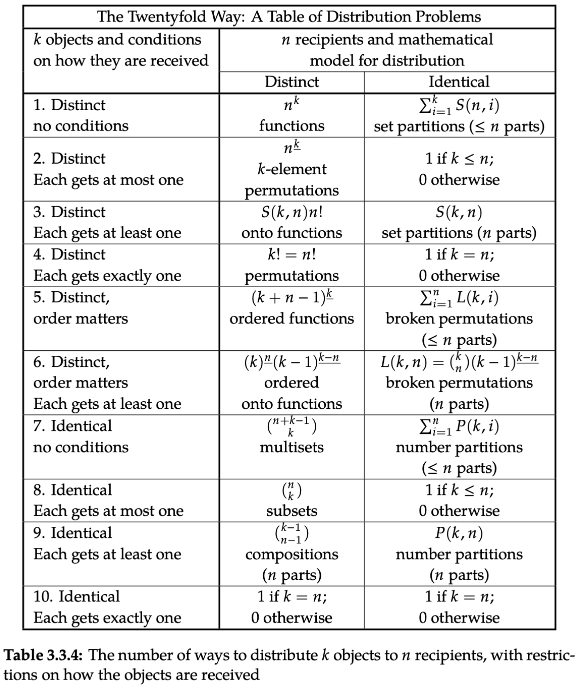
> *When we are passing out objects to recipients, we may think of the objects as being either identical or distinct. We may also think of the recipients as being either identical (as in the case of putting fruit into plastic bags in the grocery store) or distinct (as in the case of passing fruit out to children). We may restrict the distributions to those that give at least one object to each recipient, or those that give exactly one object to each recipient, or those that give at most one object to each recipient, or we may have no such restrictions. If the objects are distinct, it may be that the order in which the objects are received is relevant (think about putting books onto the shelves in a bookcase) or that the order in which the objects are received is irrelevant (think about dropping a handful of candy into a child's trick or treat bag). If we ignore the possibility that the order in which objects are received matters, we have created* $2 \times 2 \times 4 = 16$ *distribution problems. In the cases where a recipient can receive more than one distinct object, we also have four more problems when the order objects are received matters. Thus we have 20 possible distribution problems.* - Bogart, *Combinatorics Though Guided Discovery*, Chapter 3.

What I like about this approach is that the challenge with most of the basic combinatorics problems is to figure out the right way of counting. For this reason, the idea of having a unified handbook-like taxonomy is very appealing. The weakness (in my opinion) is that the language of "objects" and "recipients" is unclear because in practice it's not obvious which is which: if there are $k$ students and $n$ teachers, do the teachers receive students, or do the students receive a teacher?

A way to resolve this is to say that an object can have only one recipient, but that a recipient might receive more than one object. A more formal path is to think of the act of creating combinations in terms of functions.

- The elements of the domain are the objects.
- The elements of the range are the recipients.
- A function can be many-to-one, but it should not be one-to-many.

Figure 2.1: Bogart's Twentyfold Way



| The Twentyfold Way: A Table of Distribution Problems | | |
|---|---|---|
| $k$ objects and conditions on how they are received | $n$ recipients and mathematical model for distribution | |
| | Distinct | Identical |
| 1. Distinct<br>no conditions | $n^k$<br>functions | $\sum_{i=1}^{k} S(n,i)$<br>set partitions ($\leq n$ parts) |
| 2. Distinct<br>Each gets at most one | $n^{\underline{k}}$<br>$k$-element permutations | 1 if $k \leq n$;<br>0 otherwise |
| 3. Distinct<br>Each gets at least one | $S(k,n)n!$<br>onto functions | $S(k,n)$<br>set partitions ($n$ parts) |
| 4. Distinct<br>Each gets exactly one | $k! = n!$<br>permutations | 1 if $k = n$;<br>0 otherwise |
| 5. Distinct,<br>order matters | $(k+n-1)^{\underline{k}}$<br>ordered functions | $\sum_{i=1}^{n} L(k,i)$<br>broken permutations ($\leq n$ parts) |
| 6. Distinct,<br>order matters<br>Each gets at least one | $(k)^{\underline{n}}(k-1)^{\underline{k-n}}$<br>ordered onto functions | $L(k,n) = \binom{k}{n}(k-1)^{\underline{k-n}}$<br>broken permutations ($n$ parts) |
| 7. Identical<br>no conditions | $\binom{n+k-1}{k}$<br>multisets | $\sum_{i=1}^{n} P(k,i)$<br>number partitions ($\leq n$ parts) |
| 8. Identical<br>Each gets at most one | $\binom{n}{k}$<br>subsets | 1 if $k \leq n$;<br>0 otherwise |
| 9. Identical<br>Each gets at least one | $\binom{k-1}{n-1}$<br>compositions ($n$ parts) | $P(k,n)$<br>number partitions ($n$ parts) |
| 10. Identical<br>Each gets exactly one | 1 if $k = n$;<br>0 otherwise | 1 if $k = n$;<br>0 otherwise |

**Table 3.3.4:** The number of ways to distribute $k$ objects to $n$ recipients, with restrictions on how the objects are received

**Favorite Teachers** At a school with $k$ students and $n$ teachers, the students all have a favorite teacher. (They might all like the same one.). How many ways are there for all of the $k$ students to pick a favorite?
*Objects:* $k$ students. *Recipients:* $n$ teachers. Many students might have one favorite teacher. There are $n^k$ combinations.

**Assembling a Team** Out of a choice of $n$ athletes, a coach must assemble a team of $k$. How many ways are there to form a team?
*Objects:* n athletes. *Recipients:* team, not on the team. Many athletes can be assigned to one outcome of being on the team or not being on the team. There are $\binom{n}{k}$ combinations for the team, which is the same number as the $\binom{n}{n-k}$ selections for the bench.

## 2.2.1 Distinct Objects

### 2.2.1.1 Distinct Recipients

The $k$ objects are assigned to $n$ recipients with no conditions as to the number of objects each recipient receives. This is the same as assigning the elements of a $k$-tuple from a selection of $n$ with replacement.

$$\begin{aligned} S &= \{(a_1, a_2, ..., a_k) | a_i \in A, |A| = n\} \\ |S| &= n^k \end{aligned} \tag{2.23}$$

**Pool Balls into Labeled Buckets** All possible ways to put $k$ pool balls, which all have different numbers and colors, into $n$ labeled buckets. Some of the buckets might be empty, and others might contain more than one of the pool balls.

**Functions** All possible functions $f : x \to y$ with $\{x | x \in A, |A| = k\}$ and $\{y | y \in B, |B| = n\}$.

**Binary Strings of Length $k$** The $k$ distinct positions of a binary string $(i_1, i_2, ..., i_k)$ of length $k$ are assigned to an element of the set $A \in [0, 1]$. The number of possible binary strings of length $k$ is $2^k$.

**Subsets of a $k$-Element Set** The subsets of a set of $k$ distinct elements are formed by assigning each of its $k$ distinguishable elements to one of the two labels $A \in [\text{included}, \text{excluded}]$. The number of possible subsets, including the empty subset and the full set, is $2^k$.

#### 2.2.1.2 Indistinct Recipients

The $k$ objects are assigned to a recipient that is not distinct.

$$|A| = \sum_{i=1}^{k} S(n, i) \tag{2.24}$$

Where $S(k, n)$ is the Stirling number of the second kind that gives the number of ways that $k$ objects can be distributed across $n$ non-empty indistinct sets. The sum above takes care of the case where the $k$ objects are divided into up to $n$ collections.

A closed form expression for the Stirling Numbers of the second kind is (c.f. section 2.1.5):

$$S(k, n) = \left\{ \begin{array}{c} k \\ n \end{array} \right\} = \frac{1}{n!} \sum_{j=0}^{n} (-1)^{n-j} \binom{n}{j} j^k \tag{2.25}$$

**Pool Balls into Unlabeled Bags** All possible ways to put $k$ pool balls, which all have different numbers and colors, into $n$ unlabeled bags. Some of the bags might be empty, and others might contain more than one of the pool balls.

### 2.2.2 Distinct Objects, Every Recipient Receives At Most One

#### 2.2.2.1 Distinct Recipients

At most one of $k$ distinct objects are assigned to one of $n$ distinct recipients. This is the same as assigning the elements of a $k$-tuple from a selection of $n$ without replacement, so that first there are $n$ choices, then $n - 1$ choices, $n - 2$, etc.

$$\begin{aligned} &S = \{(a_1, a_2, ..., a_k) | a_i \in A, |A| = n, a_i \neq a_j\} \\ &|S| = \tfrac{n!}{(n-k)!} = n^{\underline{k}} \;\text{ if } k \leq n, \; 0 \text{ otherwise.} \end{aligned} \tag{2.26}$$

**At Most One Pool Ball into Labeled Buckets** All possible ways to put $k$ pool balls, which all have different numbers and colors, into $n$ labeled buckets, but the buckets can have at most one pool ball in them. In other words, you choose any $k$ out of the $n$ labeled buckets and put one pool ball into them.

**One-to-One Functions** All possible functions $f : x \to y$ with $\{x | x \in A, |A| = k\}$ and $\{y | y \in B, |B| = n\}$ subject to the constraint that $f(a) = f(b)$ implies $a = b$. That is, the functions are one-to-one, or injective.

**k-element Permutations of $n$ elements** Each of the $k$ positions in a $k$-element permutation are distinct objects. These are each assigned to one of $n$ possible values, where each value can only show up once.

**Books on a Shelf** How many ways are there to order $k$ books on a library shelf when there are $n$ different books available.

#### 2.2.2.2 Indistinct Recipients

At most one of $k$ distinct objects are assigned to one of $n$ indistinct recipients. This is the same as assigning the elements of a $k$-tuple from a selection of $n$ without replacement. Except ,since the recipients are all indistinct, there is only one type of choice of recipient for each object. Either each object finds a recipient if $k \leq n$, or it is impossible to distribute at most one object to each recipient because $n < k$.

$$\begin{aligned} &S = \{(a_1, a_2, ..., a_k) | a_i \in A, |A| = n, a_i = a_j\} \\ &|S| = 1 \text{ if } k \leq n, \; 0 \text{ otherwise.} \end{aligned} \tag{2.27}$$

**At Most One Pool Ball into Unlabeled Bags** All possible ways to put $k$ pool balls, which all have different numbers and colors, into $n$ unlabeled bags, but the bags can have at most one pool ball in them. The bags are identical, so the only result is one where there are $k$ bags with a ball in them and $n - k$ without a ball in them. If there are not enough bags, then there is no possible result.

**Distributing Candy** There are $n$ pieces of identical candy and $k$ kids. How many ways are there to give each kid a piece of candy? If there is enough candy, the answer is one. Everyone gets candy. If there is not enough candy then the answer is zero. There is no way to give everyone candy if there's not enough candy.

### 2.2.3 Distinct Objects, Every Recipient Receives at Least One

#### 2.2.3.1 Distinct Recipients

$$|A| = n!S(k,n) = n! \left\{ \begin{array}{c} k \\ n \end{array} \right\} = n! \frac{1}{n!} \sum_{j=0}^{n} (-1)^{n-j} \binom{n}{j} j^k \tag{2.28}$$

Where $S(k, n)$ denotes the Stirling function of the second kind.

**At Least One Pool Ball into Labeled Buckets** All possible ways to put $k$ pool balls, which all have different numbers and colors, into $n$ labeled buckets, so that all of the buckets have at least one ball in them. This is the same number of combinations as if the buckets were unlabeled, but multiplied with $n!$ ways of applying a label to them.

**Onto Functions** All possible functions $f : x \to y$ with $\{x | x \in A, |A| = k\}$ and $\{y | y \in B, |B| = n\}$ subject to the constraint that there is an element $x$ in the domain so that $f(x) = y$ for each element $y$ of the codomain. That is the functions are onto, or surjective.

#### 2.2.3.2 Indistinct Recipients

The number of ways to divide $k$ distinct objects into $n$ non-empty subets is given by the Stirling number of the second kind (c.f. section 2.1.5):

$$|A| = S(k,n) = \left\{ \begin{array}{c} k \\ n \end{array} \right\} = \frac{1}{n!} \sum_{j=0}^{n} (-1)^{n-j} \binom{n}{j} j^k \tag{2.29}$$

**At Least One Pool Ball into Unlabeled Bags** All possible ways to put $k$ pool balls, which all have different numbers and colors, into $n$ unlabeled bags, so that none of the bags are empty.

## 2.2.4 Distinct Objects, Every Recipient Receives Exactly One

### 2.2.4.1 Distinct Recipients

One of $k$ distinct objects are assigned to each of $n$ distinct recipients. This is the same as assigning the elements of a $k$-tuple from a selection of $n$ without replacement and with the requirement that all of the $n$ are selected.

$$\begin{aligned} &S = \{(a_1, a_2, ..., a_k) | a_i \in A, |A| = k = n, a_i \neq a_j\} \\ &|S| = n! = k! \text{ if } k = n, \text{ 0 otherwise.} \end{aligned} \tag{2.30}$$

**Exactly One Pool Ball into Each Labeled Buckets** All possible ways to put $k$ pool balls, which all have different numbers and colors, into $n$ labeled buckets, that there is one pool ball in each bucket. This is the same as putting one pool ball into unlabeled buckets and multiplying by the $n!$ ways of attaching a label to the buckets. If there isn't the same amount of balls and buckets then there is no possible way for them to be matched one-for-one.

**Bijective Functions** All possible functions $f : x \to y$ with $\{x | x \in A, |A| = k\}$ and $\{y | y \in B, |B| = n\}$ subject to the constraints that $f(a) = f(b)$ implies $a = b$ and that there is an element $x$ in the domain so that $f(x) = y$ for each element $y$ of the codomain. That is, the functions are one-to-one and onto, of bijective.

**Permutations** Since each of the $k$ objects is given to a different one of the $n$ recipients, there must be as many recipients as there are objects and $k = n$. The number of ways of assigning $k$ objects to $n$ recipients is $k! = n!$.

**Unique Identifiers** Each of $k$ entries in a database is given one of $n = k$ unique identifiers, so that each identifier leads to an entry and each entry has an identifier.

### 2.2.4.2 Indistinct Recipients

$$\begin{aligned} &S = \{(a_1, a_2, ..., a_k) | a_i \in A, |A| = k, a_i = a_j\} \\ &|S| = 1 \text{ if } k = n, \text{ 0 otherwise.} \end{aligned} \tag{2.31}$$

**Exactly One Pool Ball into Each Unlabeled Bag** All possible ways to put $k$ pool balls, which all have different numbers and colors, into $n$ unlabeled bags, that there is one pool ball in each bag. The result is that you either have $n$ bags with $n = k$ balls in them, if the two have matching numbers, or that you either have a bag or a ball left over and there is no way to match them one-for-one.

**Distribute without Leftovers** $k$ students are assigned to $n$ identical textbooks. How many ways are there for each child to have a textbook so that there are no textbooks left over?

## 2.2.5 Distinct Objects, Distributed in Ordered Groups

### 2.2.5.1 Distinct Recipients

$k$ objects are distributed to $n$ different recipients with an internal ordering, so that each recipient receives an ordered list. This is the same as creating a list of $n$ sequences that are sampled from $k$ without replacement.

$$\begin{aligned} &S = \{(\mathbf{a}_{\{i_1\}} = (a_{i_{1,1}}, a_{i_{1,1}}, ..., a_{i_{1,l}}), \mathbf{a}_{\{i_2\}}, ..., \mathbf{a}_{\{i_n\}}) | \textstyle\sum_j^n |\mathbf{a}_{i_j}| = k, a_{i_{j,i}} \in A, |A| = n\} \\ &|S| = \tfrac{(k+n-1)!}{(n-1)!} = (k+n-1)^{\underline{k}} = k!\tbinom{k+n-1}{k} \end{aligned} \tag{2.32}$$

**Books on Labeled Bookshelves** $k$ books are distributed across $n$ different bookshelves. The books may all be on the same shelf, or shelves may be empty. The ordering of the books on each of the shelves matters. This is the same as taking all the $k!$ permutations of the $k$ books and multiplying it by the way of dividing that permutation up onto $n$ shelves.

**Ordered Functions** All functions $f : x \to y$ with $\{x|x \in A, |A| = k\}$ and $\{y|y \in B, |B| = n\}$ that assign ordered sequences of elements in $x$ to elements of $y$.

#### 2.2.5.2 Indistinct Recipients

$k$ objects are broken up into $n$

$$|S| = \sum_{i=1}^{n} L(k, i) \tag{2.33}$$

where $L(k, n)$ are the Lah numbers, which describe "How many ways can $k$ objects be distributed to $n$ recipients if order matters and each recipient receives at least 1". They are given by:

$$L(k, n) = \binom{k}{n} (k-1)^{\underline{k-n}} = \frac{k!}{n!} \binom{k-1}{n-1} \tag{2.34}$$

The sum considers the case where all $k$ objects are given to $i \leq n$ recipients, with the remaining recipients receiving none. Note that the expression is more complicated than simply dividing the case for indistinct recipients by $n!$, because more than one of the recipients might receive the same number of objects, in which case i.e. the distribution $[(0), (0), (1, 2)]$ would be counted twice. The expression is analogous to the expression for distinct objects distributed to indistinct recipients, but where the objects did not have an internal ordering. In that case, instead of a sum over the Lah numbers, the sum is over the Stirling numbers of the second kind.

The Lah numbers can be derived by imagining that first one of the $k$ distinct objects is distributed to the $n$ indistinct recipients, to make sure that each recipient receives at least one, and then adding the remaining objects to the first without restrictions. After the first $n$ distinct objects have been distributed to the recipients, the recipients are no longer indistinct, because they each have been labeled by the object they have already received. There are $\binom{k}{n}$ ways of distributing the first $n$ objects and $(k-n)!\binom{k-n+n-1}{k-n}$ ways to add the $k-n$ remaining objects.

**Books into Unlabeled Boxes** $k$ books are stacked into up to $n$ different unlabeled boxes. Some of the boxes may be empty, and others may contain more than one book. The sequence in which the books are stacked in each box matters. If there are $n-r$ empty boxes and $r$ boxes that have at least one book in them, then there are $\binom{k}{r}$ ways of putting the first book in each of the boxes. Then there are $(k-r)!\binom{k-r+r-1}{k-n}$ ways to stack the remaining books on top of those first books.

**Broken Permutations $\leq n$ Parts** The permutations of $k$ distinct elements are ordered sequences of length $k$. If the sequence is cut up into up to $n$ different parts of non-zero length, then what results are *broken permutations.*

**Books into Boxes** $k$ different books are put into $n$ identical boxes. How many ways are there to pack the boxes if you keep track of the order in which the books in each box are stacked?

### 2.2.6 Distinct Objects, Distributed in Ordered Groups of At Least One

#### 2.2.6.1 Distinct Recipients

$$|S| = \frac{k!}{(k-n)!} \frac{(k-1)!}{(n-1)!} = k^{\underline{n}} (k-1)^{\underline{k-n}} \tag{2.35}$$

**Books on Labeled Bookshelves** $k$ books are distributed across $n$ different labeled bookshelves. There is at least one book on each shelf. This is the same as picking $n$ books to go on the shelves first, so that all of the shelves have at least one book on them, and then distributing the remainder without restrictions. There are $k^{\underline{n}}$ ways of picking the first $n$ books for each of the $n$ shelves, and $(k-n+n-1)^{\underline{k-n}}$ ways to add the remaining $k-n$ books.

**Ordered Onto Functions** All functions $f: x \to y$ with $\{x|x \in A, |A| = k\}$ and $\{y|y \in B, |B| = n\}$ that assign ordered sequences of elements in $x$ to elements of $y$, where every element $y \in B$ has an assignment of at least one element.

#### 2.2.6.2 Indistinct Recipients

$k$ elements are divided into $n$ ordered sequences of minimum length 1.

$$L(k,n) = \binom{k}{n}(k-1)^{\underline{k-n}} = \frac{k!}{n!}\binom{k-1}{n-1} \tag{2.36}$$

**Books into Unlabeled Boxes** $k$ books are stacked into up to $n$ different unlabeled boxes. All of the $n$ boxes have at least one book in them. The sequence in which the books are stacked in each box matters. There are $\binom{k}{n}$ ways of putting the first book in each of the boxes. Then there are $(k-n)!\binom{k-n+n-1}{k-n}$ ways to stack the remaining $k-n$ books on top of those first books.

**Broken Permutations $n$ parts** The permutations of $k$ distinct elements are ordered sequences of length $k$. If the sequence is cut up into up to $n$ different parts, then what results are *broken permutations.*

**Books into Boxes** $k$ different books are put into $n$ identical boxes, so that there is at least one book in each box. How many ways are there to pack the boxes if you keep track of the order in which the books in each box are stacked?

### 2.2.7 Identical Objects

#### 2.2.7.1 Distinct Recipients

$$|S| = \binom{k+n-1}{k} \tag{2.37}$$

This coefficient can be easily derives using the "Stars and Bars" concept. A short way of explaining this is as follows: picture putting all the $k$ identical objects in a line (stars). Picture having dividers (bars) to divide the $k$ objects into $n$ groups. For $n$ groups, it is necessary to use $n-1$ dividers. Empty groups result when the dividers sit right next to each other with no object in between them, or if a divider is at the end of a sequence. In total, the objects and the dividers make for a sequence in which $k+n-1$ spots are taken up by either an object or a divider. To pick a particular way of dividing the $k$ objects up, you can either pick the $n-1$ locations of the sequence in which the dividers are located, or, equivalently, the locations in which the objects are located. There are $\binom{k+n-1}{n-1} = \binom{k+n-1}{k}$ to do so.

**Ping Pong Balls into Labeled Buckets** $k$ identical ping pong balls are put into $n$ labeled buckets. Some of the buckets may be empty, and other buckets might have more than one ball in them.

**Multisets** Multisets are sets in which identical elements might show up several times. For example $\{a, a, b, b, b\}$. They can also be described in terms of the multiplicity of their elements. For example $[a:2, b:3, c:0]$. How many multisets can be formed with $k$ different elements of $n$ different classes?

**Integer Sums** How many different configurations of the $n$ integers $\{x_i\}_n$ satisfy $x_1+x_2+...+x_n = k$?

**Bosons in Degenerate States** In how many ways might $k$ Bosons populate $n$ degenerate states?

#### 2.2.7.2 Indistinct Recipients

$$|S| = \sum_{i=1}^{n} P(k, i) \tag{2.38}$$

It turns out that there is no known formula for $P(k, n)$.

**Ping Pong Balls into Unlabeled Bags** $k$ identical ping pong balls are put into $n$ unlabeled bags. Some of the bags might be empty, and other bags might have more than one ball in them.

**Number Partitions** How many ways are there to divide $k$ objects across up to $n$ piles. How many ways are there to divide an integer $k$ into a sum of $n$ integers (including zeros). For example: for $k = 5$, $n = 3$, the partitions are $5 + 0 + 0, 4 + 1 + 0, 3 + 2 + 0, 3 + 1 + 1, 2 + 2 + 1$.

**Unlabeled Multiplicities of Multisets** For multisets of $k$ elements with $n$ different classes, what is the number of possible multiplicities? For example, a multiset of $k = 3$ elements from $n = 2$ classes could have multiplicities $[a : 3, b : 0]$, $[a : 2, b : 1]$, $[a : 1, b : 2]$, $[a : 0, b : 3]$. If we do not care about the labels $a, b$, then the ways that the the $k$ elements migth be distributed are $[3, 0]$ and $[2, 1]$.

**Boxes of Marbles** $k$ marbles are randomly put into $n$ boxes. How many ways are there for the weight to be distributed among the boxes?

### 2.2.8 Identical Objects, Each Receives At Most One

#### 2.2.8.1 Distinct Recipients

$k$ identical objects are distributed across $n$ recipients so that each recipient recieves at most one. That amounts to choosing $k$ out of the $n$ recipients who will receive an object.

$$|S| = \binom{n}{k} \tag{2.39}$$

**At Most One Ping Pong Balls into Labeled Buckets** $k$ identical ping pong balls are distributed across $n$ different buckets, so that $k$ buckets have one ball in them and $n - k$ buckets are empty. This is the same as choosing $k$ out of the $n$ buckets, for which there are $n \times (n-1) \times (n-2) \times \ldots \times (n-k+1)$ choices for the $k$ balls, and correcting for the internal orderings of the $k$ balls by dividing by $k!$ because the balls are identical.

**Subsets** $k$ element subsets of a set of size $n$. The subsets are formed by either choosing the $k$ elements that are included or the $n - k$ elements that are excluded.

**Set Binary Labels** The problem can also be thought of as assigning a binary label to $n$ elements, where there are $k$ times $1$ and $n - k$ times 0.

#### 2.2.8.2 Indistinct Recipients

$$|S| = 1 \text{ if } k \leq n, \; 0 \text{ otherwise} \tag{2.40}$$

**At Most One Ping Pong Ball into Unlabeled Bags** $k$ identical ping pong balls are put into $n$ identical bags. The result is either $k$ bags with ping pong balls and $n - k$ empty bags, or there is no possible result if there are not enough bags (i.e. if $k > n$).

**Boxes** How many ways are there to put $k$ marbles in $n$ boxes, if each box is only big enough for one marble. One, if there are enough boxes, or zero, if there aren't enough boxes.

## 2.2.9 Identical Objects, Each Receives At Least One

### 2.2.9.1 Distinct Recipients

This problem is the same as giving each of $n$ recipients one of $k$ objects, and then distributing the remaining $k - n$ objects arbitrarily.

$$|S| = \binom{k+n-1-k}{k-n} = \binom{n-1}{k-1} \tag{2.41}$$

This can be derived by picturing first distributing one object to each of the $n$ recipients, ensuring that each recipient has at least one, and then distributing the remaining $k - n$ objects arbitrarily. There is only one way to give each recipient one of the identical objects, and then there are $\binom{k-n+n-1}{k-n}$ ways to distribute the remaining $k - n$ objects arbitrarily across the recipients.

**At Least One Ping Pong Ball into Labeled Buckets** $k$ ping pong balls are distributed into $n$ labeled buckets so that there is at least one ping pong ball in each bucket. The labeled buckets may have more than one ping pong ball in them.

**Compositions $n$ Parts** How many ways are there to assign $k$ identical objects to $n$ labeled sets of at least one object?

### 2.2.9.2 Indistinct Recipients

$$|S| = P(k, n) \tag{2.42}$$

It turns out that there is no known formula for $P(k, n)$.

**At Least One Ping Pong Ball in Unlabeled Bags** $k$ identical ping pong balls are distributed across $n$ unlabeled bags, so that none of the bags are empty.

**Partitions in $n$ Parts** How many ways are there to make $n$ piles from $k$ objects.

## 2.2.10 Identical Objects, Each Receives Exactly One

### 2.2.10.1 Distinct Recipients

$$|S| = 1 \text{ if } k = n, \ 0 \text{ otherwise} \tag{2.43}$$

**One Ping Pong Ball into Each Labeled Bucket** $k$ identical ping pong balls are distributed across $n$ labeled buckets so that each bucket has one ping pong ball in it. If $k = n$, then there is one way to put one ball in each bucket. If the numbers do not match up, then there is no way to match them one-for-one.

### 2.2.10.2 Indistinct Recipients

$$|S| = 1 \text{ if } k = n, \ 0 \text{ otherwise} \tag{2.44}$$

**One Ping Pong Ball into Each Unlabeled Bag** $k$ identical ping pong balls are distributed across $n$ identical bags so that each bags has one ping pong ball in it. If $k = n$, then there is one way to put one ball in each bag. If the numbers do not match up, then there is no way to match them one-for-one.

## 2.3 Generating Functions

Generating functions are functions that encode sequences of numbers as the coefficients of power series. One example are the moment generating functions in probability theory, though they are generally extremely useful in combinatorics problems and, almost equivalently, discrete probability problems.

Bogart (2004) approaches the concept in terms of *Picture Functions*. For each element $s \in S$, there is a picture function $P(s)$, so that, for example, the multiset $\{1, 1, 2\}$ can be written as $P(1)^1P(2)$. Collections of combinations can be rewritten in terms of sums and products, which enables factorization and overall easier accounting. Combinations with particular properties can be filtered by looking at exponents.

The picture function enables writing down combinations of elements as an enumerating function $E_P(s)$. For example, the enumerating function for all multisets that include either one or two times some element $a$ and between zero and two times some element $b$ is written:

$$\begin{aligned} E_P(s) &= P(a) + P(a)^2 + P(a)P(b) + P(a)^2P(b) + +P(a)P(b)^2 + P(a)^2P(b)^2 \\ &\left(P(a) + P(a)^2\right)\left(1 + P(b) + P(b)^2\right) \end{aligned} \tag{2.45}$$

Generating functions get much more complicated. Wilf (2013) is a good resource.

### 2.3.1 Example: Binomial Coefficients

Consider the case of a collection of $n$ indistinguishable objects $s$, and write $P(s) = x$. Then the enumerator for selecting any subset of those $n$ objects is given by:

$$E_P(s) = \prod_i^n (x^0 + x^1) = (1 + x)^n = \sum_i^n \binom{n}{i} x^i \tag{2.46}$$

Where each term $(x^0 + x^1)$ corresponds to the two options of either excluding a particular element $(x^0)$ or including a particular element $(x^1)$. The exponent of the expanded product encodes how many objects were included into a particular subset. This is one way of "proving" the binomial coefficients, and one can says $(1_x)^n$ is the generating function for the binomial coefficients $\binom{n}{i}$.

### 2.3.2 Example: Basket of Goods

An apple costs $20c$, a pear costs $25c$ and a banana costs $30c$. How many different fruit baskets can be bought for $100c$?

By replacing the picture function $P(s)$ with $x$, it was possible to identify subsets of $n$ objects by looking at the exponent of $x^n$ in the enumerating function. In this case, the exponent is supposed to show the price. This can be done by writing $P(apple) = x^{20}, P(pear) = x^{25}$ and $P(banana) = x^{30}$.

$$E_P(s) = \left(\sum_{i=0}^{5} x^{20}\right)\left(\sum_{i=0}^{4} x^{25}\right)\left(\sum_{i=0}^{3} x^{30}\right) \tag{2.47}$$

Which results in some power series of the form:

$$E_P(s) = 1x^0 + 1x^{20} + 1x^{25} + 1x^{30} + 1x^{40} + ... + 2x^{60} + ... + 1x^{290} \tag{2.48}$$

To obtain the number of combinations that correspond to a cost of exactly $100c$, one can apply the operator $\frac{1}{n!}\frac{d^n}{dx^n}$ and set $x = 0$ to obtain the desired term. This is what is done with moment generating functions in statistics.

But actually it is easier to think through what the coefficients will be so that:

$$\sum_{l=0}^{n+m+h} d_l x^l = \left(\sum_{i=0}^{n} a_i x^i\right)\left(\sum_{j=0}^{m} b_j x^j\right)\left(\sum_{k=0}^{h} c_k x^k\right) \tag{2.49}$$

$$d_l = \sum_{\substack{i,j,k \\ i+j+k=l}} a_i b_j c_k \tag{2.50}$$

If $a_i = b_j = c_k = 1$, then:

$$d_l = \sum_{\substack{i,j,k \\ i+j+k=l}} 1 = \binom{l+3-1}{l} \tag{2.51}$$

Which is the number of all multisets of size $l$ and 3 classes.

In this case, however, the coefficients are equal to 1 only for $i = a20$, $j = b25$, and $k = c30$ for $a, b, c \in \mathbb{N}_0$, and zero otherwise, so that there are 4 combinations of $a, b, c$ for which $20a + 25b + 30c = 100$. Hence, $d_{100} = 4$ for the basket of goods above.

The sum can also be rewritten:

$$d_l = \sum_{i=0}^{l} \sum_{j=0}^{l-i} a_i b_j c_{l-i-j} \tag{2.52}$$

Of course, this is the discrete version of a convolution. So, it is no surprise that the addition of random variables winds up being a convolution.

### 2.3.3 Example: Dice

How many ways are there for $n$ dice with $k$ faces to show $s$ eyes?

$$E_P = \left( \sum_{i=0}^{\infty} a_i x^i \right)^n = \sum_{i=0}^{\infty} d_i x^i \tag{2.53}$$

Where $a_i = 1 \; \forall i \in (1, k)$ and $a_i = 0$ otherwise.

$$E_P = \left( \sum_{i=1}^{k} x^i \right)^n = \left( x(1-x^k) \sum_{i=0}^{\infty} x^i \right)^n = x^n \left( \frac{1-x^k}{1-x} \right)^n = x^n \left( \sum_{i=0}^{n} (-1)^i \binom{n}{i} x^{ik} \right) \left( \sum_{j=0}^{\infty} (-1)^j \binom{-n}{j} x^j \right) \tag{2.54}$$

The coefficient for $x^s$ is given the sum:

$$d_s = \sum_{ki+j=s-n} (-1)^{i+j} \binom{n}{i} \binom{-n}{j} i \in [0, n] j \in [0, \infty] \tag{2.55}$$

Where the indices $i$ and $j$ satisfy $ki + j = s - n$. For example, for $s = 7$, $n = 2$ and $k = 6$:

$$6i + j = 7 - 2 = 5s \tag{2.56}$$

Holds for $i = 0, j = 5$:

$$d_s = (-1)^5 \binom{2}{0} \binom{-2}{5} = (-1)^{10} 1 \binom{2+5-1}{5} = \binom{6}{5} = \frac{6!}{5!1!} = 6 \tag{2.57}$$

Indeed, there are 6 ways for two d6 to add to 7:

$$[6, 1], [5, 2], [4, 3], [3, 4], [2, 5], [1, 6] \tag{2.58}$$

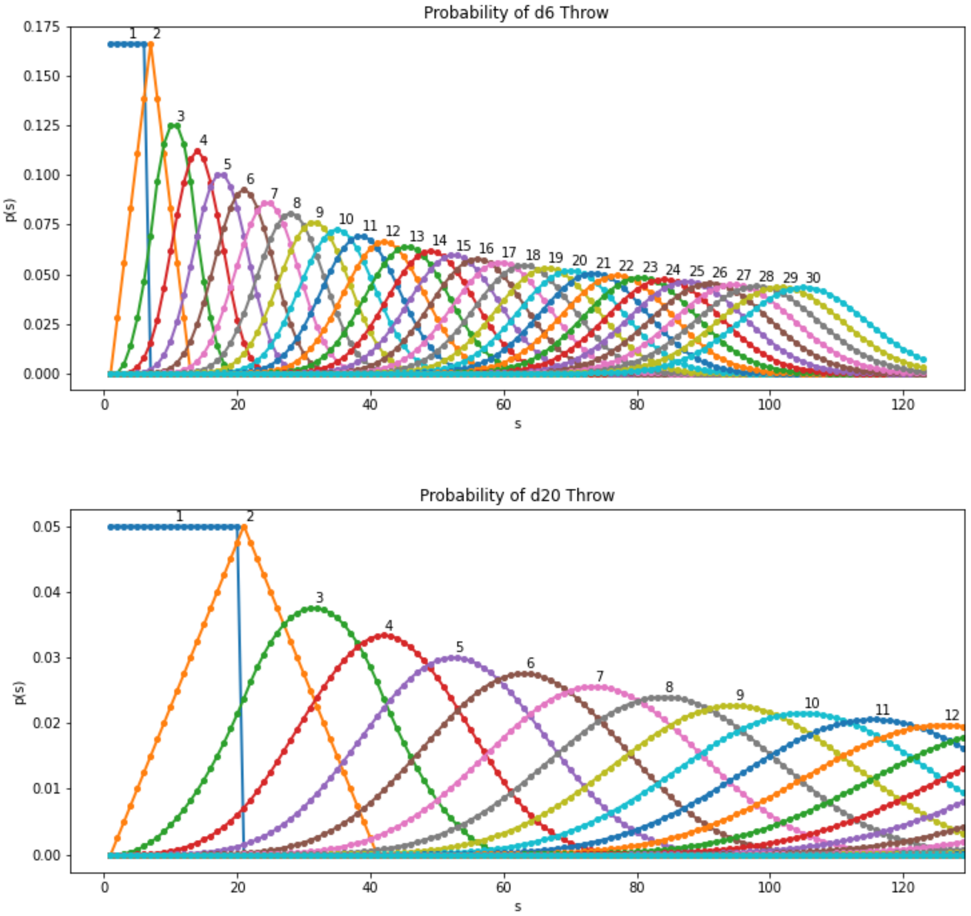


Figure 2.2: Probability of the sum of 6-sided and 20-sided dice, calculated with Eqn. 2.57. Note that these are simply the convolutions of $n = 1, 2, 3, ...$ square waves, which rapidly adopts the shape of a Bell curve.

# 3 Linear Algebra and Multivariable Calculus

**Contents of this chapter**

## 3.1 Multi-Index Notation

Multi-index notation makes high-dimensional things faster and easier. A collection is indices is represented by a tuple $\alpha = (\alpha_1, \alpha_2, \alpha_3, ...)$. The absolute value $|\alpha| = \sum_i \alpha$, partial derivatives $\partial^\alpha = \prod \partial^{\alpha_i}$, powers $\mathbf{x}^\alpha = \prod_i x_i^{\alpha_i}$.

### 3.1.1 Example: Multinomial Coefficients

Instead of:

$$\sum_{0 \leq i_1, i_2, i_3, ..., i_k \leq n} \binom{n}{i_1, i_2, i_3, ..., i_k} \tag{3.1}$$

Write:

$$\sum_{0 \leq |\alpha| \leq n} \binom{n}{\alpha} \tag{3.2}$$

### 3.1.2 Example: Taylor Expansion

For a vector valued function $\mathbf{f} : \mathbb{R}^n \to \mathbb{R}^m$ that is analytical in a neighborhood of the point $\mathbf{a}$:

$$f(\mathbf{x}) = \sum_{|\alpha| \geq 0} \frac{(\mathbf{x} - \mathbf{a})^\alpha}{\alpha!} (\partial^\alpha) f \tag{3.3}$$

## 3.2 Matrix Multiplication

Given two matrices $\mathbf{A} \in \mathbb{R}^{m \times p}$ and $\mathbf{B} \in \mathbb{R}^{p \times n}$ with entries $a_{i,j}$ and $b_{i,j}$ in their $i$th row and $j$th column, the product $\mathbf{C} \in \mathbf{AB} \in \mathbb{R}^{m \times n}$ and with elements $c_{i,j} = \sum_{k=1}^{p} a_{i,k} b_{k,j}$.

There are two ways to interpret how the rows or columns of $\mathbf{C}$ come about through this operation. The first is that a row or column vector in $\mathbf{C}$ is formed by one of the matrices acting on an individual vector from the other matrix. The second is that a row of a column vector in $\mathbf{C}$ is formed as the linear superposition of the vectors within one of the matrices, where the coefficients of the superposition are given by a row or a column within the other matrix.

Let:

$$\mathbf{A} = \begin{bmatrix} | & | & | \\ a_1^{(c)} & a_2^{(c)} & a_3^{(c)} \\ | & | & | \end{bmatrix} = \begin{bmatrix} - & a_1^{(r)} & - \\ - & a_2^{(r)} & - \\ - & a_3^{(r)} & - \end{bmatrix} \tag{3.4}$$

$$\mathbf{B} = \begin{bmatrix} | & | & | \\ b_1^{(c)} & b_2^{(c)} & b_3^{(c)} \\ | & | & | \end{bmatrix} = \begin{bmatrix} - & b_1^{(r)} & - \\ - & b_2^{(r)} & - \\ - & b_3^{(r)} & - \end{bmatrix} \tag{3.5}$$

$$\mathbf{C} = \begin{bmatrix} | & | & | \\ c_1^{(c)} & c_2^{(c)} & c_3^{(c)} \\ | & | & | \end{bmatrix} = \begin{bmatrix} - & c_1^{(r)} & - \\ - & c_2^{(r)} & - \\ - & c_3^{(r)} & - \end{bmatrix} \tag{3.6}$$

be the notation in terms of row and column vectors.

#### 3.2.0.1 Elementwise

$[\mathbf{C}]_{i,j}$ is the dot product of the $i$th row vector of $\mathbf{A}$ and the $j$ column vector of $\mathbf{B}$.

$$c_{i,j} = \begin{bmatrix} | \\ a_i^{(r)} \\ | \end{bmatrix} \cdot \begin{bmatrix} - & b_j^{(c)} & - \end{bmatrix} \tag{3.7}$$

#### 3.2.0.2 Columns of C

The $j$th column of $\mathbf{C}$ involves all of $\mathbf{A}$ and only the $j$th column of $\mathbf{B}$.

The $j$th column of $\mathbf{C}$ is the matrix product of $\mathbf{A}$ with the first column of $\mathbf{B}$.

$$\begin{bmatrix} | \\ c_j^{(c)} \\ | \end{bmatrix} = \begin{bmatrix} - & a_1^{(r)} & - \\ - & a_2^{(r)} & - \\ - & a_3^{(r)} & - \end{bmatrix} \begin{bmatrix} | \\ b_j^{(c)} \\ | \end{bmatrix} \tag{3.8}$$

The $j$th column of the product $\mathbf{C}$ is a linear superposition of the columns of $\mathbf{A}$, with the coefficients given by the $j$th column of $\mathbf{B}$.

$$\begin{bmatrix} | \\ c_j^{(c)} \\ | \end{bmatrix} = \begin{bmatrix} | \\ a_1^{(c)} \\ | \end{bmatrix} b_{1,j}^{(c)} + \begin{bmatrix} | \\ a_2^{(c)} \\ | \end{bmatrix} b_{2,j}^{(c)} + \begin{bmatrix} | \\ a_3^{(c)} \\ | \end{bmatrix} b_{3,j}^{(c)} \tag{3.9}$$

#### 3.2.0.3 Rows of C

The $i$th row of $\mathbf{C}$ involves all of $\mathbf{B}$ and only the $i$th row of $\mathbf{A}$.

The $i$th row of $\mathbf{C}$ is the matrix product of $\mathbf{B}$ acting backwards on the $i$th row of $\mathbf{A}$.

$$\begin{bmatrix} - & c_i^{(r)} & - \end{bmatrix} = \begin{bmatrix} - & a_i^{(r)} & - \end{bmatrix} \begin{bmatrix} | & | & | \\ b_1^{(c)} & b_2^{(c)} & b_3^{(c)} \\ | & | & | \end{bmatrix} \tag{3.10}$$

The $i$th row of $\mathbf{C}$ is a linear superposition of the rows of $\mathbf{B}$, with the coefficients given by the $i$th row of $\mathbf{A}$.

$$\begin{aligned} \begin{bmatrix} - & c_i^{(r)} & - \end{bmatrix} = \; & a_{i,1}^{(r)} \begin{bmatrix} - & b_1^{(r)} & - \end{bmatrix} \\ & + a_{i,2}^{(r)} \begin{bmatrix} - & b_2^{(r)} & - \end{bmatrix} \\ & + a_{i,3}^{(r)} \begin{bmatrix} - & b_3^{(r)} & - \end{bmatrix} \end{aligned} \tag{3.11}$$

## 3.3 Linear Systems of Equations

(Real numbers only this time.)

Linear equations are of the form $Ax = b$ where $A$ is a matrix and $x$ and $b$ are vectors. The rows of $A$ and $b$ form a system of equations that must be simultaneously satisfied by the entries of $x$. If $x, b \in \mathbb{R}^n$, then the solutions to the equation of a single row corresponds to an $n-1$-dimensional hyperplane. If the rows of $A$ are linearly independent, then solutions that simultaneously satisfy the equations in $k$-rows correspond to the $n-k$-dimensional intersection of $k$ $n$-dimensional hyperplanes. The solution of $x$ that satisfies all $n$ equations is a $n-n=0$-dimensional point, and so $x$ is uniquely determined. If any two rows of $A$ are not linearly independent, then the hyperplanes that correspond to values of $x$ that satisfy them overlap exactly, and their intersection is $n$ dimensional, rather than $n-1$ dimensional. In this case, the value of $x$ that satisfies all rows of $A$ is not narrowed down to a single point. The system of equations is said to be *underdetermined*:. This is equivalently the case when $A$ has $m < n$ rows.

### 3.3.1 $A \in \mathbb{R}^{n \times n}$ Square Matrices

If $A \in \mathbb{R}^{nxn}$, then the solution to the system is formally $x = A^{-1}b$, where $A^{-1}$ is the matrix inverse, satisfying $A^{-1}A = I$, where $I$ is the identity matrix.

Since $Ax = b$ is the same as expressing $b$ in terms of a linear combination of the columns of $A$, the entries of $x$ can be interpreted as the coefficients resulting from the projection of $b$ into the column space of $A$. Therefore, for an orthonormal matrix, the $A^{-1}$ is simply $A^T$

### 3.3.2 $A \in \mathbb{R}^{m \times n}$ Rectangular Matrices, Overdetermined Case

If $A \in \mathbb{R}^{mxn}$ with $m > n$ rows, then there need not be any point $x \in \mathbb{R}^n$ in which the $m$ hyperplanes all intersect. In that case, the system does not have a solution $x \in \mathbb{R}^n$, and the system is considered *overdetermined*. (The intersection of $m$ distinct hyperplanes in $n$ dimensional space would have negative dimension $(n-m) < 0$ if $m > n$, which my feeble brain can't make sense of.)

In the overdetermined case $A^{mxn}$ with $m > n$, the columns of $A$ do not span $\mathbb{R}^m$ and therefore $b \in \mathbb{R}^m$ may have some component $\epsilon$ that lies outside of the column space of $A$. In that case, no linear combination $x$ of the columns of $A$ can express $b$ perfectly, but we might look for approximate solutions $\hat{x}$ so that:

$$A\hat{x} + \epsilon = b \tag{3.12}$$

So that the error $||\epsilon||_\alpha$ is minimized. This is the starting point for linear regression from the linear algebra perspective. In practice, the approximation is usually approximated by applying an iterative gradient descent algorithm to minimize the *loss function* $||\epsilon||_\alpha$. The choice of metric $||\cdot||_\alpha$ is essentially a design choice. For $\alpha = 2$, the metric is the $L^2$ norm (cf. section 3.12.2) and an analytic solution exists, named the *normal equations*. The solution minimizes the least squares error and corresponds to the projection of $\mathbf{b}$ into the column space of $a$. The procedure is better known as ordinary least squares regression and therefore I will move a more elaborate discussion to chapter 8.

### 3.3.3 $A \in \mathbb{R}^{n \times m}$ Rectangular Matrices, Underdetermined Case

For an underdetermined system with $m < n$, there is either no solution (if the $m$ hyperplanes don't intersect), or there are infinitely many possible solutions that lie on an $n-m$ dimensional hyperplane. One idea is to pick the solution that minimizes $||\hat{x}||_2$ based on the idea that it might generalize better. The least norm solution is $\hat{x} = A^T \left(AAT\right)^{-1} b$, which is the projection of $\vec{0}$ on the solution set.

## 3.4 $\mathbf{A} = \mathbf{L}\mathbf{L}^\dagger$ Cholesky Decomposition

The Cholesky Decomposition exists when a matrix is hermitian and positive-definite. It expresses the matrix $\mathbf{A}$ as:

$$\mathbf{A} = \mathbf{L}\mathbf{L}^\dagger \tag{3.13}$$

Where $\mathbf{L}$ is a lower-triangular matrix with positive, real diagonal entries. When $\mathbf{A}$ is real, then so is $\mathbf{L}$. The Cholesky decomposition enables fast solution of a linear system, but it can also be used to create correlated random variables in Monte Carlo simulations.

### 3.4.1 Creating Correlated Random Variables

Let $\mathbf{u}_t$ be a vector of uncorrelated samples with mean 0 and standard deviation 1. If the covariance matrix of the system to be simulated is $\mathbf{\Sigma}$ with Cholesky decomposition $\mathbf{\Sigma} = \mathbf{L}\mathbf{L}^\dagger$, then the vector $\mathbf{v}_t = \mathbf{L}\mathbf{u}_t$ has the desired covariance.

## 3.5 Generalized Eigenvectors

## 3.6 $\mathbf{A} = \mathbf{V}\mathbf{\Lambda}\mathbf{V}^{-1}$ Spectral Theorems, Diagonalization

Spectral theorems deal with diagonalizable linear operators.

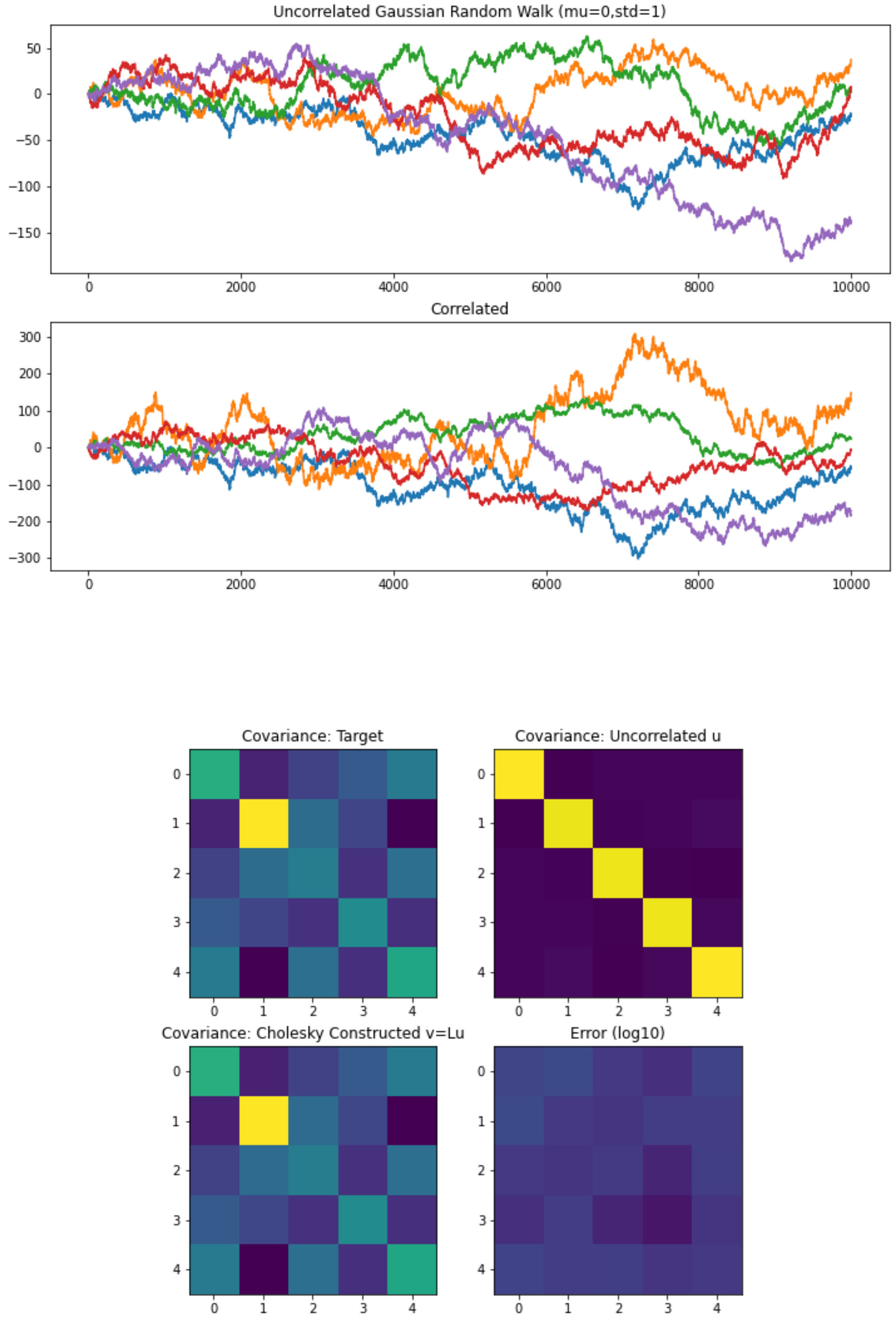


Figure 3.1: Creating correlated random variables from uncorrelated random variables using the Cholesky decomposition of the covariance matrix. The 5 uncorrelated random variables are sampled from a standard normal distribution. It is difficult to see a difference between the correlated and uncorrelated random walks.

A diagonalization of a matrix $\mathbf{A}$ is always possible when a matrix is square, and refers to a decomposition of the matrix into the matrix of eigenvectors $\mathbf{V}$ and eigenvalues $\mathbf{\Lambda}$ as

$$\mathbf{A} = \mathbf{V\Lambda V^{-1}} \tag{3.14}$$

### 3.6.1 $\mathbf{A} = \mathbf{V\Lambda V}^T$ Eigendecomposition of Symmetric Matrices

A hermitian matrix $\mathbf{A}$ has orthogonal eigenvectors, which means that $\mathbf{V}$ is unitary, meaning that $\mathbf{V}^{-1} = \mathbf{V}^\dagger$. In that case, the diagonalization is:

$$\mathbf{A} = \mathbf{V\Lambda V}^\dagger = \begin{bmatrix} | & | & \cdots & | \\ v_1 & v_2 & \ddots & v_n \\ | & | & \cdots & | \end{bmatrix} \begin{bmatrix} \lambda_1 & 0 & \ldots & 0 \\ 0 & \lambda_2 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & \ldots & \ldots & \lambda_n \end{bmatrix} \begin{bmatrix} \text{———} & v_1^\dagger & \text{———} \\ \text{———} & v_2^\dagger & \text{———} \\ \vdots & \ddots & \vdots \\ \text{———} & v_n^\dagger & \text{———} \end{bmatrix} \tag{3.15}$$

Which is the same as saying that all hermitian matrices are *similar* to a diagonal matrix (cf. section 3.10.11, two matrices $\mathbf{A}$ and $\mathbf{B}$ are similar if they are transmutable using unitary transformations as $\mathbf{A} = \mathbf{UBU}^\dagger$). The diagonal representation also shows that in order for $\mathbf{A}$ to satisfy the hermitian property $\mathbf{A}^\dagger = \mathbf{A}$ its eigenvalues $\lambda_i$ must be real. Further, it means that if $\mathbf{A}$ is hermitian, $\mathbf{A}$ can be written in terms of projections on the eigenvectors:

$$\mathbf{A} = \sum_i \lambda_i (v_i \otimes v_i) \tag{3.16}$$

Where $\otimes$ is the (complex) outer product $v_i \otimes v_i = v_i v_i^\dagger$. I have seen the existence of this representation of a hermitian matrix be described as synonmous with *spectral theorem.* Since the eigenvectors $\mathbf{V}$ are an orthonormal basis, $\sum_i v_i \otimes v_i = \mathbb{I}$.

### 3.6.2 $\mathbf{H} = \mathbf{U\Lambda U}^T$ Eigendecomposition of Hermitian Matrices

Similarly, a hermitian matrix $\mathbf{H} \in \mathbb{C}^{n\times n}$ (the complex equivalent to a symmetric matrix) has real eigenvalues and the matrix of eigenvectors is unitary, so that $\mathbf{H} = \mathbf{U\Lambda U}^T$.

### 3.6.3 Eigenvalue Sensitivity and Accuracy

#### 3.6.3.1 General Case

In general, the values of the eigenvalues of a square matrix $\mathbf{A}$ may vary wildly under a slight cange of $\mathbf{A} \to \mathbf{A} + \delta\mathbf{A}$. The sensitivity of the eigenvalues to a change in $\mathbf{A}$ can be investigated using matrix norms (Mathworks n/a). Let $||\cdot||$ denote a submultiplicative matrix norm, then:

$$\begin{aligned} \Lambda + \delta\Lambda &= \mathbf{X^{-1}}\left(\mathbf{A} + \delta\mathbf{A}\right)\mathbf{X} \\ \delta\Lambda &= \mathbf{X^{-1}}\delta\mathbf{AX} \\ ||\delta\Lambda|| &= ||\mathbf{X^{-1}}\delta\mathbf{AX}|| \le ||\mathbf{X^{-1}}||||\mathbf{X}||||\delta\mathbf{A}|| \end{aligned} \tag{3.17}$$

When $||\cdot||$ is chosen to be the operator norm with respect to $L^2$, $||\cdot||_{(2)}||$, then $||\mathbf{X^{-1}}|| = \sigma_1$ and $||\mathbf{X}|| = \frac{1}{\sigma_n}$ where $\sigma_1$ and $\sigma_2$ are the square roots of the largest and the smallest eigenvalue of $\mathbf{X}^\dagger\mathbf{X}$ respectively (cf. section 3.13 on matrix norms). In that case, the sensitivity of the eigenvalues to a change in $\mathbf{A}$ is:

$$||\delta\mathbf{\Lambda}||_{(2)} \le \frac{\sigma_1}{\sigma_n}||\delta\mathbf{A}||_{(2)} = \kappa(\mathbf{X})||\delta\mathbf{A}||_{(2)} \tag{3.18}$$

Where $\kappa(\mathbf{X})$ is the conditioning number of the matrix $\mathbf{X}$. Upper bounds on the error on individual eigenvalues can also be derived quite easily, which is shown in **??** pp 10-12.

#### 3.6.3.2 Hermitian Matrices

For hermitian (or orthogonal) matrices, the conditioning number for the individual eigenvalues $\kappa(\lambda_i, \mathbf{H}) = 1$, so that the error on an individual eigenvalue $||\lambda_i||_{(2)} \leq \kappa(\lambda_i, \mathbf{H})||\mathbf{H}||_{(2)} = 1 \times ||\mathbf{H}||_{(2)}$.

## 3.7 $\mathbf{A} = \mathbf{U\Sigma V}^\dagger$ Singular Value Decomposition

The Singular Value Decomposition (SVD) exists for <u>any</u> matrix $\mathbf{A} \in \mathbb{C}^{m \times n}$, and is a closely related alternative to the eigendecomposition (cf. section 3.6) that works for non-square matrices. The contrast to diagonalization is that the eigenvectors and eigenvalues of $\mathbf{A}^\dagger\mathbf{A}$ and $\mathbf{A}\mathbf{A}^\dagger$ found, rather than to look for the eigensystem of $\mathbf{A}$ itself. The advantage is that $\mathbf{A}^\dagger\mathbf{A}$ and $\mathbf{A}\mathbf{A}^\dagger$ always have the convenient properties of being square, hermitian and positive semidefinite.

SVD comes up incessantly in the context of data analysis. In general, the decomposition has the form:

$$\mathbf{A} = \mathbf{U\Sigma V}^\dagger \tag{3.19}$$

Where $\mathbf{U}$ and $\mathbf{V}$ are unitary, and $\Sigma$ is a diagonal matrix with real and positive entries $\sigma_i^2$ along the diagonal, so that $\sigma_1 \geq \sigma_2^2 \geq ... \geq \sigma_n^2$. $\mathbf{U}$ is the matrix of left singular vectors, which are the eigenvectors of $\mathbf{A}\mathbf{A}^\dagger$. $\mathbf{V}$ is the matrix of right singular vectors, which are the eigenvectors of $\mathbf{A}^\dagger\mathbf{A}$. The singular values are the square roots of the eigenvalues of $\mathbf{A}^\dagger\mathbf{A}$ or, equivalently, $\mathbf{A}\mathbf{A}^\dagger$. If the $\mathbf{A}$ happens to be square and symmetric ($\mathbf{A} = \mathbf{A}^\mathbf{T}$, cf. section 3.10.5), then the singular values are simply the absolute values of the eigenvalues of $\mathbf{A}$. The right and left singular vectors of the matrix are orthonormal bases that respectively span the column and row space of $\mathbf{A}$. The number of singular values of $\mathbf{A}$ is the rank of $\mathbf{A}$.
The singular values are always real and positive or zero, because $\mathbf{A}^\dagger\mathbf{A}$ and $\mathbf{A}\mathbf{A}^\dagger$ are hermitian and positive semidefinite (cf. sections 3.10.5, 3.10.1). When the matrix $\mathbf{A}$ has purely real entries, then the left and right singular vectors are also real, which is not necessarily the case for complex $\mathbf{A}$.

The SVD of a matrix is unique up to the sign columns in $\mathbf{U}$ and $\mathbf{V}$. That is, the SVD is valid under transforming $\mathbf{u}_i, \mathbf{v}_i \rightarrow -\mathbf{u}_i, -\mathbf{v}_i$ if $\mathbf{u}_i$ and $\mathbf{v}_i$ are the $i$th vectors in $\mathbf{U}$ and $\mathbf{V}$. This has the consequence that the basis of singular vectors that is found using SVD does not remain consistently oriented in the presence or noise or the slow evolution of a system. This is addressed in greater length in section 10.5 on performing SVD specifically on data matrices.

### 3.7.1 Full and Economy SVDs

While these properties are always true, unfortunately people use a range of conventions when it comes to the size of $\mathbf{U}$, $\mathbf{\Sigma}$ and $\mathbf{V}$.

The first convention is for $\mathbf{U}$ and $\mathbf{V}$ to be square, in which case they contain the full set of left and right singular vectors, and $\Sigma$ has dimension $m \times n$, with $0$ entries in rows $i > n$. This is known as the *full SVD*.

Friedman et al. (2001) uses the the convention where $\Sigma$ is square, i.e. $\mathbf{U}:\ m \times n$, $\mathbf{\Sigma}:\ n \times n$ and $\mathbf{V}:\ n \times n$. This means that $U$ does not contain the full set of $m$ left singular vectors. Note that, in this case, $\mathbf{U}^\dagger\mathbf{U} = \mathbf{I}$ but $\mathbf{U}\mathbf{U}^\dagger \neq \mathbf{I}$. Friedman et al.'s (2001) convention is of advantage in the context of data analysis, where the data matrix tends to be "tall and skinny" (i.e. $m >> n$), and only the first $n$ left singular vectors are relevant. Also, a letting $\Sigma$ be a square matrix significantly simplifies calculations. Friedman et al.'s (2001) is known as the *economy* or *compact* SVD.

The two different layouts are illustrated in Figure 3.2, which I brazenly copied from **?**.

### 3.7.2 Matrix Approximation

The SVD enables the decomposition of the matrix into a sum:

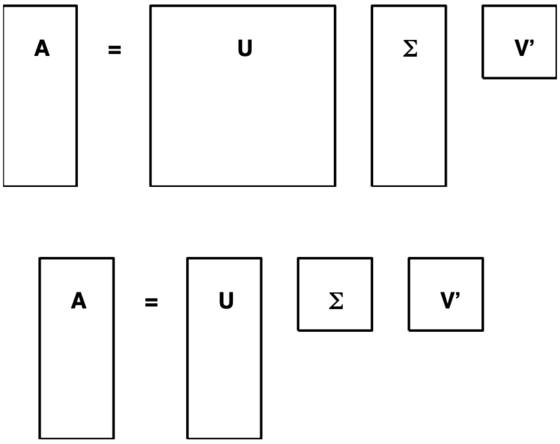


Figure 3.2: Dimensions for Full vs. Economy SVDs

$$\mathbf{A} = \sum_{i=1}^{n} \sigma_i (\mathbf{u}_i \otimes \mathbf{v}_i) \tag{3.20}$$

Which may be truncated at some rank $r$, resulting in a rank-$r$ approximation to $\mathbf{A}$. This is known as the *truncated singular value estimator* (TSVD):

$$\hat{\mathbf{A}} = \sum_{i=1}^{r} \sigma_i (\mathbf{u}_i \otimes \mathbf{v}_i) \tag{3.21}$$

Where $\mathbf{u}_i$ is the $i$th left singular vector and $\mathbf{v}_i$ is the $i$th right singular vector. Following the Eckart-Young-Mirsky Theorem, taking the first $r \leq n$ terms of this series is the best rank-$r$ approximation to $\mathbf{A}$ under the Frobenius norm of the error, i.e. $\underset{\text{rank}(\mathbf{A}^{(r)})=r\leq n}{\arg\min} ||\mathbf{A}^{\mathbf{r}} - \mathbf{A}||_F$. The Frobenius norm essentially measures the elementwise mean squares error (MSE), cf. section 3.13.3).

In the context of data analysis, the reduced rank representation of a data matrix $\mathbf{A}$ acts as a filter in which the components that explain less of the variance in the dataset are removed, with the hope being that these correspond to noise. The intuition is that noise probably originates from a random process, so that components that correspond to noise have weak correlation with the entries of the dataset and therefore small singular values. On the other hand, the signal probably systematic, so that the components that correspond to the signal have larger singular value. Since TSVD minimizes the average mean squares error (AMSE) amounts to the rank-$r$ maximum likelihood estimator of the signal under the assumption of normally distributed noise with mean zero.

By virtue of being expressible in terms of fewer base vectors, a lower rank approximation is also a very effective compression method. For a more elaborate discussion of SVD in data analysis, see section 10.5. In a practical context, the question of selecting the appropriate rank of the approximation emerges. Section 10.8 covers a few approaches to this.

### 3.7.3 Geometric Interpretation of SVD

A common intuitive interpretation of $\mathbf{U}$, $\mathbf{\Sigma}$ and $\mathbf{V}$ is to see them as a decomposition of the action of the matrix $\mathbf{A} \in \mathbb{R}^{m \times n}$ into a rotation, a stretching and another rotation. Let $\{\mathbf{x} : |\mathbf{x}| = 1\}$ be the points on the surface of a unit sphere, then $\mathbf{A}\mathbf{x}$:

- $\mathbf{V}$ rotates the unit sphere, which has no effect. However, it expresses each $\mathbf{x}$ in terms of a different coordinate system (the right singular basis of $\mathbf{A}$).

- $\mathbf{\Sigma}$ stretches the unit sphere into an ellipse with principal axes aligned to that coordinate system.
- $\mathbf{U}$ rotates the ellipse.

In the above, "rotation" might include coordinate flips that are undone by $\mathbf{U}$ following $\mathbf{V}$. Alternatively, for a vector $\mathbf{v} \in \mathbf{R}^n$, the operation $\mathbf{Av}$:

- $\mathbf{V}$ is a change of basis matrix that expresses $\mathbf{x}$ in terms of a right eigenbasis of $\mathbf{A}$
- $\mathbf{\Sigma}$ performs a stretching in that basis
- $\mathbf{U}$ is a change of basis from the right eigenspace into the left eigenspace of $\mathbf{A}$.

## 3.8 Schur Decomposition

For any square matrix $\mathbf{A}$, there exist a unitary matrix $\mathbf{H}$ so that:

$$\mathbf{T} = \mathbf{U}^{\dagger}\mathbf{A}\mathbf{U} \tag{3.22}$$

Where $\mathbf{T}$ is an upper triangular matrix. This concept is the same as saying that every square matrix is similar to an upper triangular matrix.

## 3.9 Types of Transformations

### 3.9.1 Similarity Transformations

If $\mathbf{T}$ is a nonsingular matrix, then a similarity transformation is defined as:

$$\mathbf{A} = \mathbf{T}\mathbf{B}\mathbf{T}^{-1} \tag{3.23}$$

And $\mathbf{A}$ and $\mathbf{B}$ are said to be *similar*.

### 3.9.2 Affine Transformations

Affine transformations are the combination of a linear map and a translation, which has the form $f(\mathbf{x}) = \mathbf{Ax} + \mathbf{b}$.

$$f : V \to W \tag{3.24}$$

Where $V$ and $W$ are vector spaces. Affine transformations can be expresses as matrices by adding an entry with a constant to the vectors that describe a point in space. For example, for $\mathbf{x} \in \mathbb{R}^n$, the affine transform $f(\mathbf{x}) = \mathbf{Ax} + \mathbf{b}$ with $A \in \mathbb{R}^{n,n}$ and $x, b \in \mathbb{R}^n$ can be expressed as the product of a rectangular matrix $\mathbf{M}$ and a vector $\mathbf{c}$ as:

$$\mathbf{Ax} + \mathbf{b} = \underbrace{\left[\begin{array}{c|c} \mathbf{A} & \mathbf{b} \end{array}\right]}_{\mathbf{M}} \underbrace{\begin{bmatrix} \mathbf{x} \\ 1 \end{bmatrix}}_{\mathbf{c}} \tag{3.25}$$

Where $\mathbf{c}^T = [x_1, x_2, x_3, ..., x_n, 1]$ and $\mathbf{M} \in \mathbb{R}^{n,n+1}$.

### 3.9.3 Unitary Transformations

Unitary transformations are transformations that preserve the inner product, i.e. $\hat{U}x \cdot \hat{U}y = x \cdot y$. As linear transformations, they are represented by unitary matrices (cf. section 3.10.10). Unitary transformations include translations, reflections and rotations.

### 3.9.4 Multilinear Maps

A multilinear map acts on several vectors in a way that is linear in each of its arguments. A $k$-linear map acts on $k$ vectors, where $k = 2$ are bilinear maps and $k = 1$ are linear maps.

$$f : V_1 \times V_2 \times ... \times V_n \to W \tag{3.26}$$

Where $V_1, V_2, ..., V_n$ and $W$ are vector spaces. An example would be the addition or subtraction of two or more vectors.

### 3.9.5 Multilinear Forms

Multilinear forms are multilinear maps that have a scalar output. An example is the dot product between two vectors, or summing over the elements of one or more vectors.

$$f : V_1 \times V_2 \times ... \times V_n \to K \tag{3.27}$$

Where $V_1, V_2, ..., V_n$ and $K$ is a scalar field.

## 3.10 Types of Matrices, Matrix Properties

### 3.10.1 $\operatorname{sgn}\left(\mathbf{x}^\dagger \mathbf{H}\mathbf{x}\right)$ Definite

A hermitian matrix $\mathbf{H} \in \mathbb{C}^n$ is positive definite, if for any non-zero column vector $\mathbf{x} \in \mathbb{C}^n$, the quadratic form $\mathbf{x}^\dagger\mathbf{H}\mathbf{x} > 0$, and negative definite if $\mathbf{x}^\dagger\mathbf{H}\mathbf{x} < 0$. The matrix is positive or negative *semidefinite* if $\mathbf{x}^\dagger\mathbf{H}\mathbf{x} \geq 0$ or $\mathbf{x}^\dagger\mathbf{H}\mathbf{x} \leq 0$, respectively. Definiteness plays a role in investigating the convexity of a function by looking at the Hessian (section 3.14.5). Sometimes notation with curly comparison symbols are used. The relationship between definiteness and eigenvalues is intuitive:

| | | |
|---|---|---|
| $\mathbf{H} \succ 0$ | positive definite | all eigenvalues are positive |
| $\mathbf{H} \prec 0$ | negative definite | all eigenvalues are negative |
| $\mathbf{H} \succeq 0$ | positive semidefinite | all eigenvalues are positive or 0 |
| $\mathbf{H} \preceq 0$ | negative semidefinite | all eigenvalues are negative or 0 |

The curly comparison symbols can mean other stuff though, for example in the context of partially ordered sets (order theory) or comparisons between multidimensional arrays.

### 3.10.2 Triangular

A lower triangular matrix is a matrix that has all-zero entries above the diagonal.

$$\mathbf{L} = \begin{bmatrix} l_{1,1} & & & & & 0 \\ l_{2,1} & l_{2,2} & & & & \\ l_{3,1} & l_{3,2} & \ddots & & & \\ \vdots & \vdots & \ddots & \ddots & & \\ \vdots & \vdots & & \ddots & \ddots & \\ l_{n,1} & l_{n,2} & \dots & \dots & l_{n,n-1} & l_{n,n} \end{bmatrix} \tag{3.28}$$

Upper triangular matrices are matrices that have all-zero entries below the diagonal.

### 3.10.3 $\mathbf{AB} - \mathbf{BA} = 0$ Commuting

Two matrices commute if $\mathbf{AB} = \mathbf{BA}$, or, equivalently, their *commutator* $[\mathbf{A}, \mathbf{B}] = \mathbf{AB} - \mathbf{BA}$ is zero. This means that $\mathbf{A}$ and $\mathbf{B}$ both have to be square. Matrices commute when they have the same eigenspace, i.e.

they have the same eigenvectors. This can be seen by considering the diagonal representations of $\mathbf{A}$ and $\mathbf{B}$.

Let $\mathbf{A}$ and $\mathbf{B}$ be two square matrices with the same eigenvectors $\mathbf{V}$, then they can be diagonalized as:

$$\mathbf{A} = \mathbf{V}\mathbf{\Lambda_A}\mathbf{V}^{-1}\mathbf{B} = \mathbf{V}\mathbf{\Lambda_B}\mathbf{V}^{-1} \tag{3.29}$$

They commute, because:

$$\mathbf{AB} = \mathbf{V}\mathbf{\Lambda_A}\mathbf{V}^{-1}\mathbf{V}\mathbf{\Lambda_B}\mathbf{V}^{-1} = \mathbf{V}\mathbf{\Lambda_B}\mathbf{\Lambda_A}\mathbf{V}^{-1} = \mathbf{V}\mathbf{\Lambda_B}\mathbf{V}^{-1}\mathbf{V}\mathbf{\Lambda_A}\mathbf{V}^{-1} = \mathbf{BA} \tag{3.30}$$

### 3.10.4 $\mathrm{AB} + \mathrm{BA} = 0$ Anticommuting

Two matrices anticommute if $\mathbf{AB} = -\mathbf{BA}$, or, equivalently, their *anticommutator* $\{\mathbf{A}, \mathbf{B}\} = \mathbf{AB} + \mathbf{BA}$ is zero.

### 3.10.5 $\mathrm{A}^\dagger = \mathrm{A}$ Hermitian, Symmetric

Hermitian matrices are matrices that are equal to their complex transpose. That is:

$$\mathbf{A}^{*T} = \mathbf{A}^\dagger = \mathbf{A} \tag{3.31}$$

#### 3.10.5.1 Properties

(There are more)

- By definition: $\mathbf{A} = \mathbf{A}^\dagger$
- Diagonal Entries are all real, since $a_{i,i} = a^*_{i,i}$, but not necessarily positive (physics will mislead you there...)
- Inverse is also hermitian: $\mathbf{A}^{-1} = \mathbf{A}^{-1\dagger}$
- Diagonalizable with real eigenvalues and orthogonal eigenvectors $\in \mathbb{C}^n$.

Hermitian matrices with only real entries are called symmetric matrices. In that case $\mathbf{A}^T = \mathbf{A}$. Hermitian matrices can only have real elements along their diagonal.

### 3.10.6 $\mathrm{A}^\dagger = -\mathrm{A}$ Skew Hermitian, Skew Symmetric

Skew Hermitian matrices that are equal to the negative of their complex transpose. That is:

$$\mathbf{A}^{*T} = \mathbf{A}^\dagger = -\mathbf{A} \tag{3.32}$$

Real matrices that are skew Hermitian are called skew symmetric. In that case:

$$\mathbf{A}^T = -\mathbf{A} \tag{3.33}$$

Skew Hermitian matrices can only have complex values on their diagonal, and skew symmetric matrices can only have zeros as diagonal elements.

### 3.10.7 $\mathrm{AA} = \mathrm{I}$ Involutory

Involutory matrixes are matrices that are their own inverse, so that:

$$\mathbf{AA} = \mathbf{I} \tag{3.34}$$

Involutory matrices are all square roots of the identity matrix. A famous example are the $2 \times 2$ Pauli matrices.

### 3.10.8 $||\mathbf{Ax}||_\alpha = ||\mathbf{x}||_\alpha$ Isometric

An isometric transformation with respect to some norm $||\cdot||_\alpha$ preserves that norm (it's in the name: iso-metric). The linear case is represented by isometric matrices, which satisfy:

$$||\mathbf{Ax}||_\alpha = ||\mathbf{x}||_\alpha \tag{3.35}$$

For some vector $\mathbf{x}$. Isometries are also known as distance-preserving maps. To see this, define the distance between two points $\mathbf{a}$ and $\mathbf{b}$ as $||\mathbf{a}-\mathbf{b}||_\alpha = ||\mathbf{x}||_\alpha$. Isometries are usually understood be bijective.

In terms of operator norms, isometries must satisfy $||\mathbf{A}||_{(\alpha)} = 1$ where $||\cdot||_{(\alpha)}$ denotes the operator norm (cf. section 3.13.1). However, operator norms give the upper bound on the distortion of the input, so a unit operator norm is necessary but not sufficient.

#### 3.10.8.1 Isometries with Respect to $L^2$

The $L^2$ norm is unique in that the points $\{\mathbf{x} : ||\mathbf{x}||_2 = 1\}$ lie on a perfect circle, which remains a circle regardless of the orientation of the underlying coordinate system. (A sphere looks the same regardless of angle.) This symmetry is broken for all other norms. Unitary matrices are isometries with respect to $L^2$, though unitarity is not a necessary condition for an isometry.

#### 3.10.8.2 Isometries with Respect to $L^1$ and general $L^q \neq L^2$

Isometries with respect to $L^1$ are relevant because they are permissible transformations of (classical) probability distributions. For example, the transition matrix that describes the flow of probability between the time steps of a Markov Chain has to ensure that the entries of the state space probability still sum to 1.

For norms $L^q$ with $q \neq 2$, the unit "circle" $\{\mathbf{x} : ||\mathbf{x}||_{q\neq 2} = 1\}$ is not perfectly round. Instead the symmetry is broken along the coordinate axes. That means that a rotation of the coordinate axes gives a different unit distance, and so two points $\mathbf{a}$ and $\mathbf{b}$ that are $||\mathbf{a}-\mathbf{b}||_{q\neq 2} = 1$ apart in one coordinate system may have a different separation in some other coordinate system.

In particular, for the $L^1$, or *Manhattan* norm (cf. section 3.12.1), the points $\{\mathbf{x} : ||\mathbf{x}||_1 = 1\}$ lie on a diamond with axes aligned to the axes of the coordinate system. In Manhattan, reaching a point that is 1 mile away "as the crow flies", depends on the position of that point with respect to the grid of streets and avenues. When that grid is rotated, the point may be quicker or take longer to reach.

Since this rules out rotations, the internet tells me that the only linear maps that are isometries for $L^q$ with $q \neq 2$ are signed permutation matrices. That is, matrices that assign $\hat{x} \to \hat{y}$ or $\hat{z} \to -\hat{x}$ and so on.

However, in general, any transformation that maps a point on the $L^q$ unit circle to another point on the $L^q$ unit circle is an isometry with respect to $L^q$, and this class of transformation is more general than signed permutations. The stochastic matrices are an example with respect to $L^1$.

### 3.10.9 Stochastic

Stochastic matrices are used to describe transitions between states of a Markov chain. The matrices are square and each entry is non-negative and represents a conditional probability of moving from one state to another. The entries along either the row or the column, or both, must sum to 1 according to the requirement of marginalization.

- Right stochastic has rows that sum to 1, so that $\mathbf{P1} = \mathbf{1}$. It is applied $\pi\mathbf{P}$. Column index gives "from", row index gives "to".
- Left stochastic has columns that sum to 1, so that $\mathbf{1P} = \mathbf{1}$. It is applied $\mathbf{P}\pi$. Row index gives "from", column index gives "to".
- doubly stochastic has rows and columns that sum to 1.

Products of stochastic matrices are also stochastic matrices. The spectral radius (largest eigenvalue) of a stochastic matrix is always 1. Since $\mathbf{1}$ is an eigenvector and the left and right eigenvalues of a square matrix are the same, there is at least one stationary state $\pi\mathbf{P} = \pi$ (that is, in this case of a right stochastic matrix, a left eigenvector) with eigenvalue 1.

### 3.10.10 $\mathrm{U}^\dagger = \mathrm{U}^{-1}$ Unitary, Orthogonal

Unitary matrices satisfy $\mathbf{U}^\dagger\mathbf{U} = \mathbf{U}\mathbf{U}^\dagger = \mathbf{I}$, and they have $det(\mathbf{U}) = 1$. They are diagonalizable and can be expressed as $e^{i\mathbf{H}}$ where $\mathbf{H}$ is a Hermitian matrix.

Unitary matrices that are real are called orthogonal. Orthogonal matrices satisfy $\mathbf{A}^{-1} = \mathbf{A}^T$. The rows (and columns) of $\mathbf{A}$ are an orthonormal basis in $\mathrm{R}^n$.

Unitary matrices are necessarily invertible, and have determinant $|U| = 1$ or $|U| = -1$. They represent *unitary transformations*, which means that they preserve the inner product between two vectors.

The set of $n \times n$ orthogonal matrices is known as the orthogonal group $O(n)$ and the subgroup of orthogonal matrices with determinant $1$ is known as the special orthogonal group $SO(n)$. The elements of $SO(n)$ are rotations, and the elements of $O(n)$ represent translations, reflections or rotations. Similarly, the group of $n \times n$ unitary matrices is the unitary group $U(n)$ and the subgroup of $U(n)$ that has determinant $1$ is the special unitary group $SU(n)$.

Unitary transformations preserve the $L^2$ norm of vectors.

### 3.10.11 $\mathbf{A} = \mathbf{T}\mathbf{B}\mathbf{T}^{-1}$ Similarity

Two matrices are said to be similar if they can be related through a similarity transformation $\mathbf{A} = \mathbf{T}\mathbf{B}\mathbf{T}^{-1}$ where $\mathbf{T}$ is some nonsingular matrix (cf. section 3.9.1). An important example is that square matrices are similar to diagonal matrices, see section 3.6.

## 3.11 Properties of Norms

## 3.12 $L^p$ Lebesgue Vector Norms

Let $p \geq 1$ be a real number, then the $p$-norm or $L^p$ norm of a vector $\mathbf{x} \in \mathbb{C}^n$ is defined as:

$$||\mathbf{x}||^p = \left[\sum_i^n |x_i|^p\right]^{1/p} \tag{3.36}$$

The expression can still be useful for $0 < p < 1$, but in that case the result is not a proper norm, because it is not subadditive (does not satisfy $f(x+y) \leq f(x) + f(y)$). $p$-norms are closely related to expressions for the generalized mean.

### 3.12.1 $L^1$ Taxicab / Manhattan Norm

$$||\mathbf{x}||^1 = \sum_i |x_i| \tag{3.37}$$

In the context of regression, $L^1$ loss gives the maximum likelihood estimator under the assumption of Laplacian (double exponential) distributed errors. For a dataset $\mathbf{x} \in S$, minimizing $\arg\min_{\mathbf{s}} \sum_{\mathbf{x}\in S} ||\mathbf{x}-\mathbf{s}||_1$ gives the median.

### 3.12.2 $L^2$ Euclidian Norm

$$||\mathbf{x}||^2 = \sum_i |x_i|^2 \tag{3.38}$$

In the context of regression, $L^2$ loss gives the maximum likelihood estimator under the assumption of normally distributed errors. For a dataset $\mathbf{x} \in S$, minimizing $\arg\min_{\mathbf{s}} \sum_{\mathbf{x}\in S} ||\mathbf{x}-\mathbf{s}||_2$ gives the mean.

I believe that $L^2$ norm should be the only norm that preserves the distance between two points under rotations of the coordinate system.

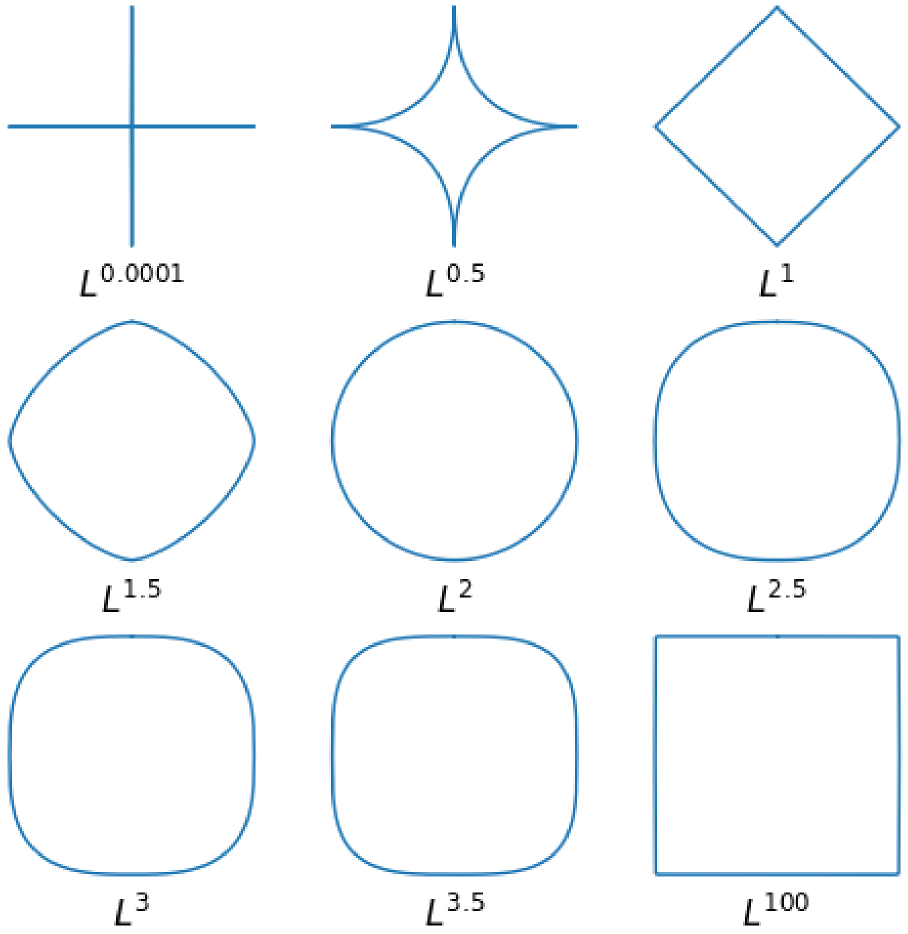


Figure 3.3: Unit Circles: Level sets $\{\mathbf{x} : ||\mathbf{x}||_q = 1\}$ for different $L^q$ norms.

| $p$ | Location $\mathbb{L}oc\{X\}$ | Dispersion $\mathbb{D}isp\{X\}$ |
|---|---|---|
| 1 | Median (32.271) | Mean absolute deviation (32.272) |
| 2 | Expectation (32.22) | Standard deviation (32.23) |
| $\infty$ | Midrange (32.274) | Maximum deviation (32.274) |
| 0 | Mode (32.278) | Variation ratio (32.279) |

Figure 3.4: Distance metrics base for different $L^q$ norms.

### 3.12.3 $L^\infty$ Maximum Norm

$$||\mathbf{x}||^\infty = \max(x_1, x_2, ..., x_n) \tag{3.39}$$

For a dataset $\mathbf{x} \in S$, minimizing $\arg\min_\mathbf{s} \sum_{\mathbf{x}\in S} ||\mathbf{x} - \mathbf{s}||_\infty$ gives the average of the maximum and the minimum value of the dataset.

### 3.12.4 $L^{-\infty}$ Minimum Norm

Formally, I only came across values $0 < p$, but it is my opinion that $-\infty$ picks out the minimum value:

$$||\mathbf{x}||^{-\infty} = \min(x_1, x_2, ..., x_n) \tag{3.40}$$

## 3.13 Operator and Matrix Norms

Matrix norms are functions $||\cdot|| : K^{m\times n} \to \mathbb{R}$ where $K$ is a field of real or complex numbers. They satisfy:

- $||\alpha A|| = |a|||A||$ (absolutely homogenous)
- $||A + B|| \leq ||A|| + ||B||$ (triangle inequality, subadditivity)
- $||A|| \geq 0$ (positive valued)
- $||A|| = 0 \implies A_{n,m} = 0$ (definiteness)

A norm is submultiplicative if it satisfies $||AB|| \leq ||A||||B||$, which Gera (2009) calls a requirement of "useful matrix norms".

The main risk of confusion is that norms for operators and vectors are different animals. Norms for operators normally measure some relationship between input and output. Norms for vectors are normally some kind of size, length or distance metric. In as far as matrices can be thought of as both operators and multidimensional vectors, norms of either type may be applied to them. People's notation and language is all over the place. Below I've used $||\cdot||_{(\alpha)}$ to denote norms in the operator sense and $||\cdot||_\alpha$ in the vector sense.

### 3.13.1 $||\mathbf{A}||_{(\alpha)}$ Operator Norm

The operator norm describes the largest change in size that it may impart on any of its inputs. That means that the operator norm is defined with respect to a definition of size in both domain and codomain. I.e., for an operator $\mathbf{A}$ and a given way of measuring size $||\cdot||_\alpha$:

$$||\mathbf{A}||_{(\alpha)} = \sup\left\{\frac{||\mathbf{Av}||_\alpha}{||\mathbf{x}||_\alpha} : \mathbf{v} \in V\right\} \tag{3.41}$$

When the operator is given by a matrix $\mathbf{A}$, and the length of the vector $\mathbf{x}$ is measured using the usual euclidian 2-norm ($||\cdot||_2$), then the operator norm is given by the square root of the largest eigenvalue of $\mathbf{A}^\mathbf{T}\mathbf{A}$. In that case, the operator norm is the same as the 2-norm (cf. section 3.13.6).

To re-emphasize, $||A||_{(q)}$ and $||A||_q$ are two different things. The former measures the change in input size, where the size of the input is measured according to the latter. That is the reason for why the 1-Norm and 2-Norms are so different from the vector norms $L_1$ and $L_2$.

Operators that preserve the length of a vector with respect to some norm $||\cdot||_\alpha$ satisfy $||\mathbf{A}||_{(\alpha)} = 1$ and are called isometries (cf. section 3.10.8).

### 3.13.2 $||\mathbf{A}||_q$ $q$-Norms

The $q$ norms for a matrix $\mathbf{A} \in \mathbb{R}^{m\times n}$ with entries $a_{i,j}$ in row $i$ and column $j$ are defined:

$$||\mathbf{A}||_q = \left( \sum_i \sum_j a_{i,j}^q \right)^{1/q} \tag{3.42}$$

For $q = 2$, this becomes the Frobenius norm (section 3.13.3). For vectors $\mathbf{v} \in \mathbb{R}^n$, the $q$-norm is more known as $p$-norm or $L^p$ norm (cf. section 3.12).

### 3.13.3 $||\mathbf{A}||_F$ Frobenius Norm

The Frobenius Norm is the sum of the squares of all entries of a matrix. Let $\mathbf{A} \in \mathbb{R}^{m\times n}$ be a matrix with entires $a_{i,j}$ in row $i$ and column $j$, then:

$$||\mathbf{A}||_F = \sqrt{\sum_i \sum_j a_{i,j}^2} \tag{3.43}$$

The Frobenius norm is invariant under rotations, and $||\mathbf{A}||_F = \sqrt{\sum_i \sigma_i^2}$ where $\sigma_i$ are the singular values of $\mathbf{A}$.

### 3.13.4 $||\mathbf{A}||_{(1)}$ (1)-Norm

Let $\mathbf{A}$ be a matrix with entires $a_{i,j}$ in row $i$ and column $j$, then:

$$||\mathbf{A}||_{(1)} = \max_{1\leq j\leq n} \sum_{i=1}^{m} |a_{i,j}| \tag{3.44}$$

That is, it is the maximum of the sums of the absolute values of any of the columns of $\mathbf{A}$.

### 3.13.5 $||\mathbf{A}||_{(\infty)}$ $(\infty)$-Norm

Let $\mathbf{A}$ be a matrix with entires $a_{i,j}$ in row $i$ and column $j$, then:

$$||\mathbf{A}||_{(\infty)} = \max_{1\leq i\leq m} \sum_{j=1}^{n} |a_{i,j}| \tag{3.45}$$

That is, it is the maximum of the sums of the absolute values of any of the rows of $\mathbf{A}$.

### 3.13.6 $||\mathbf{A}||_{(2)}$ (2)-Norm

Let $\mathbf{A} \in \mathbb{R}^{m\times n}$ be a matrix with entires $a_{i,j}$ in row $i$ and column $j$, then:

$$||A||_{(2)} = \max_{\mathbf{x}\neq 0} \frac{||\mathbf{Ax}||_2}{||\mathbf{x}||_2} \tag{3.46}$$

Which is the square root of the largest eigenvalue of $A^T A$. Or, equivalently, $||A||_{(2)} = \sigma_1$, where $\sigma_1$ is the largest singular value of the SVD of $\mathbf{A} = \mathbf{U\Sigma V}^T$. That means that $||A^{-1}|| = \frac{1}{\sigma_n}$, where $\sigma_n$ is the smallest singular value of the SVD of $\mathbf{A}$.

### 3.13.7 $||\mathbf{A}||_*$ Nuclear Norm

The nuclear norm of a matrix is the sum of its singular values. This is equivalent to the Schatten-1 Norm. Optimization problems that involve rank restriction tend to be non-convex. The nuclear norm is a convex alternative that can make such problems tractable.

$$||\mathbf{A}||_* = \sum_i \sigma_i \tag{3.47}$$

## 3.14 Vector and Matrix Derivatives

Derivatives involving matrices and vectors can look nonintuitive when the usual symbolic matrix notation is used, but can be derived handily when index notation is used. A very concise and helpful resource for this is Barnes (n/a).

### 3.14.1 Jacobian

It is particularly helpful to remember the Jacobian, which is the derivative of a function with respect of a vector. The Jacobian of some function $f : \mathbb{R}^n \to \mathbb{R}^m$ is:

$$\frac{\mathrm{d}\mathbf{f}(\mathbf{x})}{\mathrm{d}\mathbf{x}} = \left[\frac{\partial \mathbf{f}}{\partial x_1}, \dots, \frac{\partial \mathbf{f}}{\partial x_n}\right] = \begin{bmatrix} \frac{\partial f_1}{\partial x_1} & \cdots & \frac{\partial f_1}{\partial x_n} \\ \vdots & \vdots & \vdots \\ \frac{\partial f_m}{\partial x_1} & \cdots & \frac{\partial f_m}{\partial x_n} \end{bmatrix} \tag{3.48}$$

I enjoy writing the gradient $\frac{\mathrm{d}}{\mathrm{d}\mathbf{x}}$ as $\nabla_\mathbf{x}$. The relationships below can all be derived as applications of the Jacobian.

$$\begin{gathered}
\nabla_\mathbf{x}\left(\mathbf{u}^T\mathbf{x}\right) = \left[\tfrac{\partial}{\partial x_1}\left(\textstyle\sum_i u_i x_i\right), ..., \tfrac{\partial}{\partial x_n}\left(\textstyle\sum_i u_i x_i\right)\right] = \mathbf{u}^T \\
\nabla_\mathbf{x}\left(\mathbf{x}^T\mathbf{u}\right) = \left[\tfrac{\partial}{\partial x_1}\left(\textstyle\sum_i u_i x_i\right), ..., \tfrac{\partial}{\partial x_n}\left(\textstyle\sum_i u_i x_i\right)\right] = \mathbf{u}^T \\
\nabla_\mathbf{x}\left(\mathbf{x}^T\mathbf{x}\right) = \left[\tfrac{\partial}{\partial x_1}\left(\textstyle\sum_i x_i^2\right), ..., \tfrac{\partial}{\partial x_n}\left(\textstyle\sum_i x_i^2\right)\right] = 2\mathbf{x}^T \\
\nabla_\mathbf{x}\left(\mathbf{A}\mathbf{x}\right) = \begin{bmatrix} \underbrace{\frac{\partial}{\partial x_1}\left(\sum_i A_{1i} x_i\right)}_{A_{11}} & ... & \underbrace{\frac{\partial}{\partial x_n}\left(\sum_i A_{1i} x_i\right)}_{A_{1}n} \\ \vdots & \vdots & \vdots \\ \underbrace{\frac{\partial}{\partial x_1}\left(\sum_i A_{ni} x_i\right)}_{A_{n1}} & ... & \underbrace{\frac{\partial}{\partial x_n}\left(\sum_i A_{ni} x_i\right)}_{A_{nn}} \end{bmatrix} = \mathbf{A}
\end{gathered} \tag{3.49}$$

### 3.14.2 Inverse Function Theorem

The inverse function theorem gives a sufficient condition for the invertibility of a function near some point in its domain. If the derivative $f'$ of a function $f$ is continuous and non-zero near some point $a$ within its domain, then the function is invertible near that point. If $b = f(a)$, then:

$$\frac{d\left[f^{-1}(b)\right]}{dx} = \frac{1}{\frac{df(a)}{dx}} \tag{3.50}$$

That is, the derivative of the inverse function at a point $b = f(a)$ of the range, is the reciprocal of the derivative of the function near the point $a$ in the domain. This extends to multivariable calculus. Given a function $\mathbf{f} : \mathbf{x} \to \mathbf{y}$:

$$\nabla_{\mathbf{y}} \left[\mathbf{f}^{-1}\right] = \left[\nabla_{\mathbf{x}}\mathbf{f}\right]^{-1} \tag{3.51}$$

In words: the Jacobian of the inverse function at the point $\mathbf{b} = \mathbf{f}(\mathbf{a})$ is the matrix inverse of the Jacobian of the function at the point $\mathbf{a}$. The sufficient condition is that the Jacobian $\nabla_{\mathbf{x}}\mathbf{f}$ is continuous and *nonsingular* near $\mathbf{a}$.

### 3.14.3 Critical Points

Critical points are points where the Jacobian does not have maximal rank. In case of a square Jacobian, this means that the Jacobian is singular.

### 3.14.4 Differential Volume Element, Change of Variables

The Jacobian is used when transforming between different coordinate systems. Consider a transformation $\mathbf{x} = \mathbf{H}(\mathbf{y})$, then:

$$\mathrm{d}^n x = \left|\nabla_{\mathrm{y}}\mathbf{H}\right| \mathrm{d}^n y \tag{3.52}$$

And:

$$\int_{\mathbf{x}} \mathrm{d}^n \mathbf{x} f(\mathbf{x}) = \int_{\mathbf{y}} \mathrm{d}^n \mathbf{y} \left|\nabla_{\mathrm{y}}\mathbf{H}\right| f(\mathbf{H}(\mathbf{y})) \tag{3.53}$$

Alternatively, if $\mathbf{y} = \mathbf{H}^{-1}(\mathbf{x})$:

$$\begin{aligned} \mathrm{d}^n y &= \left|\nabla_{\mathrm{x}}\mathbf{H}^{-1}(\mathbf{x})\right| \mathrm{d}^n x \\ &= \left|\left[\nabla_{\mathrm{y}}\mathbf{H}(\mathbf{y})\right]^{-1}\right| \mathrm{d}^n x \\ \mathrm{d}^n x &= \frac{1}{\left|\left[\nabla_{\mathrm{y}}\mathbf{H}(\mathbf{y})\right]^{-1}\right|} \mathrm{d}^n y \end{aligned} \tag{3.54}$$

The Jacobian has to be nonsingular within the domain of integration. This implies that $\mathbf{x}$ and $\mathbf{y}$ have to have the same dimension. In the context of probability theory that sometimes requires artificially defining additional variables so that $\mathbf{H}$ is bijective because the quantity of interest has lower dimension (for example, if you calculate the mean of a random variable).

### 3.14.5 Hessian

The Hessian is the second derivative of a scalar valued function $f : \mathbb{R}^n \to \mathbb{R}$ with respect to a vector, i.e. $\nabla \cdot \nabla f$. The elements are $\mathbf{H}(f)_{i,j} = \frac{\partial^2 f}{\partial x_i \partial x_j}$.

$$\mathbf{H}(f) = \begin{bmatrix} \frac{\partial^2 f}{\partial x_1^2} & \frac{\partial^2 f}{\partial x_1 \partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_1 \partial x_n} \\ \frac{\partial^2 f}{\partial x_2 \partial x_1} & \frac{\partial^2 f}{\partial x_2^2} & \cdots & \frac{\partial^2 f}{\partial x_1 \partial x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 f}{\partial x_n \partial x_1} & \frac{\partial^2 f}{\partial x_n \partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_n^2} \end{bmatrix} \tag{3.55}$$

The Hessian of a vector valued function $f : \mathbb{R}^n \to \mathbb{R}^m$ is a third order tensor with elements $\mathbf{H}(\mathbf{f})_{i,j,k} = \frac{\partial^2 f_k}{\partial x_i \partial x_j}$.

#### 3.14.5.1 Testing Convexity

The definiteness (cf. section 3.10.1) of the Hessian is used to test convexity.

| | |
|---|---|
| $\mathbf{H} \succeq 0$ | convex |
| $\mathbf{H} \succ 0$ | strictly convex |
| $\mathbf{H} \preceq 0$ | concave |
| $\mathbf{H} \prec 0$ | strictly concave |

If this holds at a point, the property is local (for example at a local maximum or minimum), and if it holds everywhere on the domain, then the property is global.

## 3.15 Fundamental Theorems of Calculus

## 3.16 Leibnitz Integral Rule

Differentiation under the integral sign.

$$\frac{\partial}{\partial x} \int_{\Omega} f(x,\omega) d\omega = \int_{\Omega} f(x,\omega) d\omega \tag{3.56}$$

Where $X \subseteq \mathbb{R}$ is an open subset, $\Omega$ is a measure space, and the following conditions are satisfied:

- xx

For example, consider:

# 4 Probability

### Contents of this chapter

## 4.1 Interpretations and Definitions of Probability

Weirdly enough, every book I open has a different approach to introducing probability. I found Wasserman (2013) the best resource so far in terms of being clear and applied at the same time. Wasserman's language and notation is congruent with probability theory founded in measure theory. Lindgren (2006) is the best resource so far in introducing mathematical probability in the context of stochastic processes. The Bright Side of Mathematics (2019) has an accessible lecture series on measure theory and a very calming German accent.

### 4.1.1 Blitzstein's Naive and Non-Naive Definitions of Probability

Like most probability texts, Blitzstein & Hwang (2019) starts with what he calls the *naive* definition of probability, which is to look at the fraction of the event space that corresponds to a particular event. He then introduces a *non-naive* definition of probability, which is congruent with measure theoretic probability.

#### 4.1.1.1 Naive Probability, Uniform Probability

The event space $S$ consists of a collection of equally likely outcomes. The actual outcome $s_{actual} \in S$. The probability of an outcome in a subset $A \subseteq S$, $P(s_{actual} \in A)$ corresponds to the fraction of events in $A$ out of $S$.

$$P_{naive}(A) = \frac{|A|}{|S|} = \frac{\text{number of outcomes favorable to A}}{\text{total number of outcomes in S}} \tag{4.1}$$

I've seen this called combinatorial probability or counting probability, presumably because this is how you calculate the probability in situations where events correspond to a finite number of equally likely combinations. I call it *job interview probability*. More helpfully, Wasserman (2013) calls it the *uniform probability distribution*.

The method of counting the possible outcomes $s_{actual}$ within $S$ and $A$ is not appropriate when the outcomes are not equally likely (because counting them weighs them all equally). Also, the definition is, by itself, unclear on how to deal with infinite probability spaces (for example: $s_outcome \in \mathbb{R}$).

#### 4.1.1.2 Blitzstein's Non-Naive Definition of Probability

This definition is from (Blitzstein & Hwang 2019) and is congruent with measure theoretic probability. A probability space consists of a sample space $S$ in addition to a *probability function.* The job of the probability function is to take an event $A \subseteq S$ and map it to a number between 0 and 1. I.e. $P : S \rightarrow [0, 1]$.

The function $P$ must satisfy:

- $P(\emptyset) = 0, P(S) = 1$
- If $A_1, A_2, ...$ are *disjoint* events, then:

$$P\left(\bigcup_{j=1}^{\infty} A_j\right) = \sum_{j=1}^{\infty} P(A_j) \tag{4.2}$$

In other words, all you need to work with probability is a set of possible outcomes $S$ and some function to apply to parts of that set.

### 4.1.2 Frequentist Probability

The *frequentist* interpretation of probability is that it represents the frequency of an outcome of a particular experiment upon running the experiment many times. On one hand, this view on probability is arguably empirically grounded. On the other hand it relies on the impractical notion of identically repeating something a large number of times. Wasserman (2013) describes *frequentist inference* as inference with guaranteed frequency behavior.

### 4.1.3 Bayesian Probability

The *Bayesian* view on probability is that it represents a degree of belief in a particular outcome. On one hand, this implies a degree of subjectivity. On the other hand, this notion of probability is much more broadly applicable, because it does not require repeatable experiments. It also permits the introduction of subjective biases through priors. Wasserman (2013) describes *Bayesian inference* as statistical methods that use data to update belief.

I'm not sure whether the supposed "battle" between frequentists and Bayesians has ever been of any consequence for me. Also, the frequentist perspective strikes me as inherently contradictory, because supposed empirical proof can strictly speaking never be practically delivered.

In practice, it seems that frequentist approaches tend to be structured around parametrization of solutions in terms of summary statistics, while Bayesians have a tendency to approach problems in terms of full probability distributions. Social scientists seem to be leaning towards frequentist methods, while physicists and engineers seem to be more interested in Bayesian data analysis.

I suspect that an additional source of bias is that frequentist methods tend to be computationally lighter and that many standard accepted recipes exist for common problems like hypothesis testing. In contrast, Bayesian approaches are usually derived ad-hoc in the context of a particular problem. They tend to involve more explicit assumptions about the data generating process.

### 4.1.4 Kolmogorov's Axioms

### 4.1.5 Cox' Theorems

## 4.2 Measure Theoretic Probability

The modern approach to probability is rooted in measure theory, and use the following notation and definitions. Kolmogorov's three probability axioms overlap with them. Cox' theorems, an alternative approach to formalizing probability that is rooted in propositional logic, imply them. Conveniently, the other famous child of measure theory happens to be Lebesgue Integration, which allows for rigorous treatment of infinite sample spaces, distribution functions, and so on, all based within the same framework.

### 4.2.1 Sample Space, Outcomes, Events

- $\Omega$ is the *state space*, the set of all possible outcomes, and maybe some impossible ones too. It doesn't matter, as long as all the possible outcomes are included. Sometimes also called *event space* or *sample space*.
- A point or element $\omega \in \Omega$ is an *outcome*. Sometimes also called *element*, *sample outcome* or *realization*.
- A subset $A \subseteq \Omega$ is an *event*.

$\Omega$, $A$ and $\omega$ are sets and elements of sets, and the correct operations on them are set operations like forming unions and intersections. This is different from random variables, which might be numbers that can be added, subtracted, multiplied etc.

### 4.2.2 Probability Distributions

Probability is a *measure* that assigns a real number between $0$ and $1$ to events within the state space. It turns out that it is not always possible to define a meaningful measure to arbitrary collections subsets of the state space $\Omega$, so that the probability measure is instead assigned to a $\sigma$-algebra $\mathscr{A}$, which is a collection of subsets of $\Omega$ that fulfills criteria that guarantee the existence of a measure. $\sigma$-algebras are also called $\sigma$-fields. The combination $(\Omega, \mathscr{A})$ is called a measurable space.

Given a measurable space $(\Omega, \mathscr{A})$, a *probability distribution* or *probability measure* $\mathbb{P}$ is a measure $\mathbb{P} : \mathscr{A} \to \mathbb{R}$. The following criteria follow from the definition of a measure:

- Null-Event: $\mathbb{P}(\emptyset) = 0$. The empty subset $\emptyset \in \mathscr{A}$ is also called the *null event*.
- Additivity: $\mathbb{P}(\bigcup_i A_i) = \sum_i \mathbb{P}(A_i)$ if $A_i \cap A_j = \emptyset$ when $i \neq j$. The probability of the union of a finite collection of disjoint subsets $A_i \in \mathscr{A}$ is additive.
- $\sigma$-additivity: $\mathbb{P}(\bigcup_i^\infty A_i) = \sum_i^\infty \mathbb{P}(A_i)$ if $A_i \cap A_j = \emptyset$ when $i \neq j$. The probability of the union of a countable, possibly infinite collection of disjoint subsets $A_i \in \mathscr{A}$ is also additive. (This derives from a feature of $\sigma$-algebras, which have to be closed under taking the countable union of subsets.) This is Kolmogorov's third axiom.

In addition, a probability measure satisfies:

- Positivity: $\mathbb{P}(A) \geq 0$ for all subsets $A \in \mathscr{A}$. Though positivity is also sometimes included in the definition of a measure in general. Positivity is part of Kolmogorov's first probability axiom, which requires probability to be a positive, real number.
- Unitarity, Unit Measure: $\mathbb{P}(\Omega) = 1$. This represents certainty that something within the sample space must happen. Additivity ensures that probability is never greater than $1$. This is Kolmogorov's second axiom.

The combination of a sample space, a $\sigma$-algebra and a probability measure $(\Omega, \mathscr{A}, \mathbb{P})$ is a *probability space*. The important case of the $\sigma$-algebra generated by the real numbers $\sigma(\mathbb{R})$ is called the Borel $\sigma$-algebra, $\mathscr{B}(\mathbb{R})$.

### 4.2.3 Random Variables

Random variables are *measurable functions* defined on a probability space that assigns a real number to each outcome. That is, given $(\Omega, \mathscr{A}, \mathbb{P})$, a random variable is some function $x(\omega), \omega \in \Omega$, $X : \Omega \to \mathbb{R}$. The fact that $X(\omega)$ is a *measurable function* (also called *measurable map*) is that both the domain and the range are measurable spaces, and that for each element of the range, the preimage is contained contained within a $\sigma$-algebra, and hence measurable. This allows one to talk about the *distribution* of a random variable by applying the probability measure to subsets of its preimage. For example, one was interested in the probability that $X(\omega)$ will fall on the interval $[0, a]$ for some $a \in \mathbb{R}$, one feeds the subset of outcomes $A_{0 \leq X \leq a} = \{\omega : \omega = X^{-1}(x), x \in [0, a]\}$ to the probability measure:

$$\mathbb{P}(0 \leq X \leq a) = \mathbb{P}(A_{0 \leq X \leq a}) = \mathbb{P}(\{\omega : 0 \leq X(\omega) \leq a\}) \tag{4.3}$$

### 4.2.4 Cumulative Distribution Function (CDF)

The cumulative distribution function (CDF) of a random variable $X(\omega), \omega \in \Omega$ is given by $F_X(x) = \mathbb{P}(X \leq x)$. It measures the subset $\{\omega : \omega \in \Omega, X(\omega) \leq x\}$. A CDF uniquely determines its underlying random variable. It satisfies the condition:

- $F_X$ is normalized: $0 \leq F_X(x) \leq 1$, $\lim_{x \to -\infty} F_X(x) = 0$, $\lim_{x \to \infty} F_X(x) = 1$
- $F_X$ is non-decreasing. That is, $F(x_1) \leq F(x_2)$ if $x_1 \leq x_2$.
- Right-continuous: $F(x) = F(x^+)$

With respect to the probability of $X(\omega)$ assuming particular values, it follows:

- $\mathbb{P}(X = x) = F_X(x) - F_X(x^-)$ where $F_X(x^-) = \lim_{y \uparrow x} F_X(y)$
- $F_X(b) - F_X(a) = \mathbb{P}(a < x \leq b)$
- $\mathbb{P}(X > x) = 1 - F_X(x)$

If $X$ is continuous, then:

- $F_X(b) - F_X(a) = \mathbb{P}(a < X < b) = \mathbb{P}(a \leq X \leq b) = \mathbb{P}(a \leq X < b) = \mathbb{P}(a < X \leq b)$

### 4.2.5 Probability Mass Function (PMF)

If a random variable $X$ takes on discrete values, then it has a probability mass function (PMF) $f_X(x) = \mathbb{P}(X = x)$. The probability mass function measures the preimage of $x$, which is the subset $\{\omega : \omega \in \Omega, \omega = X^{-1}(x)\}$. The relationship to the CDF is $F_X(x) = \sum_{x_i \leq x} f_X(x_i)$.

### 4.2.6 Probability Density Function (PDF)

If a random variable $X$ takes on continuous values, then it has a probability density function (PDF) $\mathbb{P}(a \leq X \leq b) = \int_a^b f_X(x) \mathrm{d}x$. It measures the size of the subset $\{\omega : \omega \in \Omega, a \leq X(\omega) \leq b\}$. The relationship to the CDF is $F_X(x) = \int_{-\infty^x f_X(x) \mathrm{d}x}$, or $\frac{\mathrm{d}}{\mathrm{d}x} F(x) = f_X(x)$.

### 4.2.7 Quantile Function (Inverse CDF)

The quantile function, or inverse CDF, is given by $F_X^{-1}(q) = \inf\{x : F_X(x) > q\}$. It gives the value of $X(\omega)$ at which the CDF assumes the value $q$.

- First quartile: $q = 1/4$
- Second quartile / Median: $q = 1/2$
- Third quartile: $q = 3/4$

### 4.2.8 Joint, Conditional and Marginal Distributions

Let $X(\omega)$ and $Y(\omega)$ be two random variables defined on some $(\Omega, \mathscr{A}, \mathbb{P})$. The joint distribution mesaures the probability of some subset bounded by conditions on both $X$ and $Y$. For example, in the discrete case, the joint distribution $f_{XY}(x,y) = \mathbb{P}(X = x, Y = y)$ measures $\{\omega : X(\omega) = x, Y(\omega) = y\}$. When the value of one of the variables is considered fixed, then the joint distribution is the conditional distribution $f_{X|Y}(x) = \mathbb{P}(X = x|Y = a)$, which measures the subset $\{\omega : X(\omega) = x, Y(\omega) = a\}$. The marginal distribution is the distribution of a single random variable in a setting in which multiple random variables are defined on an event space. Also written $f_X$.

## 4.3 Mutually Exclusive Events, Disjoint Sets

Two events $A$ and $B$ are mutually exclusive if $A \cap B = \emptyset$. That is, $A$ and $B$ are disjoint sets.

## 4.4 Independent Events

Interestingly, independence can, with the exception of specific symmetric cases, generally not be read off from a Venn diagram (Wasserman 2013). A set of events $A_i$ is *independent* if:

$$\mathbb{P}\left(\bigcap_i A_i\right) = \prod_i \mathbb{P}(A_i) \tag{4.4}$$

Wasserman (2013) writes independence using the coproduct symbol $A \coprod B$. Independence can be either assumed or it can be proven by verifying $\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B)$.

## 4.5 Conditional Probability

Conditional probability $\mathbb{P}(A|B)$ "probability of $A$ given $B$" is the ratio of the probability measure applied to the subsets $A \cap B$ and $B$:

For $\mathbb{P}(B) > 0$:

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)} \tag{4.5}$$

For dependent events, this can be interpreted as the fraction of $B$ that overlaps with $A$. For independent events, $\mathbb{P}(A|B) = \mathbb{P}(A)$.

## 4.6 Law of Total Probability

Let $A_1, ..., A_n$ be a partition of $\Sigma$ so that $\mathbb{P}(\bigcup_i A_i) = \mathbb{P}(\Omega) = 1$. Then:

$$\mathbb{P}(B) = \sum_i \mathbb{P}(B|A_i)\mathbb{P}(A_i) \tag{4.6}$$

## 4.7 Bayes' Theorem

Let $A_1, ..., A_n$ be a partition of $\Sigma$ so that $\mathbb{P}(\bigcup_i A_i) = \mathbb{P}(\Omega) = 1$ and $\mathbb{P}(A_i) > 0$ for each $A_i$. Then:

$$\mathbb{P}(A_i|B) = \frac{\mathbb{P}(B|A_i)\mathbb{P}(A_i)}{\sum_j \mathbb{P}(B|A_j)\mathbb{P}(A_j)} = \frac{\mathbb{P}(B|A_i)\mathbb{P}(A_i)}{\mathbb{P}(B)} \tag{4.7}$$

The theorem is so important that the different terms have names:

- $\mathbb{P}(A_i|B)$ is the posterior
- $\mathbb{P}(B|A_i)$ is the likelihood
- $\mathbb{P}(A_j)$ is the prior
- $\mathbb{P}(B)$ is the evidence

## 4.8 Functions of Random Variables, Derived Distributions

(This section needs revision)

I find changes of variables to be the easiest to understand by writing down a joint probability distribution using a delta function to express the conditional probability for the new variables based on the old variables. The next step is to perform a change of variables in the delta function, and integrate. The delta function means that the relationship between the old and the new variables is deterministic. If the relationship is non-deterministic you can use whatever expression for the conditional probability is appropriate.

Let's say that you want to know the probability distribution of the function $\vec{u} = \mathbf{H}(\vec{x})$ of some random variable $\vec{x}$ for which the probability distribution is known.

I personally find it least confusing to approach this problem by thinking about the joint probability distribution, and then obtaining $p(\vec{u})$ through marginalization:

$$p(\vec{u}) = \int dx^n p(\vec{x}, \vec{u}) = \int dx^n p_x(\vec{x}) p_u(\vec{u}|\vec{x}) \tag{4.8}$$

Since $\vec{u}$ is a deterministic function of $\vec{x}$, the conditional probabiltiy $p_u(\vec{u}|\vec{x}) = \delta(\vec{u} - \mathbf{H}(\vec{x}))$. Explicitly:

$$p(\vec{u}) = \int dx^n p_x(\vec{x}) \delta(\vec{u} - \mathbf{H}(\vec{x})) \tag{4.9}$$

What this does, is to integrate over all the points $p_x(\vec{x})$ where the argument of the delta function is zero. The important thing is that care needs to be taken when the argument of the delta function is itself a function. In that case, a change of variables has to be performed, so that this is no longer the case. On wikipedia, this is done by defining the new variable $du = |\frac{d}{dx} g(x)| dx$, from which follows:

$$\delta(g(x)) = \sum_{x_0} \frac{\delta(x - x_0)}{|g'(x_0)|} \tag{4.10}$$

Where it is necessary to sum over each point $x_0$ for which $g(x_0) = 0$.

Which makes sense. Except, if you rewrite in terms of the new variable, you get an expression that does not strike me as necessarily the same:

$$\int f(x)\delta(g(x))dx = \sum_n \int f(g_n^{-1}(u))\delta(u) \left| \frac{d}{du} g^{-1}(u) \right| du \tag{4.11}$$

Where the sum $n$ is over all functions $g_n^{-1}$ that satisfy $g(g_n^{-1}(u)) = u$. For example, if $u = g(x) = sin(x)$ then $g_n^{-1} = asin(u) + n2\pi$, for any integer $n$. I feel more comfortable with the second path. The answer to my confusion is most likely the inverse function theorem (cf. section 3.14.2). In other words: this section needs revision! But hey, I flagged it.

Note that a delta function with a vector argument can be written as a product of the delta functions along each dimension.

Define a new set of variables $\vec{a} = \vec{u} - \mathbf{H}(\vec{x})$. It follows that $\vec{x} = \mathbf{H_n^{-1}}(\vec{u} - \vec{a})$ and the volume element $dx^n = \left| \frac{d}{d\vec{a}} \mathbf{H}_n^{-1}(\vec{u} - \vec{a}) \right| da^n$. Here, $H_n^{-1}$ are all the functions that satisfy $\mathbf{H}(\mathbf{H}_n^{-1}(\vec{u})) = \vec{u}$. The integral becomes:

$$p(\vec{u}) = \int da^n p_x(\mathbf{H}_n^{-1}(\vec{u} - \vec{a})) \left| \frac{d}{d\vec{a}} \mathbf{H}_n^{-1}(\vec{u} - \vec{a}) \right| \delta(\vec{a}) \tag{4.12}$$

At this point it is safe to evaluate the delta function integral. The result is:

$$p(\vec{u}) = p_x(\mathbf{H}_n^{-1}(\vec{u}) \left| \frac{d}{d\vec{a}} \mathbf{H}_n^{-1}(\vec{u} - \vec{a}) \right|_{(\vec{a}=0)} \tag{4.13}$$

### 4.8.1 Example: Sum of Random Variables

If the map $\mathbf{H}$ is not bijective, for example because $\vec{u}$ has lower dimensionality than $\vec{x}$, then a bijective map can be artifically constructed by introducing additional variables that are then also marginalized out. For example, if the goal is to calculate the probability of measuring some sum of random variables $s$, then you can define:

$$u_0 = s - \sum(x_i) u_1 = x_1 u_2 = x_2 \vdots u_{(n-1)} = x_{n-1} \tag{4.14}$$

The inverse is:

$$x_1 = u_1 x_2 = u_2 \vdots x_n = s - u_0 - \sum_{i<n} x_i \tag{4.15}$$

The argument in the delta function is transformed $\delta(s - \sum(x_i)) \to \delta(u_0)$. The determinant of the Jacobian $|J| = |\frac{d}{d\vec{u}} H^{-1}(\vec{u})|$ is given by $[\frac{d}{du_1}\mathbf{H^{-1}}, \frac{d}{du_2}\mathbf{H^{-1}}, ..., \frac{d}{du_{n-1}}\mathbf{H^{-1}}]$. That's an $nxn$ matrix where the first row is all $-1$, the lower left is an $(n-1)x(n-1)$ identity matrix, and the lower right is a $(n-1)x1$ vector of 0s. The determinant is one. Consequently the probability distribution of measuring a sum $s$ is given by:

$$p(s) = \int dx^{n-1} p_x(x_1, x_2, x_3, ..., s - \sum_{x<n} x_i) \tag{4.16}$$

Which turns out to be the convolution when the variables are independent.

### 4.8.2 Example: Lower Dimensional Random Variable

This was already the case for the sum of several random variables, in which case the dimensionality of the problem was reduced from many to one.

Consider the case many to fewer. Again, you just need to perform a change of variables. The new set of variables needs to have the same dimensionality as the old set of variables. The rest should follow pretty obviously.

Linear example: $\vec{y} = \mathbf{M}\vec{x}$ where $y$ has 2 dimensions and $x$ has 3.

Transform:

$$u_1 = y_1 - \vec{M}_1 \cdot \vec{x} u_2 = y_2 - \vec{M}_2 \cdot \vec{x} u_3 = x_3 \tag{4.17}$$

You can write this in terms of some invertible matrix $\mathbf{W} = [\mathbf{M}, [0,0,1]]$ as $\vec{u} = \mathbf{W}\vec{x}$, so that $dx^n = |\mathbf{W}^{-1}|du^n|$. The integral is then:

$$p(y) = \int du^n p_x(\mathbf{W}^{-1}(\vec{y} - \vec{u}))\delta(u_1)\delta(u_2)|\mathbf{W}^{-1}| \tag{4.18}$$

## 4.9 Indicator Variables

Indicator variables are useful devices that simplify probability calculations involving binary outcomes, namely wether the outcome lies in a certain region of event space or not. Specifically, they allow one to translate set expressions into algebraic expressions.

Let $A \subseteq S$ be a subset of event space $S$ (ex. $A \equiv$ "it rains tomorrow"). Then the indicator variable $I_A$:

$$I_A = \begin{cases} 1 \text{ if outcome in A} \\ 0 \text{ if outcome in A}^c \end{cases} \tag{4.19}$$

Which means that:

$$\mathbb{E}(I_A) = 1 \times p(A) + 0 \times p(A^c) = p(A) \tag{4.20}$$

To indicator variables for different events $A, B, C...$ can also be combined:

$$I_{A\cap B\cap C\cap...} = I_A I_B I_C... \tag{4.21}$$

And indicator variables for the complement can be constructed trivially:

$$I_{A^c} = 1 - I_A \tag{4.22}$$

Whenever accounting for complicated combinations of events becomes overwhelming, indicator variables are often a good approach.

### 4.9.1 Example: The Party Problem

The party problem is a typical interview question, so I will include the two easier problems that do not make use of indicator variables. (This version of my notes also still has bitterness included.)

There are $n$ drunk kids at a party that is presumably getting shut down by the fun police in Cambridge, MA. They (meaning the kids, presumably) grab their coats at random, and the problem is built around thinking about how many people wind up with the correct coat.

#### 4.9.1.1 Lame Interview Question 1: Every one finds their coat

The common version of this problem asks "what is the probability that all of the kids wind up with the right coat". This is much easier than the general case. Imagine the list of party guests as a sequence $(1, 2, 3, 4, ..., n-1, n)$ and the coats they grab as a random permutation of that sequence, such as $\alpha = (6, 1, 40, 21, 9, ...)_n$. There are $n!$ such permutations, and they are all equally likely. The probability of all guests getting the correct coat is the probability that the permutation happens to be the single correct one, i.e. $p(\alpha = (1, 2, 3, 4, ..., n-1, n)_n)$, which is the probality of one particular permutation, i.e. $p(\alpha = (1, 2, 3, 4, ..., n-1, n)_n) = \frac{1}{n!}$.

#### 4.9.1.2 Lame Interview Question 2: At least r people find their coat

This is still pretty easy, because the permutations that are correct are easily counted. If at least $r$ coats are correctly assigned, then there are $\begin{pmatrix} n \\ r \end{pmatrix}$ ways of choosing wich of the $r$ people wind up with the right coat. Then, while the location of $r$ indices in the sequence is fixed, $n-r$ indices can be assigned arbitrarily.

The number of permutations where at least $r$ are assigned correctly are then:

$$\begin{pmatrix} n \\ r \end{pmatrix} (n-r)! \tag{4.23}$$

And the probability of at least $r$ people finding their coat is the number of permutations multiplied with the probability of an individual permutation (that is, $\frac{1}{n!}$):

$$p(\# \text{ correct} \geq r) = \begin{pmatrix} n \\ r \end{pmatrix} \frac{(n-r)!}{n!} = \frac{1}{r!} \tag{4.24}$$

Where $r \leq n$.

#### 4.9.1.3 Not lame: Exactly r people find their coat

So, then, what's the probability that exactly nobody finds their coat? What's the probability that 2 people find their coat but nobody else does? What's the probability that $r$ out of $n$ people find their coat? After thinking quickly on your feet for two seconds, you realize that the answer is obviously:

$$p(\# \text{ correct} = r) = \frac{1}{r!}\sum_{s=0}^{n-r}\frac{(-1)^s}{s!} \tag{4.25}$$

The interviewer grunts ambiguously. They never call you back. You never find out why. You can't sleep. You can't eat. You become an anarchist and you declare war on the system.

The first two versions here are what I've come across in interview prep-type materials. I find them a bit annoying, because they represent particular cases that are much simpler than the general case. Applicants in the habit of studying stupid interview questions are rewarded because those answers are easily memorized (false positive). Applicants who intuit the complexity of the general problem, and who don't know the answer beforehand, might become overwhelmed during an interview and fail (false-ish negative).

/ rant

My approach here follows the extraordinary lecture notes https://mast.queensu.ca/ stat455/ by Glen Takahara at the University of Queensland. The difficulty of the problem is that events of a coat being picked up are interrelated: whether one person picks up their correct coat alters the probability of another person also picking up their correct coat. The beauty of this solution is that, rather than messing about with conditional probabilities, it looks at subsets of the event space, and then uses indicator variables to translate set expressions into algebraic expressions.

It uses the properties:

$$I_A = \begin{cases} 1 \text{ if outcome in A} \\ 0 \text{ if outcome in A}^c \end{cases} \tag{4.26}$$

Which means that:

$$\mathbb{E}(I_A) = 1 \times p(A) + 0 \times p(A^c) = p(A) \tag{4.27}$$

To indicator variables for different events can also be combined:

$$I_{A\cap B\cap C\cap ...} = I_A I_B I_C ... \tag{4.28}$$

And indicator variables for the complement can be constructed trivially:

$$I_{A^c} = 1 - I_A \tag{4.29}$$

Let $A_i$ be the region of state space in which guest $i$ grabbed the right coat. Let's say a *particular* subset $\{i\}_r$ of $r$ guests grabs their correct coats (for example, $\{i\}_r = \{5, 9, 11, 24, ...\}_r$), and that the set of remaining $n-r$ guests $\{j\}_{n-r} = \{1, 2, 3, ...\}_n \setminus \{i\}_r$ grab the wrong coat. The area of state space that corresponds to this outcome is:

$$A_{\{i\}_r,\{j\}_{n-r}} = \bigcap_{i\in\{i\}_r} A_i \bigcap_{j\in\{j\}_{n-r}} A_j^c \tag{4.30}$$

The event that the outcome lies within that region of configuration space can be described with an indicator function:

$$I_{A_{\{i\}_r,\{j\}_{n-r}}} = \prod_{\{i\}_r} I_{A_i} \prod_{\{j\}_{n-r}} (1 - I_{A_j}) \tag{4.31}$$

And the probability of those *particular* $r$ people finding their coat is it's expectation value, $p(A_{\{i\}_r,\{j\}_{n-r}}) = \mathbb{E}(I_{A_{\{i\}_r,\{j\}_{n-r}}})$. Good stuff.

The expression above will consist of a bunch of products of indicator variables that describe whether a particular guest wound up with the correct coat. We know that the product of indicator variables corresponds to an indicator variable for the *intersection* of the corresponding subsets of the state space. Explicitly, if it is a product of $s$ indicator variables for some particular set of coats coat $\{k\}_s$:

$$\prod_k I_{A_k} = I_{\bigcap_k A_k} \tag{4.32}$$

And $\mathbb{E}(I_{\bigcap_k A_k}) = p(\bigcap_k A_k)$ is the probability of a particular $s$ coats being picked up correctly, which is $(n-s)!/n!$. (There is no binomial factor, because it's one *specific* set of $s$ coats).

Products of the sort $\prod^n (1-x_i)$ can be expanded:

$$\prod^n (1-x_i) = \sum_{s=0}^{n} (-1)^s \sum_{1\le i_1,\ldots,i_s\le n} x_{i_1} x_{i_2} \ldots x_{i_s} \tag{4.33}$$

The sum $\sum_{1\le i_1,\ldots,i_s\le n}$ is over all possible sets of up to $s$ indices that can be drawn from $\{1,2,3,\ldots,n\}$. A simple example with $n=3$:

$$(1-x_1)(1-x_2)(1-x_3) = \underbrace{1}_{s=0} - \underbrace{(x_1+x_2+x_3)}_{s=1} + \underbrace{(x_1x_2+x_2x_3+x_1x_3)}_{s=2} - \underbrace{(x_1x_2x_3)}_{s=3} \tag{4.34}$$

The number of terms of order $s$ is the amount of ways that $s$ indices can be sampled from $n$ indices, $\left(\begin{array}{c} n \\ s \end{array}\right)$.

Returning to the original problem, then:

$$\begin{aligned} I_{A_{\{i\}_r,\{j\}_{n-r}}} &= \prod_{\{i\}_r} I_{A_i} \prod_{\{j\}_{n-r}} (1-I_{A_j}) \\ &= \sum_{s=0}^{n-r} (-1)^s \underbrace{\sum_{n-r\le j_1,\ldots,j_s\le n}}_{\text{sum of } \left(\begin{array}{c} n-r \\ s \end{array}\right) \text{ terms}} \underbrace{\prod_{\{i\}_r} I_{A_i} \prod_{\{j\}_s} I_{A_j}}_{\text{product of r+s terms}} \end{aligned} \tag{4.35}$$

By linearity of expected value, and using the relationship of the expected value of indicator variables to their probabilities:

$$\begin{aligned} \mathbb{E}(I_{A_{\{i\}_r,\{j\}_{n-r}}}) &= \sum_{s=0}^{n-r} (-1)^s \underbrace{\sum_{n-r\le j_1,\ldots,j_s\le n}}_{\text{sum of } \left(\begin{array}{c} n-r \\ s \end{array}\right) \text{ terms}} p\left(\underbrace{I_{\bigcap_{\{i\}_r} A_i \bigcap_{\{j\}_s} A_j}}_{\text{r+s coats picked up correctly}}\right) \\ &= \sum_{s=0}^{n-r} (-1)^s \left(\begin{array}{c} n-r \\ s \end{array}\right) \frac{(n-r-s)!}{n!} \end{aligned} \tag{4.36}$$

Now, this is already a pretty neat expression for the probability of a *particular* $r$ coats being picked up correctly (i.e. Joe, Mary, and "Hans" picked up the right coat). We don't really care which of the $r$ guests got lucky, though, so since there are $\left(\begin{array}{c} n \\ r \end{array}\right)$ ways of $r$ coats having been picked out correctly, you sum over all of them:

$$\begin{aligned} p(\#\text{ correct} = r) &= \left(\begin{array}{c} n \\ r \end{array}\right) \sum_{s=0}^{n-r} (-1)^s \left(\begin{array}{c} n-r \\ s \end{array}\right) \frac{(n-r-s)!}{n!} \\ &= \sum_{s=0}^{n-r} (-1)^s \frac{n!}{r!(n-r)!} \frac{(n-r)!}{s!(n-r-s)!} \frac{(n-r-s)!}{n!} \\ &= \frac{1}{r!} \sum_{s=0}^{n-r} \frac{(-1)^s}{s!} \end{aligned} \tag{4.37}$$

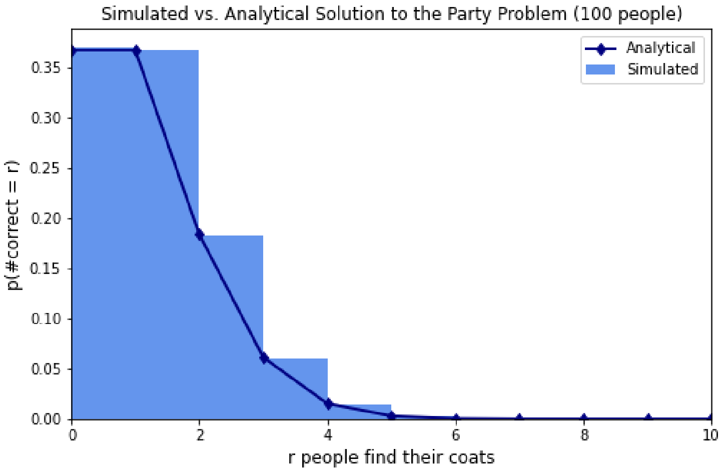


Figure 4.1: Simulated and analytically calculated probabilities that exactly $r$ people at a $n = 100$ people party randomly pick up their coat.

## 4.10 Copulas

## 4.11 Relationships Between Distributions

## 4.12 Large Deviation Theory

### 4.12.1 Gaertner-Ellis Theorem

### 4.12.2 Example: Sum of Uniform Random Variables

## 4.13 Point Mass Distributions

If the entire probability is concentrated at a single point $\alpha$, then that can be expressed in terms of the Kronecker Delta in the discrete case and the Dirac Delta function in the continuous case.

### 4.13.1 Kronecker Delta $\delta_\alpha$

In the discrete case:

$$\delta_\alpha = \begin{cases} 1 \text{ if } x = \alpha \\ 0 \text{ else} \end{cases} \tag{4.38}$$

### 4.13.2 Dirac Delta Function $\delta(x - \alpha)$

In the continuous case:

$$\int_A \delta(x - \alpha) \mathrm{d}x = \begin{cases} 1 \text{ if } \alpha \in A \\ 0 \text{ else} \end{cases} \tag{4.39}$$

The Dirac Delta function is actually a distribution with extremely useful properties that I ought to write up.

## 4.14 Uniform Distributions

### 4.14.1 Discrete Uniform $\mathrm{Uniform(1, k)}$

With $k > 0$ be some integer, the PMF is:

$$f(x) = \begin{cases} 1/k \text{ if } x \in \{1, ..., k\} \\ 0 \text{ else} \end{cases} \tag{4.40}$$

The CDF is:

$$F(x) = \begin{cases} 0 & x < 1 \\ \frac{x}{k} & 1 \leq x \leq k \\ 1 & x > k \end{cases} \tag{4.41}$$

### 4.14.2 Continuous Uniform $\mathrm{Uniform(a, b)}$

With $a, b \in \mathbb{R}$ and $b > a$, the PDF is:

$$f(x) = \begin{cases} 1/(b-a) \text{ if } x \in [a, b] \\ 0 \text{ else} \end{cases} \tag{4.42}$$

The CDF is:

$$F(x) = \begin{cases} 0 & x < a \\ \frac{x-a}{b-a} & a \leq x \leq b \\ 1 & x > b \end{cases} \tag{4.43}$$

## 4.15 Bernoulli Processes

A Bernoulli process is a number of discrete trials with binary outcome, for example a series of coinflips. This is typically framed in terms of success $X = 1$ with probability $p$ and failure $X = 0$ with probability $1-p$. The trials are independent from each other. Thinking of the process as a sequence in time, this means that the process is memoryless: the number of successes or time since the last success have no bearing on the future. Thinking about the process as a sequence in space, the successes are independently scattered over a grid like randomly flipped bits in a string of bits. That is, the events are uniformly distributed.

### 4.15.1 Bernoulli $\mathrm{Bernoulli(p)}$

The PMF of a single binary outcome $X = 1$ with probability $p$ and $X = 0$ with probability $1 - p$. The PMF is:

$$f(x) = p^x(1-p)^{1-x} = \begin{cases} p & x = 1 \\ (1-p) & x = 0 \end{cases} \tag{4.44}$$

For $x \in \{0, 1\}$ and $p \in [0, 1]$.

### 4.15.2 Binomial $\mathrm{Binomial}(n, p)$

The binomial distribution is the PMF for the number of successes with probability $p$ among $n$ trials. That is, if $X_i \sim \mathrm{Bernoulli}(p)$, then $X = \sum_i^n X_i$ has distribution:

$$f(x) = \binom{n}{x} p^x (1-p)^{n-x} \tag{4.45}$$

It follows that if $X_1 \sim \text{Binomial}(n_1, p)$ and $X_2 \sim \text{Binomial}(n_2, p)$, then $X_1 + X_2 \sim \text{Binomial}(n_1 + n_2, p)$. The binomial distribution can be interpreted of the probability that there will be $x$ successes among $n$ draws with replacement.

### 4.15.3 Geometric $\mathrm{Geom}(p)$

The PMF for the number of Bernoulli trials with parameter $p \in (0, 1)$ is given by:

$$f(x) = p(1-p)^{x-1} \tag{4.46}$$

In the time picture, it models the number of intervals $x$ until the first success occurs, or, equivalently, the number of trials between two successes. In the space picture, it models the distance between successes that are independently scattered on a grid.

### 4.15.4 Pascal, Negative Binomial $\mathrm{NB}(r, p)$

The negative binomial distribution with parameters $r$ and $p$ gives the sum of $r$ geometric random variables with parameter $p$. In the time picture, if $X \sim \mathrm{NB}(r, p)$, $X$ gives the probability for the number of failures in a sequence of Bernoulli trials until there are $r$ successes. In the space picture, it the probability for the width of the interval of a grid between $r$ successes (not counting the spots taken up by the successes). Its PMF is given by:

$$f(x) = \binom{x+r-1}{x} p^r (1-p)^x \tag{4.47}$$

For $r = 1$, the distribution is the same as the geometric distribution with parameter $p \to 1 - p$. The negative binomial distribution gives the probability that, when drawing with replacement, it will take $x$ failures until there have been $r$ successes, where each success has probability $p$.

## 4.16 Bernoulli Processes Without Replacement

Bernoulli Processes were a sequence of independent trials, which can be thought of as a sequence of samples with replacement. The "hyper"- distributions treat the analogous case where samples are not replacemed.

### 4.16.1 Hypergeometric $\mathrm{Hypergeom(N, K, n)}$

The hypergeometric distribution gives the probability for the number of successes when drawing $n$ times from a population of $N$, of which $K$ correspond to successes.

$$f(k) = \frac{\binom{K}{k}\binom{N-K}{n-k}}{\binom{N}{k}} \tag{4.48}$$

It is the analogue to the binomial distribution for sampling without replacement.

### 4.16.2 Negative Hypergeometric $\mathrm{NH(N, K, n)}$

The negative hypergeometric distribution gives the probability that, when sampling without replacement, it will take $r$ failures until there have been $k$ successes, if the total population is $N$ and the number of elements corresponding to a success is $K$.

$$f(k) = \frac{\binom{k+r-1}{k}\binom{N-r-k}{K-k}}{\binom{N}{K}} \tag{4.49}$$

The negative hypergeometric distribution is the analogue to the negative binomial distribution for sampling without replacement.

## 4.17 Poisson Point Processes

Poisson point processes are the continuous analogue to Bernoulli processes. Rather than looking at the outcome of a number of binary trials, they look at the number or spacing of independent point events over a continuous interval. In one dimension, Poisson Processes very often model rare events happening over a given time interval. In higher dimensions, the Poisson process can be thought of as independently scattered points in some volume. The process is typically characterized by the parameter $\Lambda$, called the *rate* or *intensity* of the process. It can be written $\Lambda = \nu\lambda$ where $\nu$ is a Lebesgue measure (assigning a length or volume to a set) and $\lambda$ is a constant. For a temporal process, $\nu$ would be the time interval. For a spatial process, $\nu$ would be a volume. The events have uniform distribution over the measure. This is also called a homogenous Poisson process.

### 4.17.1 Poisson $\mathrm{Poisson}(\Lambda)$

The poisson distribution is the continuous analogue to the Binomial distribution. It measures the number of events for a process with intensity $\Lambda$.

$$f(x) = e^{-\Lambda}\frac{\Lambda^x}{x!} \quad \mathrm{x} \geq 0 \tag{4.50}$$

Just like for the Binomial distribution, if $X_1 \sim \mathrm{Poisson}(\Lambda_1)$ and $X_2 \sim \mathrm{Poisson}(\Lambda_2)$, then $X_1 + X_2 \sim \mathrm{Poisson}(\Lambda_1 + \Lambda_2)$.

### 4.17.2 Exponential $\mathrm{Exp}(\lambda)$

The exponential distribution is the continuous analogue to the Geometric distribution. In one dimension, if $X \sim \mathrm{Exp}(\Lambda)$, then $X$ can be interpreted as the distance between two independently scattered points on the real line. In the time picture, that would be the time that elapses between two events, or, equivalently, the time until the next event, where $\lambda$ is the rate or intensity.

$$f(x) = \lambda e^{-\lambda} \tag{4.51}$$

### 4.17.3 Gamma $\mathrm{Gamma}(\alpha, \beta)$

The Gamma Distribution is the continuous analogue to the negative binomial distribution. For integer $\alpha$, it gives the probability of the value of the sum of $\alpha$ exponentially distributed random variables with parameter $\beta = \frac{1}{\lambda}$. That is, if $X_i \sim Exp(\frac{1}{\beta})$ then $X = \sum_i^\alpha X_i \sim \mathrm{Gamma}(\alpha, \beta)$. The PDF is given by:

$$f_x = \frac{1}{\beta^\alpha \Gamma(\alpha)} x^{\alpha-1} e^{-x/\beta} \tag{4.52}$$

With $\alpha, \beta > 0$.

As one would expect, if $X_1 \sim \Gamma(\alpha_1, \beta)$ and $X_2 \sim \Gamma(\alpha_2, \beta)$ then $X_1 + X_2 \sim \Gamma(\alpha_1 + \alpha_2, \beta)$.

## 4.18 $t$-Distribution $\mathrm{t}_\nu$

### 4.18.1 Cauchy Distribution $t_1$

The Cauchy Distribution does not have a mean. It is infinite.

## 4.19 Chi$^2$-Distribution $\chi^2_p$

The $\chi^2$ distribution is the probability distribution of the sum of squares of standard normally distributed random variables. If $Z_i$ has standard normal distribution, then $X = \sum_i^p Z_i \sim \chi^2_p$.

## 4.20 Normal $\mathscr{N}(\mu, \sigma^2)$

### 4.20.1 Standard Normal Distribution $\mathscr{N}(0, 1)$

The normal distribution is so ubiquitous that there is notation specifically for the *standard normal distribution*, which is the normal distribution with parameters $\mu = 0$ and $\sigma = 1$. It is always possible to transform a random variable $X$ to have standard normal distribution by performing a coordinate transform. The PDF of the standard normal distribution is written $\phi(x)$ and the CDF $\Phi(x)$. The PDF is given by:

$$\phi(x) = \frac{1}{\sqrt{2\pi}} \exp -x^2 \tag{4.53}$$

The CDF has no closed form expression.

By convention, $Z$ is the random variable that has standard normal distribution. If $X \sim \mathscr{N}(\mu, \sigma^2)$, then $Z = \frac{X-\mu}{\sigma} \sim \mathscr{N}(0, 1)$. This standardizing transformation allows for calculations using lookup tables.

#### 4.20.1.1 Example: Interval $\mathbb{P}(a < X < b)$

$$\mathbb{P}(a < X < b) = \mathbb{P}(a' < Z < b') = \Phi(b') - \Phi(a') \tag{4.54}$$

with $a' = \frac{a-\mu}{\sigma}$ and $b' = \frac{b-\mu}{\sigma}$.

#### 4.20.1.2 Example: Quantile $x = F^{-1}(q)$

$$z = \Phi^{-1}(q) \tag{4.55}$$

with $z = \frac{x-\mu}{\sigma}$, so that $x = \sigma\Phi^{-1}(q) + \mu$. The interquartile range is calculated by transforming $(z_1 = \Phi^{-1}(0.25), z_3 = \Phi^{-1}(0.75))$, and so on.

## 4.21 Log Normal Distribution

The log normal distribution has the odd property that is “thin tailed" for low variance and “fat tailed" for high variance.

## 4.22 Categorial Processes, Multinoulli Processes

A Multnoulli process is a number of discrete trials that may assume a number of different categorical outcomes, for example a series of dice throws or a random sequence of letters. It is the generalization of the Bernoulli process to processes with more than two possible outcomes. The trials are independent from each other. Thinking of the process as a sequence in time, this means that the process is memoryless: the sequence of future outcomes is independent of past outcomes. Thinking about the process as a sequence in space, it is a uniformly random assignment of one of a set number of outcomes to grid points.

### 4.22.1 Categorical, Multinoulli $\mathrm{Categorical}(\mathbf{p})$

The categorial or Multinoulli distribution is the generalization of the Bernoulli distribution to more than two possible outcomes. The PMF is given by:

$$f(\mathbf{x}) = \prod_{i}^{k} p_i^{[x=k]} \tag{4.56}$$

With $p_i \in \mathbf{p}$ a discrete probability distribution and $[x = k]$ is the Iverson bracket.

### 4.22.2 Multinomial $\mathrm{Multinomial}(\mathrm{n}, \mathbf{p})$

The multinomial distribution is the multivariate generalization of the binomial process. Rather than a binary outcome, there are $k$ possible outcomes. If $\mathbf{X} = (X_1, X_2, ..., X_k)$ has multinomial distribution with parameter vector $\mathbf{p} = (p_1, p_2, ..., p_k)$, then it gives the probability that out of $n$ trials, $x_j$ will be of type $j$, where the probabilities of the different classes is given by the discrete probability distribution $\mathbf{p}$.

$$f(\mathbf{x}) = \binom{n}{x_1, x_2, ..., x_k} p_1^{x_1} p_2^{x_2} ... p_k^{x_k} = \binom{|\mathbf{x}|}{\mathbf{x}} \mathbf{p}^{\mathbf{x}} \tag{4.57}$$

Where the final expression uses multi-index notation. The binomial process corresponds to $k = 2$ classes (success and failure). The marginal distribution of $X_j$ is $Binomial(n, p_j)$.

## 4.23 Beta $\mathrm{Beta}(\alpha, \beta)$

The PDF for the beta distribution is:

$$f(x) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1} \;\; 0 < x < 1 \tag{4.58}$$

With $\alpha, \beta > 0$.

The Beta distribution is the conjugate prior to the Bernoulli and Binomial distributions.

## 4.24 Dirichlet $\mathrm{Dirichlet}(\alpha)$

The Dirichlet Distribution is the multivariate generalization of the Beta Distribution. Given $\mathbf{X} = (X_1, ..., X_K)$ and the parameter vector $\alpha = (\alpha_1, ..., \alpha_K), \;\alpha_j > 0$, the PDF is:

$$f(\mathbf{x}, \alpha) = \frac{1}{B(\alpha)} \prod_{i=1}^{K} x_i^{\alpha_i - 1} \tag{4.59}$$

Where $B(\alpha)$ is the multivariate Beta function:

$$B(\alpha) = \frac{\prod_{i=1}^{K} \Gamma(\alpha_i)}{\Gamma(\sum_{i=1}^{K} \alpha_i)} \tag{4.60}$$

The Dirichlet distribution is the conjugate prior to the Multinoulli/Categorical and Multinomial distributions. The marginal distributions are Beta distributions, i.e. $X_i \sim \mathbf{Beta}(\alpha_i, \alpha_0 - \alpha_1)$ where $\alpha_0 = \sum_j \alpha_j$.

## 4.25 Multivariate Normal $\mathcal{N}(\mu, \Sigma)$

The multivariate normal distribution is the multivariate generalization of the normal distribution. The PDF is given by:

$$f\mathbf{x} = \frac{1}{(2\pi)^{\frac{k}{2}} |\Sigma|^{\frac{1}{2}}} \exp\{-\frac{1}{2}(\mathbf{x} - \mu)^T \mathbf{\Sigma}^{-1}(\mathbf{x} - \mu)\} \tag{4.61}$$

### 4.25.1 Covariance Matrix $\mathbf{\Sigma}$

The covariance matrix $\mathbf{\Sigma}$ is symmetric and positive definite. Therefore, $\mathbf{\Sigma}^{\frac{1}{2}}$ exists and is real, symmetric, and

- $\mathbf{\Sigma}^{\frac{1}{2}}\mathbf{\Sigma}^{\frac{1}{2}} = \mathbf{\Sigma}$
- $\mathbf{\Sigma}^{\frac{1}{2}}\mathbf{\Sigma}^{-\frac{1}{2}} = \mathbf{\Sigma}^{-\frac{1}{2}}\mathbf{\Sigma}^{\frac{1}{2}} = \mathbf{I}$

### 4.25.2 Marginal Distribution

If $\mathbf{X} \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma})$, then the marginal distribution of $X_a$ is:

$$X_a \sim \mu, \tag{4.62}$$

### 4.25.3 Conditional Distribution

If $\mathbf{X} \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma})$, then the conditional distribution of $X_b$ given $X_a = x_a$ is:

$$X_b|X_a = x_a \sim \mathcal{N}(\mu_b + \Sigma_{ba}\Sigma_{aa}^{-1}(x_a - \mu_a), \Sigma_{bb} - \Sigma_{ba}\Sigma_{aa}^{-1}\Sigma_{ab}) \tag{4.63}$$

### 4.25.4 Vector Multiplication

If a is a vector, then:

$$\mathbf{a}^T\mathbf{X} \sim \mathcal{N}(\mathbf{a}^T\mu, \mathbf{a}^T\mathbf{\Sigma}\mathbf{a}) \tag{4.64}$$

### 4.25.5 Relationship to Chi$^2$

$$V = \mathbf{Z}^T\mathbf{Z} = (\mathbf{X} - \mu)^T\mathbf{\Sigma}^{-1}(\mathbf{X} - \mu) \sim \chi_k^2 \tag{4.65}$$

### 4.25.6 Multivariate Standard Normal Distribution $\mathcal{N}(0, \mathbf{\Sigma})$

The standard multivariate normal distribution for $\mathbf{Z} = (Z_1, ..., Z_k)$ with $\mathbf{Z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ is:

$$f(z) = \prod_{i=1}^{k} f(z_i) = \frac{1}{(2\pi)^{\frac{k}{2}}} \exp\{\frac{1}{2}\sum_{j=1}^{k} z_j\} = \frac{1}{(2\pi)^{\frac{k}{2}}} \exp\{\frac{1}{2}\mathbf{z}^{\mathbf{T}}\mathbf{z}\} \tag{4.66}$$

The transformation to a general multivariate normal random variable $\mathbf{X}$ is that if $\mathbf{Z} \sim\ , \mathbf{I}$, then $\mathbf{X} = \mu + \mathbf{\Sigma}^{\frac{1}{2}}\mathbf{Z}$. Conversely, if $\mathbf{X} \sim \mu, \mathbf{\Sigma}$ then $\mathbf{\Sigma}^{-\frac{1}{2}}(\mathbf{X} - \mu) \sim \mathcal{N}(0, \mathbf{I})$.

## 4.26 Expectations

### 4.26.1 Law of the Unconscious Statistician

Given a random variable $X$ with distribution $f(x)$ and some function $r(x)$, the expected value $\mathbb{E}(r)$ is given by:

$$\mathbb{E}(r) = \int r(x)\mathrm{d}F_X(x) \tag{4.67}$$

### 4.26.2 Linearity of Expeced Value

$$\mathbb{E}\left(\sum_i a_i X_i\right) = \sum_i a_i \mathbb{E}(X_i) \tag{4.68}$$

## 4.27 Moments, Central Moments

The $k$th moment is $\mathbb{E}X^k$. The $k$th central moment is $\mathbb{E}(X-\mu)^k$.

## 4.28 Variance

In one dimension:

$$\sigma^2 = \mathbb{V}(X) = \mathbb{E}(X-\mu)^2 = \int (x-\mu)^2 \mathrm{d}F(x) \tag{4.69}$$

The standard deviation is $\mathrm{sd}(X) = \sqrt{\mathbb{V}(X)}$.

- $\mathbb{V}(X) = \mathbb{E}(X^2) - \mu^2$.
- If $a$ and $b$ are constants, $\mathbb{V}(aX+b) = a^2\mathbb{V}(X)$.
- If $X_1, ..., X_n$ are independent and $a_1, ..., a_n$ are constants, then $\mathbb{V}\left(\sum_{i=1}^n a_i X_i\right) = \sum_{i=1}^n a_i^2 \mathbb{V}(X_i)$

## 4.29 Mean and Variance of Vector Valued Random Variables

The mean is simply $\mathbb{E}(\mathbf{X}) = (..., \mathbb{E}(X_i), ..., )$. In higher dimensions, the variance is a matrix with entries $\Sigma_{i,j} = \mathrm{Cov}(X_i, X_j)$.

- $\mathbb{E}(\mathbf{a}^T\mathbf{X}) = \mathbf{a}^T\mu$
- $\mathbb{E}(\mathbf{A}\mathbf{X}) = \mathbf{A}\mu$
- $\mathbb{V}(\mathbf{a}^T\mathbf{X}) = \mathbf{a}^T\mathbf{\Sigma}\mathbf{a}$
- $\mathbb{V}(\mathbf{A}\mathbf{X}) = \mathbf{A}\mathbf{\Sigma}\mathbf{A}^T$

## 4.30 Covariance, Correlation

$$\mathrm{Cov}(X,Y) = \mathbb{E}(XY) - \mathbb{E}X\mathbb{E}Y \tag{4.70}$$

## 4.31 Sample Mean and Sample Variance

If $X_1, ..., X_n$ are random variables, the *sample mean* is:

$$\overline{X}_n = \frac{1}{n}\sigma_i X_i \tag{4.71}$$

The *sample variance* is:

$$S_n^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i - \overline{X}_n)^2 \tag{4.72}$$

The sample mean and sample variance are random variables. The mean and variance are fixed properties of the underlying distribution.

$$\begin{aligned}\mathbb{E}(\overline{X}_n) &= \mu \\ \mathbb{V}(\overline{X}_n) &= \tfrac{\sigma^2}{n} \\ \mathbb{E}(S_n^2) &= \sigma^2\end{aligned} \tag{4.73}$$

## 4.32 Law of Iterated Expectation

Given two random variables $X$ and $Y$, the law of iterated expectation states:

$$\mathbb{E}(X) = \mathbb{E}\mathbb{E}(X|Y) \tag{4.74}$$

### 4.32.1 Applied to Variance

$$\mathbb{V}(Y) = \mathbb{E}\mathbb{V}(Y|X) + \mathbb{V}\mathbb{E}(Y|X) \tag{4.75}$$

## 4.33 Moment Generating Functions, Laplace Transforms

The moment generating function (MGF) of a random variable $X$ is:

$$\psi_X(t) = \mathbb{E}(e^{tX}) = \int e^{tx} dF(x) \tag{4.76}$$

The MGF allows for interchanging the operations of differentiation and "taking expectation".

## 4.34 Cumulant Generating Functions

The cumulants provide an alternative to the moments of the distribution.

$$K_X(t) = \log \mathbb{E}(e^{tX}) = \log \int e^{tx} dF(x) \tag{4.77}$$

Then the $n$th cumulant $\kappa_n = K_X^{(n)}(0)$ is the $n$th derivative of the cumulant generating function. The first cumulant is the mean, the second cumulant is the variance.

## 4.35 Characteristic Function

$$C_X(t) = \mathbb{E}(e^{itX}) = \int e^{itx} dF(x) \tag{4.78}$$

The characteristic function has the advantage that it is well-defined for all real values of $t$ even when $\mathbb{E}e^{tX}$ is not well defined.

## 4.36 Inequalities for Probabilities

### 4.36.1 Markov's Inequality

For any $t > 0$:

$$\mathbb{P}(X > t) \leq \frac{\mathbb{E}(X)}{t} \tag{4.79}$$

This is a general result, independent of the distribution of $X$.

### 4.36.2 Chebyshev's Inequality

$$\mathbb{P}(|X-\mu| \geq t) \leq \frac{\sigma^2}{t^2} \tag{4.80}$$

and

$$\mathbb{P}(|Z| \geq k) \leq \frac{1}{k^2} \tag{4.81}$$

where $Z = (X-\mu)/\sigma$. This is a general result that can be proven using Markov's inequality.

### 4.36.3 Hoeffding's Inequality

Let $Y_1, ..., Y_n$ be independent observations so that $\mathbb{E}Y_i = 0$ and $a_i \leq Y_i \leq b_i$. Then, for any $t > 0$,

$$\mathbb{P}\left(\sum_{i=1}^{n} Y_i \geq \epsilon\right) \leq e^{-t\epsilon} \prod_{i=1}^{n} e^{t^2(b_i-a_i)^2/8} \tag{4.82}$$

Hoeffding's inequality provides a sharper bound than Markov's inequality and relies on bounds of the random variables rather than their variance.

#### 4.36.3.1 Example: Bernoulli Random Variables

Let $X_1, ..., X_n \sim \text{Bernoulli}(p)$. Assume that the goal is to measure the parameter $p$, which is estimated by the sample mean. This could for example be the error rate of a binary classifier. For the Bernoulli Random variable, $0 \leq X_1 \leq 1$, so that the probability that the estimate of the error rate is off by more than $\epsilon$ is given by:

$$\mathbb{P}(|\bar(X)_n - p| > \epsilon) \leq 2e^{-2n\epsilon^2} \tag{4.83}$$

### 4.36.4 Mill's Inequality

Mill's inequality is useful for normal random variables. Let $Z \sim \mathcal{N}(0,1)$, then the probability that $|Z|$ will be greater than some value $t$ is bounded by:

$$\mathbb{P}(|Z| > t) \leq \frac{2}{\pi} \frac{e^{-t^2/2}}{t} \tag{4.84}$$

## 4.37 Inequalities for Expectations

### 4.37.1 Cauchy-Schwarz Inequality

$$\mathbb{E}|XY| \leq \sqrt{\mathbb{E}(X^2)\mathbb{E}(Y^2)} \tag{4.85}$$

### 4.37.2 Jensen's Inequality

Jensen's inequality provides bounds for expectations for non-linear functions of a random variable. Given a random variable $X$ and a function $g(X)$, if $g(X)$ is convex:

$$\mathbb{E}g(X) \geq g(\mathbb{E}X) \tag{4.86}$$

If $g(X)$ is concave:

$$\mathbb{E}g(X) \leq g(\mathbb{E}X) \tag{4.87}$$

## 4.38 Asymptotic Theory

Asymptotic theory, large sample theory, or limit theory, deals with the question of the limiting behavior of sequences of random variables.

The notion of convergence for random variables is more involved than is usual for, for example, a sequence of numbers or functions in calculus. There are different types of convergence, and they do not necessarilty imply one another.

The chain of implication is:

$$\text{quadratic mean} \rightarrow \text{probability} \rightarrow \text{distribution} \tag{4.88}$$

For point mass distributions only, convergence in distribution implies convergence in probability. Similarlty, implication rules exist for functions of random variables when convergence for the underlying random variables is given. See Theorems 5.4, 5.5 and 5.17 in Wasserman (2013) (pages 73-74, 81).

### 4.38.1 Preasymptotics

The behavior in the intermediate regime, for example where $n$ large but not yet asymptotic. Asymptotic behavior kicks in at very different rates for different underlying processes. The “baseline" in terms of speed is represented by normally distributed random variables.

### 4.38.2 Convergence in Probability

$$\mathbb{P}(|X_n - X| > \epsilon) \rightarrow 0 \tag{4.89}$$

as $n \rightarrow \infty$. Convergence in Probability means that the distribution of $X_n$ becomes sharper and sharper around $X$ as $n \rightarrow \infty$. At $n = \infty$, it has point mass distribution concentrated at $X$.

### 4.38.3 Convergence in Distribution

Where $F$ is the CDF of $X_n$,

$$\lim_{n \rightarrow \infty} F_n(t) = F(t) \tag{4.90}$$

For all $t$ for which $F$ is continuous. That means, convergence is satisfied even when the equality is violated at points of discontinuity.

### 4.38.4 Convergence in Quadratic Mean, Convergence in $L_2$

$$\mathbb{E}(X_n - X)^2 \rightarrow 0 \tag{4.91}$$

as $n \rightarrow \infty$.

### 4.38.5 Almost Sure Convergence

$X_n$ converges *almost surely* to $X$ if:

$$\mathbb{P}(\{\omega : X_n(\omega) \rightarrow X(\omega)\}) = 1 \tag{4.92}$$

Which I would read as the value of the measurable map $X_n$ converging to the value of the measurable map $X$ everywhere on the probability space except possibly on a set of probability measure $0$.

### 4.38.6 $L_1$ Convergence

$L_1$ convergence requires $\mathbb{E}|X_n - X| \to 0$ as $n \to 0$.

### 4.38.7 Weak Law of Large Numbers

The sample mean of i.i.d. variables $\overline{X}_n$ converges in probability to the mean of the distribution $\mu$i. It can be proven with Chebyshev's inequality that:

$$\mathbb{P}(|\overline{X}_n - \mu| > \epsilon) \leq \frac{\mathbb{V}(\overline{X}_n)}{\epsilon^2} = \frac{\sigma^2}{n\epsilon^2} \tag{4.93}$$

In words, the probability that the sample mean deviates from the population mean by more than $\epsilon$ has an upper bound that decreases inversely proportional to $n\epsilon^2$. The probability becomes more and more centered around the mean $\mu$.

### 4.38.8 Strong Law of Large Numbers

The strong law of large number gives almost surely convergence of the sample mean to the population mean.

If $\mathbb{E}|X_1| < \infty$, then $\overline{X_n} \xrightarrow{as} \mu$.

### 4.38.9 Central Limit Theorem

The distribution of the sample mean converges in distribution to a normal distribution with variance $\sigma^2/n$ and mean $\mu$.

If $Z_n = \frac{\overline{X}_n - \mu}{\sqrt{\mathbb{V}(\overline{X}_n)}}$ then $\lim_{n\to\infty} \mathbb{P}(Z_n \leq z) = \Phi(z)$, where $\Phi(z)$ is the CDF of a standard normal distribution.

It turns out that when $Z_n$ is obtained by normalizing not by the (most likely unknown) population variance $\sigma$ but by the sample variance $S_n^2$, the CLT still holds. The accuracy of this is given by the Berry-Esséen Inequality.

### 4.38.10 Multivariate Central Limit Theorem

Given $\mathbf{X_1}, ..., \mathbf{X_n}$ i.i.d random vectors where each vector:

$$\mathbf{X_i} = \begin{pmatrix} X_{1i} \\ X_{2i} \\ \vdots \\ X_{ki} \end{pmatrix} \tag{4.94}$$

Then the population mean:

$$\mu = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_k \end{pmatrix} = \begin{pmatrix} \mathbb{E}(X_{i1}) \\ \mathbb{E}(X_{i2}) \\ \vdots \\ \mathbb{E}(X_{ik}) \end{pmatrix} \tag{4.95}$$

The variance is given by the matrix $\Sigma$ as before. The sample mean:

$$\overline{\mathbf{X}} = \begin{pmatrix} \overline{X_1} \\ \overline{X_2} \\ \vdots \\ \overline{X_k} \end{pmatrix} \tag{4.96}$$

Then $\sigma^{-\frac{1}{2}}(\overline{X} - \mu)$ converges in distribution ot $\mathcal{N}(0,1)$.

### 4.38.11 Proof of the Central Limit Theorem

Wasserman (2013), page 81.

Given i.i.d random variables $X_i$, the transformation $Y_i = \frac{X_i - \mu}{\sigma}$ gives i.i.d. random variables with zero mean and unit variance. Let $\psi(t)$ be the MGF of $Y_i$. Since $Y_i$ are i.i.d., the sum $\sum_{i=1}^{n} Y_i$ has MGF $\psi(t)^n$. The normalized sample mean $Z_n = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} Y_i$ has MGF $\Xi_n(t) = \psi(t/\sqrt{n})^n$. Two random variables that have the same MGF in an open interval about the point $t = 0$ have the same distribution, probably because the Laplace transform is injective. Therefore, if $\psi_n(t) \to \psi(t)$ in some open interval around $t = 0$, then their underlying random variables $Z_n \xrightarrow{dist} Z_n$ converge in distribution. Taking the Taylor expansion of $\epsilon_n(t)$:

$$\epsilon_n(t) = \left(1 + 0 + \frac{t^2}{2!n} + ...\right)^n \to e^{t^2/2} \tag{4.97}$$

Which is the MGF of $\mathcal{N}(0,1)$

### 4.38.12 Delta Method

The delta method allows statements regarding the convergence of functions of random variables, whenever the input random variable converges in distribution to a normal distribution.

If $Y_n$ has a limiting normal distribution, and $g(Y_n)$ is a smooth function so that $g'(\mu) \neq 0$, then if:

$$\frac{\sqrt{n}(Y_n - \mu)}{\sigma} \xrightarrow{dist} \mathcal{N}(0,1) \tag{4.98}$$

Then:

$$\frac{\sqrt{n}(g(Y_n) - g(\mu))}{|g'(\mu)|\sigma} \xrightarrow{dist} \mathcal{N}(0,1) \tag{4.99}$$

Rewriting, if $Y_n \xrightarrow{dist} \mathcal{N}(\mu, \frac{\sigma^2}{n})$ then $g(Y_n) \xrightarrow{dist} \mathcal{N}(g(\mu), (g'(\mu))^2 \frac{\sigma^2}{n})$.

#### 4.38.12.1 Multivariate Delta Method

If $\mathbf{Y_n} \xrightarrow{dist} \mu, \mathbf{\Sigma}$ then the scalar-valued function $g(\mathbf{Y_n}) \xrightarrow{dist} \mathcal{N}(g(\mu), \frac{1}{n}(\nabla g(\mu))^T \mathbf{\Sigma} (\nabla g(\mu)))$.

Use case would be functions of several sample means, where the underlying samples have non-trivial covariance (cf. (Wasserman 2013), p. 80).

## 4.39 Classes of Distributions

### 4.39.1 Stable Distributions

Let $X_1, X_2, ...X_n$ be i.i.d. random variables. Let $S_n = \sum_i X_i$. $X_s$ has stable distibution if, for sufficiently large $n$, the (properly normalized) sum $S_n$ converges in distribution to the distribution of $X_s$.

$$\frac{S_n - \alpha_n}{\beta_n} \xrightarrow{dist} X_s \tag{4.100}$$

### 4.39.2 Sub-exponential

### 4.39.3 Exponential

### 4.39.4 Elliptical

# 5 Information Theory

**Contents of this chapter**

## 5.1 Entropy

## 5.2 Mutual Information

## 5.3 Kullback-Leibler Divergence

# 6 Stochastic Processes and Time Series Analysis

**Contents of this chapter**

In the discrete case, time series $X$ is a set of random variables $X = \{X_1, X_2, X_3, ...\}$, or alternatively $X = \{X_t : t \in T\}$ where $T$ is the index set.

## 6.1 Branching Processes

## 6.2 Markov Chains

Markov Chains are time series in which the probability of a state depends only on the state preceding it. That is:

$$p(X_{t+1}|X_t, Xt-1, ..., X_1, X_0) = p(X_{t+1}|X_t) \tag{6.1}$$

## 6.3 Martingales

Martingales are time series in which the expected value of a state is identical to the expected value of the state preceding it. That is:

$$\mathbb{E}X_{t+1} = \mathbb{E}X_t \tag{6.2}$$

### 6.3.1 Martingale Convergence Theorem

## 6.4 Hidden Markov Models

## 6.5 Ito Calculus

## 6.6 Chaos Analysis

## 6.7 Noise

### 6.7.1 White Noise

White noise contains all frequencies with the same intensity. That is, it has constant power spectral density. Infinite bandwidth white noise is a purely theoretical concept, but if noise has a flat power spectrum across the frequencies of interest, it is referred to as white noise.

In discrete time, white noise is a series of uncorrelated random variables with zero mean and finite variance. When the variables all have normal distribution, the noise is referred to as Gaussian white noise.

### 6.7.2 Gaussian Noise

Gaussian noise is typically understood to be white noise in which the shocks have a normal distribution.

### 6.7.3 Brownian Noise

Brownian noise is noise corresponding to random motion (i.e. Brownian motion). The power spectrum is proportional to $\frac{1}{f^2}$.

## 6.8 Empirical Processes

## 6.9 Mean Field Theory, Fluid Approximation, Deterministic Approximation

In modeling phenomena with Markov Chain's, the number of elements constituting a system and the possible local states of that system may grow very large. This so-called *state space explosion* is sometimes approached by approximating the system with a continuous state space. Those approaches include *fluid approximation, mean field approximation* or sometimes the *deterministic approximation.*

## 6.10 Backtesting

# 7 Statistical Inference

## Contents of this chapter

## 7.1 Parametric and Nonparametric Models

A statistical model $\mathfrak{F}$ is a family of functions. *Parametric models* can be parametrized by a finite number of parameters. For example, the Gaussian Normal distribution $\mu, \sigma$ has the two parameters $\mu$ and $\sigma^2$. The general form is:

$$\mathfrak{F} = \{f(x;\theta) : \theta \in \Theta\} \tag{7.1}$$

Where $\theta$ is a vector of parameters and $\Theta$ is the parameter space. Elements of $\theta$ that are not of interest are called *nuisance parameters*. *Nonparametric models* cannot be described by a finite number of parameters. An example is an interpolating spline.

## 7.2 Fundamental Concepts in Inference

### 7.2.1 Point Estimators

Point estimation refers to providing a single best guess of some quantity of interest. The point estimator is denoted with a had, i.e. $\hat{\theta}$. The point estimator for some quantity based on $n$ datapoints:

$$\hat{\theta}_n = g(X_1, X_2, ..., X_n) \tag{7.2}$$

### 7.2.2 Bias

A point estimator $\hat{\theta}$ has bias:

$$\text{bias}(\hat{\theta}_n) = \mathbb{E}_\theta(\hat{\theta}) - \theta \tag{7.3}$$

Where $\theta$ is the "true" value $\theta_n \to \theta$ as $n \to \infty$. $\mathbb{E}_\theta(r(X)) = \int r(x)f(x;\theta)\mathrm{d}x$.

### 7.2.3 Consistency

A point estimator $\hat{\theta}_n$ is consistent if $\hat{\theta}_n \xrightarrow{P} \theta$. That is, it converges in probability to $\theta$. $\hat{\theta}_n$ is consistent with both bias and standard error approach 0 as $n \to \infty$.

### 7.2.4 Sampling Distribution, Standard Error

The distribution of $\hat{\theta}_n$ is the *sampling distribution.* The standard deviation of the distribution of $\hat{\theta}_n$ is the *standard error,* denoted se.

$$\text{se}(\hat{\theta}_n) = \sqrt{\mathbb{V}(\hat{\theta}_n)} \tag{7.4}$$

The standard error might depend on the unknown population CDF, in which case it is estimated. The point estimator for the standard error is then $\hat{\text{se}}$.

#### 7.2.4.1 Example: Bernoulli Distribution

Let $X_1, X_2, X_3 \sim \text{Bernoulli}(p)$. The point estimator for $p$ based on $n$ datapoints is $\hat{p}_n = \frac{1}{n}\sum_{i=1}^n X_i$. Then the expected value of the point estimator $\mathbb{E}(\hat{p}_n) = \frac{1}{n}\sum_{i=1}^n \mathbb{E}(X_i) = p$, so that $\hat{p}_n$ is unbiased. The standard error is $\text{se} = \sqrt{\mathbb{V}(\hat{p}_n)} = \sqrt{\frac{p(1-p)}{n}}$. The estimated standard error is $\hat{\text{se}} = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$.

### 7.2.5 Mean Squared Error

The quality of a point estimator is often measured using the *mean squared error*:

$$\text{MSE} = \mathbb{E}_\theta(\hat{\theta}_n - \theta)^2 = \text{bias}^2(\hat{\theta}_n) + \mathbb{V}_\theta(\hat{\theta}_n) \tag{7.5}$$

### 7.2.6 Asymptotically Normal Estimators

An asymptotically normal estimator satisfies:

$$\frac{\hat{\theta}_n - \theta}{\text{se}} \xrightarrow{dist} \mathcal{N}(0,1) \tag{7.6}$$

### 7.2.7 Confidence Sets

A confidence set $C_n$ is the subset of parameters $\theta$ that has a greater than $1-\alpha$ probability of containing the true value of $\theta$.

$$\mathbb{P}_\theta(\theta \in C_n) \geq 1-\alpha \text{ for all } \theta \in \Theta \tag{7.7}$$

#### 7.2.7.1 Normal-Based Confidence Intervals

If $\hat{\theta}_n \approx \mathscr{N}(\theta, \hat{\text{se}}^2)$ and $z_{\frac{\alpha}{2}} = \Phi^{-1}(1-\frac{\alpha}{2})$ the value of the standard normally distributed random variable $Z$ at which $\mathbb{P}(-\frac{\alpha}{2} < Z < \frac{\alpha}{2}) = 1-\alpha$, then, transforming backwards, $\mathbb{P}(\hat{\theta}_n - z_{\frac{\alpha}{2}}\hat{\text{se}} < \theta < \hat{\theta}_n + z_{\frac{\alpha}{2}}\hat{\text{se}}) = 1-\alpha$. Hence, the confidence interval for a normally distributed point estimator $\hat{\theta}_n$ is:

$$C_n = (\hat{\theta}_n - z_{\frac{\alpha}{2}}\hat{\text{se}}, \hat{\theta}_n + z_{\frac{\alpha}{2}}\hat{\text{se}}) \tag{7.8}$$

For a 95% confidence interval $\alpha = 0.05$ and $z_{\frac{\alpha}{2}} = 1.96 \approx 2$, so that the confidence interval is approximately $\hat{\theta}_n \pm 2\hat{\text{se}}$.

#### 7.2.7.2 Pointwise and Uniform Asymptotic Confidence Intervals

A *pointwise asymptotic* confidence interval requires:

$$\liminf_{n\to\infty} \mathbb{P}_\theta(\theta \in C_n) \geq 1-\alpha, \quad \forall\theta \in \Theta \tag{7.9}$$

A *uniform asymptotic* confidence interval requires:

$$\lim_{n\to\infty} \inf_{\theta\in\Theta} \mathbb{P}_\theta(\theta \in C_n) \geq 1-\alpha \tag{7.10}$$

### 7.2.8 Pivots, Pivotal Quantities

Pivotal quantities are functions of the sample that are independent of the distribution parameters of the sample. For example, if $X = (X_1, X_2, ..., X_n)$ is a random sample from a distribution with parameters $\theta$, then the random variable $g(X, \theta)$ is a pivot if it has the same distribution regardless of the choice of $\theta$. An example is the $z$-score $z = \frac{x-\mu}{\sigma}$. The $z$-score may require population parameters to be known (in this case $\mu$,$\sigma$) but it's distribution is independent of them.

## 7.3 Non-Parametric Estimation of the CDF and Statistical Functionals

### 7.3.1 Empirical Distribution Function

It may be necessary to perform non-parametric estimation of the CDF $F$ of a set of random variables $X_1, X_2, ..., X_n \sim F$.

The empirical distribution function $\hat{F}_n$ is the CDF that puts mass $1/n$ at each point $X_i$.

$$\hat{F}_n(x) = \frac{\sum_{i=1}^n I(X_i \leq x)}{n} \tag{7.11}$$

Where $I(X_i \leq x) = \begin{cases} 1 \text{ if } X_i \leq x \\ 0 \text{ if } X_i > x \end{cases}$.

The empirical CDF is discrete, even when the random variable it is based on may be continuous.
At a given point $x$, $\hat{F}_n(x)$ is an unbiased estimator of $F(x)$.

- $\mathbb{E}\hat{F}_n(x) = F(x)$

- $\mathbb{V}\hat{F}_n(x) = 0 + \text{MSE} = \frac{F(x)(1-F(x))}{n}$
- $\hat{F}_n(x) \xrightarrow{P} F(x)$

The Glivenko-Cantelli Theorem guarantees that, if $X_1, X_2, ..., X_n \sim F$, then:

$$\sup_x |\hat{F}_n(x) - F(x)| \xrightarrow{P} 0 \tag{7.12}$$

### 7.3.2 Confidence Measures for the Empirical CDF

A confidence interval for the empirical CDF is given through the Dvoretzky-Kiefer-Wolfowitz (DKW) Inequality:

$$\mathbb{P}\left(\sup_x |F(x) - \hat{F}_n(x)| > \epsilon\right) \leq 2e^{-2n\epsilon^2} \tag{7.13}$$

A nonparametric $1 - \alpha$ confidence band is then:

$$\begin{aligned} L(x) &= \max\{\hat{F}_n - \epsilon_n, 0\} \\ U(x) &= \min\{\hat{F}_n + \epsilon_n, 1\} \end{aligned} \tag{7.14}$$

with $\epsilon_n = \sqrt{\frac{1}{2n}\log\left(\frac{2}{\alpha}\right)}$.

$$\mathbb{P}\left(L(x) \leq F(x) \leq U(x)\right) \geq 1 - \alpha \tag{7.15}$$

### 7.3.3 Statistical Functionals

A functional is, roughly speaking, a function of a function. The fourier transform of a function is a functional. A *statistical functional* is any function of the CDF, $F$. Examples are the mean, the variance, or the median.

### 7.3.4 Plug-in Estimator

The plug-in estimator of $\theta = T(F)$ is given by $\hat{\theta} = T(\hat{F}_n)$. In other words, the estimate of the CDF is used instead of the true $F$, resulting in an estimator.

### 7.3.5 Linear Functionals

Functionals of the form $T(F) = \int r(x)dF(x)$ are linear functionals.

### 7.3.6 Plug-in Estimator for Linear Functionals

$$T(\hat{F}_n) = \int r(x)d\hat{F}_n(x) = \frac{1}{n}\sum_{i=1}^{n} r(X_i) \tag{7.16}$$

### 7.3.7 Examples: Mean, Variance, Sample Variance, Sample Correlation

Wasserman (2013) pp. 100

## 7.4 The Bootstrap

Bootstrap allows for investigating the variance of a statistic by calculating the statistic many times on samples drawn from the empirical CDF. This amounts to sampling from the dataset with replacement (i.e. datapoints may be included in a sample several times).

There are two approximations in play: first, the true CDF is approximated using the empirical CDF. Second, the variance of the statistic is estimated based on a sample.

Uncertainty in terms of bootstrap estimates may be calculated in terms of normal, pivotal or percentile intervals. The normal interval is valid when the statistic is approximately normally distributed.

## 7.5 Jackknife

The Jackknife method is less computationally expensive, but less general than the bootstrap. For a dataset of $n$ elements, the Jackknife method calculates the statistic $n$ times, each time removing one of the estimates from the calculation. The jackknife does not produce consistent estimators of the standard error of sample quantiles.

## 7.6 Parametric Inference

In parametric inference, the quantity of interest might be some function $T(\theta)$. The sought-after parameters are *parameters of interest* and additional parameters that emerge as part of the model are *nuisance parameters*. Parametric inference deals with creating parametric estimators.

### 7.6.1 Method of Moments

The method of moments relies on a system of linear equations to link estimators of moment to sample moments. The $j$th moment is given by:

$$\alpha_j = \int x^j dF_\theta(x) \tag{7.17}$$

The $j$th sample moment is given by:

$$\hat{\alpha}_j = \frac{1}{n}\sum_{i=1}^{n} X_i^j \tag{7.18}$$

The method of moments estimator $\hat{\theta}_n$ is defined to be the value of $\theta$ so that:

$$\begin{aligned} \alpha_1(\hat{\theta}_n) &= \hat{\alpha}_1 \\ \alpha_2(\hat{\theta}_n) &= \hat{\alpha}_2 \\ \alpha_3(\hat{\theta}_n) &= \hat{\alpha}_3 \\ \alpha_4(\hat{\theta}_n) &= \hat{\alpha}_4 \\ &\vdots \end{aligned} \tag{7.19}$$

The method of moments estimator satisfies:

1. The estimate $\hat{\theta}_n$ exists with probability tending to 1.
2. The estimate is consistent: $\hat{\theta}_n \xrightarrow{P} \theta$ (it converges in probability)
3. The estimate is asymptotically normal (cf. (**?**), pp.122)

### 7.6.2 Maximum Likelihood Estimation

The maximum likelihood estimator is the value $\hat{\theta}$ that maximizes the joint probability density of the data, called the likelihood function.

For i.i.d. random variables with pdf $f(x;\theta)$, the likelihood function is:

$$\mathscr{L}_n(\theta) = \prod_{i=1}^{n} f(X_i;\theta) \tag{7.20}$$

And the log likelihood functin is $l_n(\theta) = \log \mathscr{L}_n(\theta)$. Since $\log$ is a monotonic function, maximizing the log-likelihood yields the same estimator as maximizing the likelihood directly. Log-likelihood is often easier to deal with, and alleviates numerical issues associated with the often sharply spiked likelihood function.

MLE estimators have a flurry of desirable properties under certain smoothness conditions on the density function.

- Consistency: convergence in probability upon the true value
- Equivariance: if $\hat{\theta}_n$ is the MLE of $\theta$ then $g(t\hat{heta}_n)$ is the MLE of $g(\theta)$
- Asymptotically Normal
- Asymptotically Optimal / Efficient: smallest variance, at least for large samples.
- Approximately the Bayes estimator.

### 7.6.3 Parametric Confidence Intervals

Confidence intervals for infered parameters in the parametric setting can be derived, for example, using the delta method (assuming that the estimators are asymptotically normal) or using parametric bootstrap. In the nonparametric case, bootstrap sampled from the empirical CDF. In the parametric case, bootstrap sample from the density $f(X;\hat{\theta})$ where $\hat{\theta}$ is the estimator.

## 7.7 Score Function, Fisher Information

Given some pdf $f(X;\theta)$, the score function is given by:

$$s(X;\theta)\frac{\partial \log f(X;\theta)}{\partial \theta} \tag{7.21}$$

The Fisher information is the variance of the score function at each datapoint:

$$I_n(\theta) = \mathbb{V}_\theta \left( \sum_{i=1}^{n} s(X_i;\theta) \right) \tag{7.22}$$

## 7.8 Directional Statistics

### 7.8.1 Mean Direction

The mean direction of a collection of $i$ vectors $\{\mathbf{x}\}_i$ is (Damask 2019):

$$\langle \mathbf{x} \rangle = \frac{\mathbf{x_s}}{||\mathbf{x_s}||_2} \tag{7.23}$$

Where

$$\mathbf{x_s} = \sum_i \mathbf{x_i} \tag{7.24}$$

The mean vector is not defined in case $||\mathbf{x_s}||_2 = 0$. The alternative way to calculate the mean direction might make use of angles but apparently that creates ambiguity with respect to the choice of a "zero" angle (cf. footnote in Damask (2019)).

### 7.8.2 Dispersion

Dispersion is a measure of the variance on the direction of a set of vectors $\{\mathbf{x}\}_i$. For a system of vectors, for example an eigenbasis, there are common and differential modes of dispersion. One way of measuring directional dispersion is to look at the *mean resultant length*:

$$\mu_r = \frac{||\mathbf{x_s}||}{N} \quad 0 \leq \mu_r \leq 1 \tag{7.25}$$

*Circular variance* may be defined as $\sigma_c = 1 - \mu_r$, but apparently there is an issue with generalization to higher dimensions. (TO DO)

An alternative is a model based approach, for example based on the von Mises - Fisher Distribution.

In the case of a basis, Damask (2019) states that, in order to understand the drivers for variation, dispersion parameters should be calculated for each descending subspace of the basis, first including all of the eigenvectors, then excluding the first eigenvector, etc.

## 7.9 Features

Features are sources of information that hopefully allow conclusions towards some kind of quantity of interest. Some people like to call them independent variables.

### 7.9.1 Dense Features, Sparse Features

Sparsity and density refer to the fraction of a matrix that is zeros. In the context of features, sparse features typically refer to feature vectors that have many zeros in them. I.e., $[1, 2, 5, 2, 6, 3, 0]$ would be a dense feature vector and $[2, 0, 0, 0, 0, 3, 0, 0, 1]$ would be a sparse feature vector. Typically, sparsity is touted as an advantage because it allows for lossless compressed representations of high-dimensional data, which has computational advantages. Another advantage, though, is that sparse representations can be more interpretable by containing information in an inherently more condensed fashion. They are also thought to prevent overfitting: many regularization techniques aim to minimize the number of parameters or features used in a predictive model, which amounts to biasing an algorithm towards learning sparse coefficients. An example is ridge regression.

## 7.10 Multicollinearity

So far this section is based on Software (n/a).

Multicollinearity, or collinearity, occurs when features are linearly correlated with each other. The effects are horrible. They include inaccurate estimates of the regression coefficients, higher standard errors of the regression coefficients, lower partial t-tests for the regression coefficients, falsely insignificant p-values and decreased predictive power of the model. And other things.

#### 7.10.0.1 Detection

**Scatter Plots** Scatterplots provide a visual test for collinearity by hopefully exposing relationships between independent variables. This is subjective and unreliable, but people love plots.

**Variance Inflation Factors (VIF)** A VIF over 10 it said to indicate collinear variables.

**Eigenvalues of the Correlation Matrix** Linear relationships between two or more variables cause the corresponding rows of the correlation matrix to be identical or very similar. Correspondingly, the matrix will be singular or near-singular, which will manifest itself through zero or near-zero eigenvalues. The conditioning number, given by the largest eigenvalue divided by the smallest eigenvalue, are a quick way to test for this. A large conditioning number indicates collinearity.

**Regression Coefficients** Collinearity increases the standard error of the regression coefficients because it allows for the variation of the dependent variable to be explained in terms of a greater variety of different weights assigned to the collinear variables. Counterintuitive results for the regression coefficients may be the result of collinearity.

### 7.10.1 Sources

**Data Collection**

**Physical Constraints**

**Over-defined Model**

**Model Choice or Specification**

**Outliers**

#### 7.10.1.1 Remedies

**Dimensionality Reduction** SVD, PCA, NNMF and other dimensionality reduction techniques allow for the feature space to be shrunk in a way that aims to optimally preserve information. Any technique worth its salt will either collapse or filter collinear variables into a reduced-rank representation of the information in the dataset.

**Regularization** Certain forms of regularization are similar in spirit to dimensionality reduction, except that, rather than addressing the dataset, they reduce the parameters used by a model to fit to the data. The canonical example is ridge regression. Ridge regression penalizes the use of a larger number of parameters and, if two variables are collinear, will tend to push the weight of one of them towards zero. It is important to standardize the variables before fitting the model, so that the regression weights for different variables are on the same scale. Ridge regression remains controversial (Software n/a).

## 7.11 Fat Tails

Roughly speaking, the consequences of "fat tails" is that rare events tend to play a disproportionally large role in determining the statistical properties of a sample. This is in contrast to "thin tails", where, as the sample size increases, pretty quickly no single observation modifies the statistical properties of the sample.

For "thin tails", extreme observations tend to result from the combination of several very unlikely events. For "fat tails", an extreme observation tend result from a single very unlikely event.

**Thin Tails** Two people are randomly selected, and their combined height is 4 meters. Most likely this resulted from the selection of two people who are two meters tall, rather than one person that is 10 cm and another that is 3,90 m tall.

**Fat Tails** Two people are randomly selected, and their combined net worth is $ 40M. The probability of having selected two people with a net work of $ 20M is less likely than having selected one person with a net worth of $ 200k and another person with a net worth of $ 39,8 M.

Fat tails do not imply that rare events are more frequent, but that they have greater impact when they do happen. In fact, “fattening" the tails of a Gaussian distribution results in a larger number of observations within one standard deviation.

### 7.11.1 Consequences of Fat Tails

**The Law of Large Numbers works too Slowly**

**The Sample Mean will rarely correspond to the Distribution Mean** For example, for an 80/20 power law, 92% of the observations will fall below the distribution mean. The sample mean will tend to underestimate the distribution mean because the distribution mean is heavily biased by rare observations that will tend to be underrepresented in the sample.

**Metrics such as Sample Mean and Sample Variance will be Unusable**

**In finance, metrics like Sharp etc. are unusable**

**Gauss-Markov Theorem fails** Therefore, linear least squares regressions do not work.

**Maximum Likelihood Methods can still work** For example in case of a Pareto distribution, it is possible to fit the tail exponent using maximum likelihood methods, and estimating the mean from there. Direct observation of the mean would be misleading. The tail exponent intelligently extrapolates the fat tails of the distribution.

**Absence of Evidence $\neq$ Evidence of Absence**

**PCA is going to cause spurious factors and loads**

**Method of Moments does not work** Approximating a distribution by matching its moments does not work when higher moments are undefined or cannot be reliably estimated.

**There is no “typical" large deviation**

### 7.11.2 Maximum to Sum

The "maximum to sum" or MS plot allows to see the behavior of the relationship between the observed maximum to the sum for a particular moment as the number of observations increases.

### 7.11.3 Maximum Domain of Attraction

The “maximum domain of attraciton" is, so to speak, the “right endpoint of the distribution":

$$x^* = \sup\{x : F(x) < 1\} \tag{7.26}$$

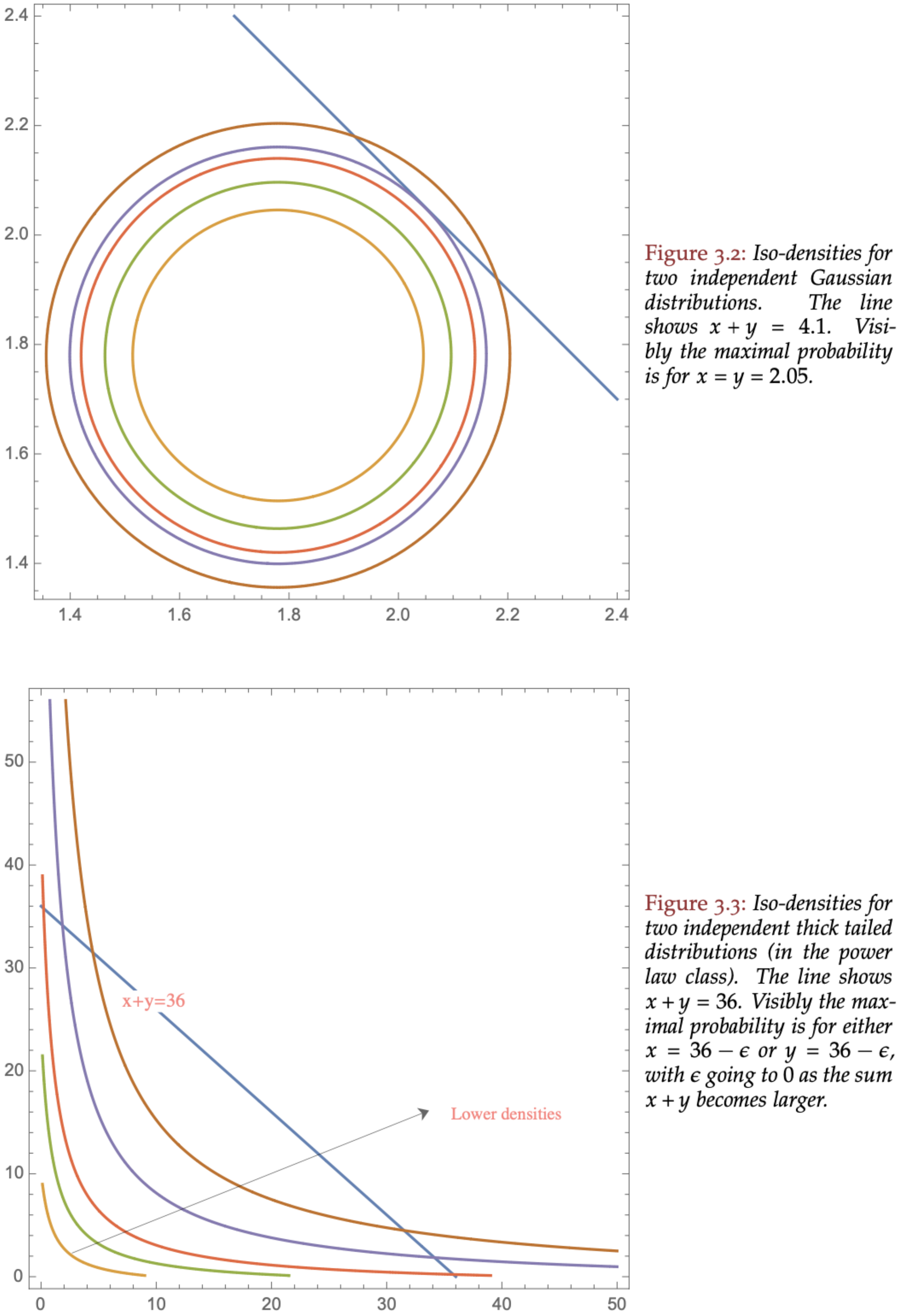


Figure 3.2: *Iso-densities for two independent Gaussian distributions. The line shows $x + y = 4.1$. Visibly the maximal probability is for $x = y = 2.05$.*

Figure 3.3: *Iso-densities for two independent thick tailed distributions (in the power law class). The line shows $x + y = 36$. Visibly the maximal probability is for either $x = 36 - \epsilon$ or $y = 36 - \epsilon$, with $\epsilon$ going to 0 as the sum $x + y$ becomes larger.*

Figure 7.1: Probability densities of two independent thin tailed and thick tailed random variables (brazenly copied from Taleb, 2020). Compare to plot of Lp norms. For the thin tailed random variables, the observation of a particular sum is most likely to result from a balanced contribution from both random variables. For the fat tailed random variables, the observation of a particular sum is most likely to result from the contribution of one of the variables.

### 7.11.4 Hidden Tail, Problems in Estimating Moments

The Glivenko-Cantelli theorem guarantees uniform convergence of the empiral cdf to the true CDF, however, the empirical distribution is necessarily bounded by the values of the minimum and maximum observations. This results in an unobserved contribution to moments $p > 0$ that does not necessarily have to be negligible. To illustrate, take $K_n$ to be the maximum observed value.

$$\mathbb{E}(X^p) = \underbrace{\int_L^{K_n} x^p \phi(x) \mathrm{d}x}_{observed} + \underbrace{\int_{K_n}^{\infty} x^p \phi(x) \mathrm{d}x}_{unobserved} \tag{7.27}$$

## 7.12 $R^2$ Value

## 7.13 Regression Diagnostics

## 7.14 t-Statistics

## 7.15 AIC and BIC

## 7.16 Factor Regression

## 7.17 Factor Models

## 7.18 How to Combine Estimates with Different Uncertainties

A series of classic papers exist on this subject.

## 7.19 How to Tackle Very High Dimensional Feature Spaces

In practice, high-dimensional feature spaces amplify the risk of overfitting. The answer, framed in the most general way, is to introduce limitations on the family of functions that are fit by the estimator of choice. The generic methods are regularization (i.e. penalizing variance), variable selection (i.e. filtering out irrelevant features) and dimensionality reduction through feature transformation (i.e. finding a lower-rank representation of the data). It's not possible to draw neat distinctions between this methods in terms of their effects. For example, regularization often amounts to variable selection through soft (or hard) thresholding. Drop out regularization in neural networks biases the model towards particular types of feature transformations.

# 8 Linear Regression

**Contents of this chapter**

There is no escaping from linear regression but hardly anyone seems to agree on how to do it properly. Even worse, people seem to expect you to know things like the normal equations, which you'll never, ever need outside of a job interview.

## 8.1 Least Squares Regression ($L^2$)

### 8.1.1 The Normal Equations, Analytical Least Squares Estimator

Section 3.3 discussed the case of the linear system of equations:

$$\mathbf{A}\mathbf{x} = \mathbf{b} \tag{8.1}$$

When $\mathbf{A}^{m \times n}$ with $m > n$, so that the system is overdetermined. This is, of course, the starting point for least squares regression, only that the convention is to use different letters:

$$\mathbf{X}\beta = \mathbf{y} \tag{8.2}$$

And that, seeing that the system is overdetermined, one looks for an approximate solution $\hat{\beta}$, so that

$$\mathbf{X}\beta + \epsilon = \mathbf{y} \tag{8.3}$$

A natural approach for picking an approximate solution $\hat{\beta}$ is to look for the projection of $\mathbf{y}$ in the column space of $\mathbf{X}$. That is, since the column rank $\leq n$ of $\mathbf{X}$ is insufficient to express $m$-dimensional $\mathbf{y}$ exactly in terms of only $m$ coefficients $\beta$, we look for the $n$-dimensional shadow $\hat{\beta}$ of some hypothetical higher dimensional exact solution.

The projection has the property that it maximizes the dot product $(\mathbf{X}\hat{\beta}) \cdot \mathbf{y}$, and hence minimizes the length of the difference vector $\epsilon$. In turn, the length of the difference vector $\epsilon$ is $\sqrt{\epsilon \cdot \epsilon}$, which is monotonic to $\epsilon \cdot \epsilon = \sum_i^m \epsilon_i^2$. That means that finding the projection of $\mathbf{y}$ in the column space of $\mathbf{X}$ minimizes the $L_2$ norm of $\epsilon$, also known as *least squares error*.

There are two ways to go about finding $\hat{\beta}$.

### 8.1.2 The Quick Way to $\hat{\beta}$

By construction, the vector $\epsilon$ is orthogonal to the column space of $\mathbf{X}$. Which means:

$$\begin{aligned} \mathbf{X}^T\epsilon &= 0 \\ \mathbf{X}^T\left(\mathbf{X}\hat{\beta}-\mathbf{y}\right) &= 0 \\ \mathbf{X}^T\mathbf{X}\hat{\beta} &= \mathbf{X}^T\mathbf{y} \\ \hat{\beta} &= \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{y} \end{aligned} \tag{8.4}$$

Making use of the fact that $\mathbf{X}^T\mathbf{X}$ is square and therefore hopefully invertible.

### 8.1.3 The Long Way to $\hat{\beta}$

Loss functions play a central role in computational statistics (for example when regularization is introduced), and therefore it is of interest to approach finding $\hat{\beta}$ by instead minimizing the least square error. This requires:

$$\frac{d}{d\hat{\beta}}L_2(\epsilon) = 0 \tag{8.5}$$

where

$$\begin{aligned} L_2(\epsilon) &= \left(\mathbf{X}\hat{\beta}-\mathbf{y}\right)^T\left(\mathbf{X}\hat{\beta}-\mathbf{y}\right) \\ &= \hat{\beta}^T\mathbf{X}^T\mathbf{X}\hat{\beta} - \hat{\beta}^T\mathbf{X}^T\mathbf{y} - \mathbf{y}^T\mathbf{X}\hat{\beta} + \mathbf{y}^T\mathbf{y} \end{aligned} \tag{8.6}$$

Taking derivatives with respect to a vector is covered in section 3.14.
It follows:

$$\begin{aligned} &\tfrac{d}{d\hat{\beta}}\left(x^T\mathbf{X}^T\underbrace{\mathbf{X}\hat{\beta}}_{u(\hat{\beta})}\right) = \tfrac{d}{du}\left(u^Tu\right)\tfrac{d}{d\hat{\beta}}u = 2u^T\tfrac{d}{d\hat{\beta}}u = 2\hat{\beta}^T\mathbf{X}^T\mathbf{X} \\ &\tfrac{d}{d\hat{\beta}}\hat{\beta}^T\mathbf{X}^T\mathbf{y} = \mathbf{y}^T\mathbf{X} \\ &\tfrac{d}{d\hat{\beta}}\mathbf{y}^T\mathbf{X}\hat{\beta} = \mathbf{y}^T\mathbf{X} \\ &\tfrac{d}{d\hat{\beta}}\mathbf{y}^T\mathbf{y} = 0 \end{aligned} \tag{8.7}$$

So that

$$\begin{aligned} \tfrac{d}{dx}L_2(\epsilon) = 0 &= 2\hat{\beta}^T\mathbf{X}^T\mathbf{X} - 2\mathbf{y}^T\mathbf{X} \\ \hat{\beta}^T\mathbf{X}^T\mathbf{X} &= \mathbf{y}^T\mathbf{X} \\ \mathbf{X}^T\mathbf{X}\hat{\beta} &= \mathbf{X}^T\mathbf{y} \\ \hat{\beta} &= \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{y} \end{aligned} \tag{8.8}$$

### 8.1.4 Projection Matrix

If $\hat{\mathbf{y}} = \mathbf{X}\hat{\beta}$ is the projection of $\mathbf{y}$ in the column space of $\mathbf{X}$, then, based on the result for $\hat{\beta}$, the projection matrix is $\mathbf{P} = \mathbf{X}\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T$. In a fully determined system, $\mathbf{P} = \mathbf{I}$. Projection matrices have eigenvalues that are either $1$ or $0$, corresponding to dimensions that are kept or discarded during the projection operation.

### 8.1.5 Bayesian Perspective on Least Squares Regression

### 8.1.6 Q-plots

### 8.1.7 Variance Inflation Factor

## 8.2 Total Least Squares

While least squares regression only allows for errors in the dependent variable, total least squares regression allows for measurement errors on both variables.

## 8.3 Ridge Regression (Tikhonov Regularization, $\lambda||\beta||^2$)

Ridge Regression ads the $L^1$ norm of the weight vector to

### 8.3.1 Analytical Ridge Estimator

### 8.3.2 Bayesian Perspective on Ridge Regression

## 8.4 Least Absolute Shrinkage and Selection Operator Regression (LASSO)

## 8.5 Least Absolute Deviation Regression (LAD, $L^1$)

## 8.6 Generalized Linear Models

# 9 Bayesian Data Analysis

**Contents of this chapter**

## 9.1 Markov Chain Monte Carlo Methods

Markov Chain Monte Carlo (MCMC) methods allow for the approximate solution of Bayesian inference problems by drawing samples from the posterior distribution.

Plain vanilla MCMC works as follows:

1. Make an initial guess $\theta_0$ for the value of the latent variables. This is the starting point for the Markov Chain, which could be picked randomly or, for example, the maximum a posteriori estimate.
2. Calculate the probability of observing the data based on these parameters $(p(\{\text{data}\}|\theta_0))$.
3. Suggest values $\theta$ where the Markov Chain might jump next. (The way guesses are generated is where optimized sampling might comes in.)
4. Calculate $p(\{\text{data}\}|\theta)$.
5. Calculate a probability of jumping to the new values, $p_{\text{jump}} = \min\left(\frac{p(\{\text{data}\}|\theta)}{p(\{\text{data}\}|\theta_0)}, 1\right)$.
6. With probability $p_{\text{jump}}$, let the Markov Chain jump $\theta_0 \to \theta$.
7. Repeat steps 3-6

It can be shown that upon convergence, the probability of the Markov Chain reaching particular values of the latent variables is given by the posterior distribution. In other words, MCMC is a trick to use likelihood (and priors) to sample the posterior distribution. Certain pathological posteriors can make it difficult or impossible for the Markov Chain to sample the full posterior, and there is also no completely certain way to say that convergence has been achieved. I personally sample using multiple independent chains, and assume convergence when the posteriors sampled by all chains looks identical. I also test how robust the results are to changes in the priors.

# 10 Unsupervised Learning

**Contents of this chapter**

## 10.1 Blind Source Separation

Blind Source Separation or Blind Signal Separation (BSS) is the separation of mixed signals. An example is the cocktail party problem, where a microphone is recording the conversations of many people speaking at the same time, making the recording unintelligible. Successful BSS would be able to extract the voice of only a single person.

## 10.2 Clustering

### 10.2.1 K-Means

### 10.2.2 Affinity Propagation

## 10.3 Independent Component Analysis (ICA)

Independent Component Analysis (ICA) is a decomposition of data into components that have vanishing mutual information.

## 10.4 Compressed Sensing

## 10.5 $\mathbf{X} = \mathbf{U\Sigma V^T}$ Eigenanalysis, Singular Value Decomposition (SVD)

Singular Value Decomposition (SVD) is a matrix decomposition of the form $\mathbf{A} = \mathbf{U}\Sigma\mathbf{V}^T$, where $\mathbf{U}$ and $\mathbf{V}$ are both unitary (cf. section 3.7). The decomposition always exists for a general complex matrix $\mathbf{A} \in \mathbb{C}^{m\times n}$.

If $\mathbf{X} \in \mathbb{R}^{N \times p}$ is a data matrix with $N$ samples in the rows and $p$ features in the columns, then SVD allows for the decomposition of the data into $p$ linearly independent components, ranked by their explained variance (i.e. their strength in the dataset). The basis of the decomposition turns out to be the same as for PCA (cf. section 10.6) except that it is arrived at slightly differently, because PCA involves diagonalizing the covariance matrix., nor are the singular values the same as the explained variances of the dataset. The decomposition of the dataset can be understood as follows. In the SVD of the data matrix:

$$\mathbf{X} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T \tag{10.1}$$

The individual rows of $\mathbf{X}$, i.e. the individual data points, are expressed as:

$$\mathbf{x}_i^T = \sum_{j=1}^{p} u_{i,j}\sigma_j \mathbf{v}_j^T \tag{10.2}$$

Where $\mathbf{x}_i$ is the $i$th data point. Hence the right singular vectors $\mathbf{v}_i$ are the normalized basis vectors, and the columns of $\mathbf{U}$ give the coefficients the basis expansion, together with the singular values $\sigma_i$, which give a measure of the global strength of the corresponding basis vector.

Truncating the expansion after the $r$th term results in a rank $r$ approximation of $\mathbf{X}$ known as the *truncated singular value decomposition* (TSVD) estimator. According to the Eckart-Young Mirsky-Theorem, the estimator is the best rank-$r$ approximation under the Frobenius norm of the error (cf. 3.13.3), which is the average mean square error across $\mathbf{X}$.

$$\hat{\mathbf{X}} = \sum_{i=1}^{r} \sigma_i(\mathbf{u}_i \otimes \mathbf{v}_i) \tag{10.3}$$

Since it minimizes the average mean square error, $\hat{\mathbf{X}}$ is the rank $r$ maximum likelihood approximation to $\mathbf{X}$ under the assumption of normally distributed noise. Optimal truncation is discussed in section 10.8.

SVD is scalable to very large datasets and finds many applications in the wild, including page rank, facial recognition, recommendation algorithms, and others. Randomized SVD (cf. section 10.5.4) is an approximate method that gives an even faster speedup.

### 10.5.1 Example: Eigenfaces and Facial Recognition

One famous result are the so-called eigenfaces. The data matrix $\mathbf{X} \in \mathbb{R}^{N \times p}$ consists of $N$ pictures of faces that each have $p$ pixels. The right singular eigenvectors yield $p$ eigenfaces in terms of which any of the $N$ pictures can be expressed.

Below are the first eigenfaces extracted from the "Labeled Faces in the Wild" dataset, which includes 13233 portraits with 62x47=2914 pixels each. Running SVD on the data matrix $\mathbf{X} \in \mathbb{R}^{13233 \times 2914}$ yields the eigenfaces shown in Figure 10.1.

Facial recognition may be performed by projecting a new face into the space of eigenfaces and matching to the coefficients of a known subject. This could be done using the euclidean distance, or it could be done using a classifier. In case of classificiation, the dimensionality reduction that is enabled by approximating images with a smaller set of eigenvectors may be critical to making the problem tractable by overcoming the curse of dimensionality. When the data matrix is not centered (that is, the mean is not subtracted from it before performing the SVD), then the first right singular vector is vector close to the average row in the matrix.

### 10.5.2 Tracking an Eigensystem over Time

As discussed in section 3.7, the sign of the basis vectors can be flipped without affecting the validity of the singular value decomposition. That means that if the SVD is performed on some system may equally well return, say, a left-handed or a right-handed coordinate system. This becomes a problem when the results of repeated SVDs are supposed to be compared, for example to study the evolution of a system over time. Figure 10.2 shows the eigenbasis of a bivariate Gaussian with principal axes slowly rotating over

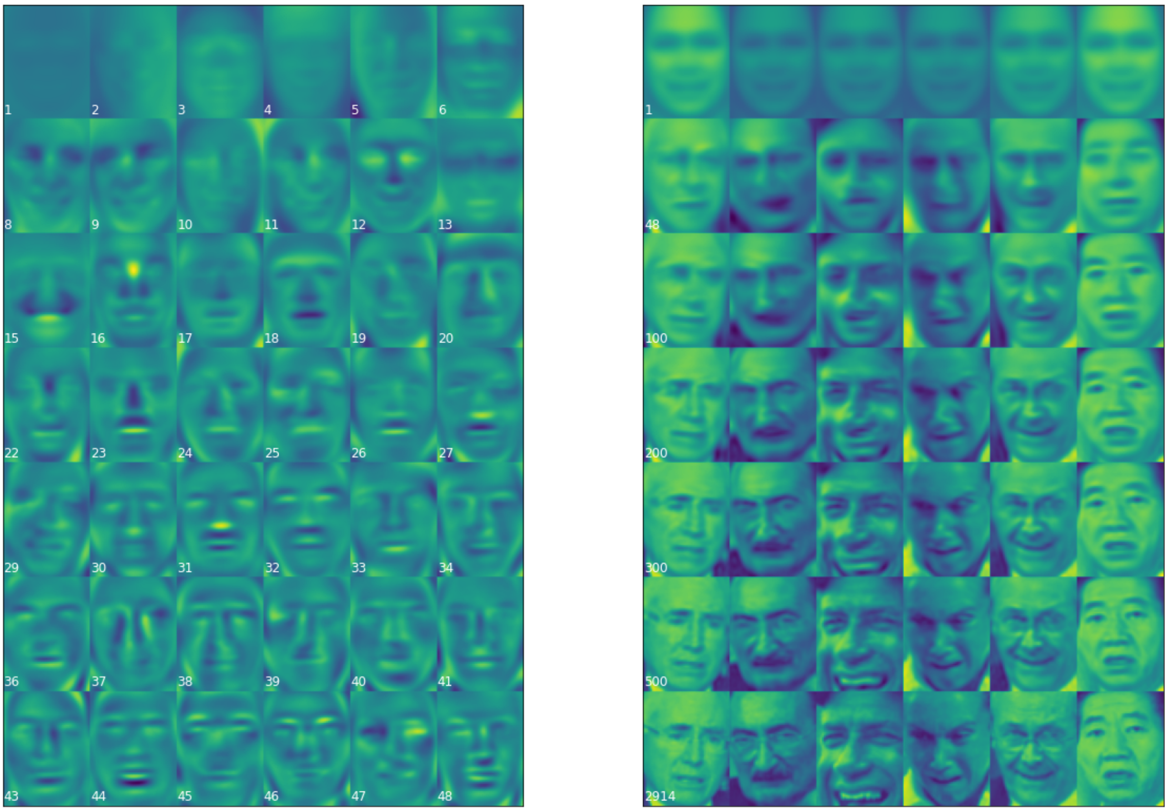


Figure 10.1: Left: The first 48 eigenfaces. The colorscale is consistent across the images. As can be expected, the eigenfaces seem to have an ordering from more general features, that are highly prevalent in the dataset, towards more specific features. Right: Five sample portraits from the dataset approximated using different numbers of eigenvectors. 2914 corresponds to the original image, which had 2914 pixels (degrees of freedom).

a 180-degree angle. The top shows how the eigenbases that is found flips back and forth, so that there is a double-trajectory corresponding to results with positive and negative sign on an eigenvector. The trajectory is not described by an injective function, which complicates analysis.

#### 10.5.2.1 Heuristic Method

Say an SVD is performed on a slowly-evolving system at time $t = 0$ and at time $t = 1$. To compensate the sign flips, a heuristic method is to first match the basis vectors at $t = 0$ to the basis vectors at $t = 1$ using a distance metric that is immune to the sign of the vectors, for example the absolute value of the dot product. (This assumes that the vectors are similar enough at time $t = 0$ and $t = 1$ that the matching an be done unambiguously. For the higher-order singular vectors this might be a problem, because they have lower numerical certainty.) Once the vectors are matched, the sign of the inner product between the vectors at $t = 0$ and $t = 1$ may be compared, and the sign flipped accordingly.

This method works, but the problem is that the sign of the vectors is somewhat arbitrarily pinned relative to the result at $t = 0$. That is, if, for example, the coordinate system found at $t = 0$ was left-handed, then the time series of eigenbases will be adjusted to be left-handed. If the point of comparison had instead been the SVD performed at some other time $t = t'$, then one might have wound up with a right-handed coordinate system instead. It is desirable to find a consistent orientation.

#### 10.5.2.2 Consistent Method

Damask (2019) recently developed a method that can be used to find a consistently oriented basis. Consistent orientation in this case means, roughly, that the eigenbasis is always flipped to obey the convention of being right-handed. The method relies on reconstructing the rotations and reflections necessary to transform the eigenbasis in question to the natural basis $\mathbf{I}$ as reference. While rotations preserve the orientation of a basis, but reflections do not. When the reflections of an eigenbasis with respect to the natural basis are known, then they can be undone by flipping them back in place.

#### 10.5.2.3 Rank Order Changes

The singular vectors or eigenvectors of a system are labeled only in terms of their associated singular values or eigenvalues. In order to track a singular vector throughout rank order changes, it is necessary to figure out a way to attach a separate label, for example by looking at the "content" of a particular singular vector. How well that works has to be figured out in context, I am not currently aware of a generally valid solution.

### 10.5.3 Bayesian SVD

To do! Cf. https://ieeexplore.ieee.org/document/7336426

It's awful to be tied to the assumption of normality and it's awful to not know the uncertainty of my singular vectors!

### 10.5.4 Randomized SVD

Randomized SVD is a method to very efficient and effective way to perform approximate singular value decompositions using random matrix theory. The premise is a very large data matrix $\mathbf{X} \in \mathbb{R}^{n \times m}$ that has low *intrinsic rank* $r$, so that $\mathbf{X} \approx \mathbf{U_r \Sigma_r V_r}^T$ will be a good approximation.

#### 10.5.4.1 Step 1: Sample Column Space of $\mathbf{X}$ with $\mathbf{P}$

Given a random projection matrix $\mathbf{P} \in \mathbb{R}^{m \times r}$, the matrix:

$$\mathbf{Z} = \mathbf{XP} \tag{10.4}$$

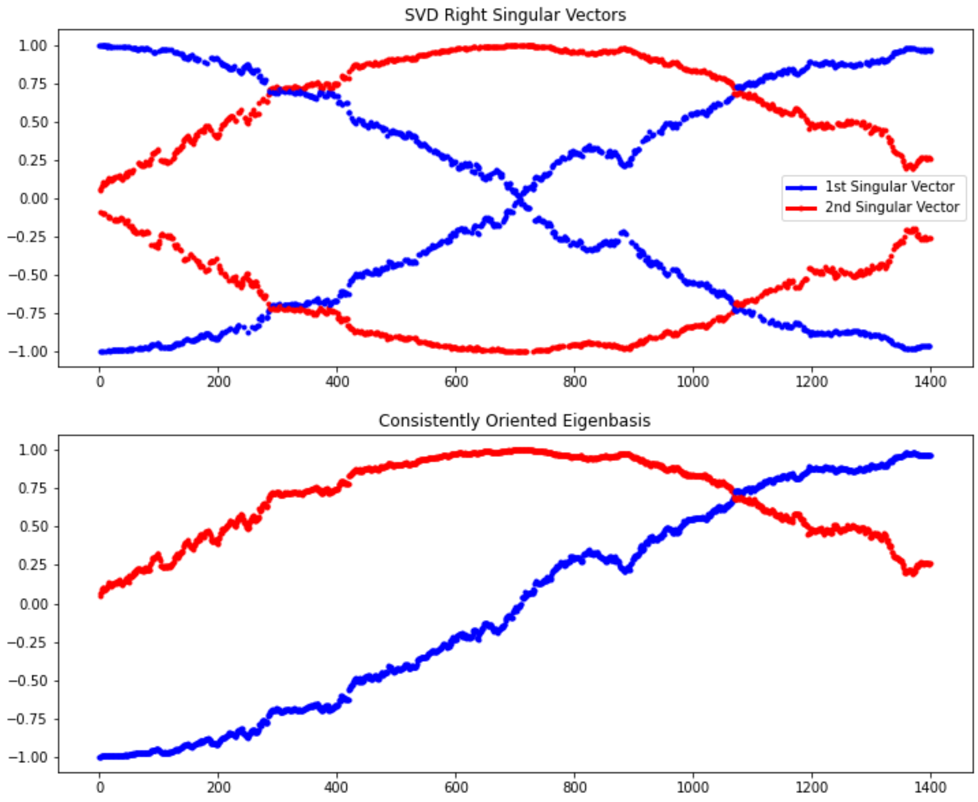


Figure 10.2: Right singular vectors extracted using SVD, showing sign flips (top) and with a consistently oriented basis (bottom). The underlying data is a bivariate gaussian with principal axes undergoing a 180 degree rotation. With the consistently oriented basis, the singular vectors trace out a one-to-one trajectory that can be analyzed.

will, because of the properties of random matrices, likely have the same rank $r$ dominant column space as $\mathbf{X}$. SVD is usually calculated via the QR-decomposition of a matrix. Performing the QR decomposition of $\mathbf{Z}$ is much less laborious, because $\mathbf{Z} \in \mathbb{R}^{m \times r}$ is much smaller than $\mathbf{X}$. We obtain:

$$\mathbf{X} = \mathbf{QR} \tag{10.5}$$

Where $\mathbf{Q} \in \mathbb{R}^{m \times r}$ and $\mathbf{R} \in \mathbb{R}^{r \times r}$.

#### 10.5.4.2 Step 2: Compute SVD on Projected $\mathbf{Y} = \mathbf{Q}^{\mathbf{T}}\mathbf{X}$

The next step consists of projecting $\mathbf{Z}$ into $\mathbf{Q}$:

$$\mathbf{Y} = \mathbf{Q}^T\mathbf{X} \tag{10.6}$$

Resulting in the matrix $\mathbf{Y} \in \mathbb{R}^{r \times n}$. Performing the SVD on $\mathbf{Y}$ (which is, again, much smaller than $\mathbf{X}$), gives:

$$\mathbf{Y} = \mathbf{U}_Y \mathbf{\Sigma}_r \mathbf{V}_r^T \tag{10.7}$$

Where $\mathbf{\Sigma}_r$ and $\mathbf{V}_r$ turn out to likely be the same as those that would be obtained from the SVD of $\mathbf{X}$ itself. In the last step, $\mathbf{U}_r$ is recovered using the mathrix $\mathbf{Q}$, via $\mathbf{U}_r = \mathbf{Q}\mathbf{U}_Y$.

Guaranteed error bounds for the low rank approximation obtained with this technique. There are two approaches to improve the accuracy of randomized SVD. The first is to oversample, by letting the projection matrix $\mathbf{P}$ have a rank larger than $r$. The second is to "sharpen" the singular value spectrum by using the matrix $(\mathbf{X}\mathbf{X}^{\mathbf{T}})^{\mathbf{q}}\mathbf{X}$ instead of $\mathbf{X}$. Calculating the power results in a matrix with a much faster drop off in the singular values, but is much more computationally expensive. This technique is useful when the singular values of $\mathbf{X}$ decay only slowly. The column space of the power iteration $\ldots\mathbf{X}\mathbf{X}^{\mathbf{T}}\mathbf{X}\mathbf{X}^{\mathbf{T}}\mathbf{X}\mathbf{X}^{\mathbf{T}}\mathbf{X}$ has the same column space, so that the trick with the projection operation works here too, only that the dominant subspace will be emphasized.

```python
def rSVD(X,r,q,p):
    """
    Randomized SVD Code
    """
    # Step 1: Sample column space of X with P matrix
    ny = X.shape[1]
    P = np.random.randn(ny,r+p)
    Z = X @ P
    for k in range(q):
        Z = X @ (X.T @ Z)

    Q, R = np.linalg.qr(Z,mode='reduced')

    # Step 2: Compute SVD on projected Y = Q.T @ X
    Y = Q.T @ X
    UY, S, VT = np.linalg.svd(Y,full_matrices=False)
    U = Q @ UY

    return U, S, VT
```

Empirically, randomized SVD outperforms SVD significantly the lower the rank of the TSVD.

## 10.6 Principal Component Analysis (PCA)

Principal Component Analysis (PCA) is a decomposition of data into linearly uncorrelated components, ordered by their explained variance. It turns out that these axes are the right-singular eigenvectors $\mathbf{V}$

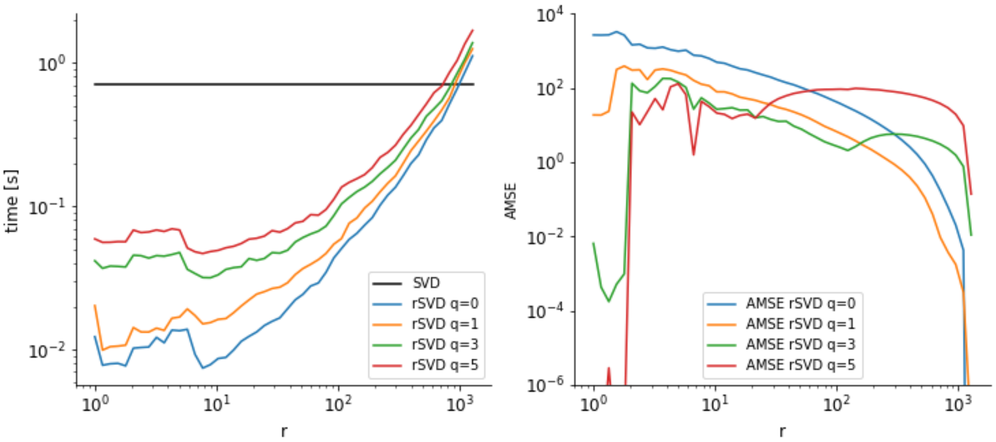


Figure 10.3: Left: Time taken to perform conventional SVD (black line) and randomized SVD (colored lines) on the photo of the two young aristocrats which had a data matrix with rank $r = 1277$. The different colors correspond to different numbers of power iterations performed ($q$). Right: Average mean square error (AMSE) of truncated SVD using matrices derived with randomized SVD compared with conventional SVD. Especially for small $r$, rSVD is significantly faster even after a few power iterations. For very small $r$, power iterations easily reduce the error introduced by the randomized method by several orders of magnitude, but at some point the method seems to increase the error. Perhaps this is due to numerical stability. I'm not sure. To somewhat account for random fluctuations, the curves shown are averaged over 20 trials.

that are found from SVD. SVD and PCA have correspondence (cf. section 10.6.1). My guess is that the advantage of PCA is that its results are interpretable in terms of commonly used summary statistics of the distribution (namely the variances or Pearson correlations of the data).

More elaborately, for a data matrix $\mathbf{X} \in \mathbb{R}^{N \times p}$, PCA extracts the ordered, rank-$p$, orthonormal basis in which the $p \times p$ covariance matrix of $\mathbf{X}$ is diagonal. The basis vectors are called the principal axes or principal directions of the data, and their ordering is by the magnitude of the variance that they explain.

The property that the covariance matrix is diagonal in the principal component basis means that projecting the data onto any of the basis vectors extracts a linearly uncorrelated component of the data that has variance corresponding to the corresponding eigenvalue of the covariance matrix.

Just like with SVD, the ordering of the principal component basis in terms of their explained variance allows for lower-rank approximations of the data matrix to be constructed (section 3.7). Just like SVD, the lower rank approximations $\mathbf{X}$ are the best possible approximations with respect to the Frobenius Norm (cf. section 3.13.3).

The $p \times p$ covariance matrix $\mathbf{C}$ of $\mathbf{X}$ is:

$$\mathbf{C} = \frac{(\mathbf{X} - \langle \mathbf{X} \rangle)^T (\mathbf{X} - \langle \mathbf{X} \rangle)}{n-1} \tag{10.8}$$

Where $(\mathbf{X} - \langle \mathbf{X} \rangle)^T$ is often referred to as the *centered* data matrix. The covariance matrix is a symmetric, positive-definite matrix that can be diagonalized with orthonormal eigenvectors $\mathbf{V}$ and positive (or vanishing) eigenvalues $\lambda_i$:

$$\mathbf{C} = \mathbf{V} \boldsymbol{\Lambda} \mathbf{V}^T \tag{10.9}$$

Where the eigenvectors in $V$ are ordered so that the eigenvectors along the diagonal of $\boldsymbol{\Lambda}$ have decreasing magnitude. The eigenvectors are the principal axes or principal directions of the data.

That PCA corresponds to the diagonalization of the covariance matrix implies a description of the underlying data in terms of a multivariate Gaussian distribution. In other words, it assumes normally distributed noise, and the components describe best-fit linear subspaces in the data under the $L^2$ norm. Whenever the Gaussian assumption is unjustified, it makes sense to look for best-fit linear subspaces under different norms. For example, $L^1$-PCA is more robust to outliers (cf. section 10.6.3).

### 10.6.1 Relationship between PCA and SVD

This is based on a great Stack Exchange answer (amoeba 2015).

Let the singular value decomposition of the centered data matrix be:

$$(\mathbf{X} - \langle \mathbf{X} \rangle) = \mathbf{U\Sigma V}^T \tag{10.10}$$

Then:

$$\mathbf{C} = \frac{\mathbf{V\Sigma U}^T \mathbf{U\Sigma V}^T}{n-1} = \mathbf{V}\frac{\Sigma^2}{n-1}\mathbf{V}^T \tag{10.11}$$

That means that:

- The principal axes are the right-singular vectors $\mathbf{V}$ that are obtained during SVD.
- The singular values and the eigenvalues of the covariance matrix are related via $\lambda_i = \frac{\sigma_i^2}{n-1}$.

### 10.6.2 Tracking Principal Components over Time

Given the direct parallel between PCA and SVD, the identical issue with sign flips emerges with the principal component basis, and the remedy is the same. See section 10.5.2.

### 10.6.3 $L^1$-Norm Principal Component Analysis ($L^1$ PCA)

$L1$-Norm Principal Component Analysis is an alternative to conventional PCA that aims to provide better robustness to outliers. While L2-PCA is uniquely defined, a cursory search reveals a zoo of approaches to what $L^1$ PCA is supposed to be (Brooks & Jot n.d.). $L^1$-PCA may also be a natural choice in situations in which the $L^1$ norm is the proper way of measuring distances in a space, for example in cellular automata models or perhaps even in case of categorical data.

**?** establishes that the PCA under $L^1$ norm is found by successive fitting of hyperplanes under $L^1$ error in progressively smaller subspaces. Much like the $L^2$ distance of a point to a hyperplane is given by the radius of a circle around that point, the $L^1$ distance of a point to a hyperplane is given by its intersection of a rhombus. This has the (kind of quirky) consequence that the distance of all points to a hyperplane is always measured directly along one of the coordinate axes (and it is the same coordinate axes for each of the points). This, in turn, implies, that in some subspace $\mathbb{R}^m$, the $L^1$ best-fit hyperplane can be found by performing an $L^1$ regression with each of the $m$ dimensions serving as the explanatory variable, and then selecting the regression result that had the smallest residual.

It is worth noting that this approach is within the spirit of minimizing the taxicab-like distances of a point to some sub-space, which assigns great significance to the coordinate grid. This goes against the spirit of PCA, in that it typically aims to extract some kind of natural coordinate system from the data, irrespective of the coordinate system in which the raw data was expressed. **?** method is within keeping of the definition of the $L^1$ norm, but its different from simply avoiding assigning quadratic importance to far-away points to achieve greater robustness.

## 10.7 $\mathbf{X} = \mathbf{WH}$ Non-Negative Matrix Factorization

Non-Negative Matrix Factorization (NNMF) is an approximate decomposition technique for positive data. It can be used, for example, for dimensionality reduction, blind source detection or topic extraction. The

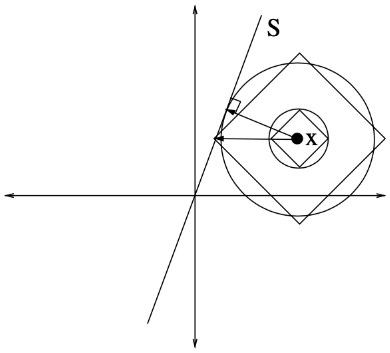


**Fig. 1.** Level sets for the $L_1$ and $L_2$ norms.

Figure 10.4: Distance of a point to a hyperplane in $L^1$ and $L^2$. The $L^1$ distance in this case is just the distance along the $x$ axis. If the hyperplane was cutting a shallower angle, then the distance would be measured along the $y$ axis. The best fit hyperplane to many points in $\mathbb{R}^2$ would be found by regressing the data points once with $y = \beta x + \epsilon$ and once with $x = \beta y + \epsilon$ and selecting the result with the smaller $||\epsilon||_1$. (The figure is from J. P. Brooks (2014).

reputation is that it can often generate decompositions that are "sparse and meaningful".

Following the convention in Gillis (2014), it works by expressing a positive data matrix $\mathbf{X} \in \mathbb{R}_+^{p\times n}$ in terms of two smaller positive matrices $\mathbf{W} \in \mathbb{R}_+^{p\times r}$ and $\mathbf{H} \in \mathbb{R}_+^{r\times n}$. Here the $n$ columns of the data matrix are data points and the rows are the $p$ features. An excellent introduction to NNMF is found in Colyer's (2019) blog post and in Gillis (2014). Dimensionality reduction derives from the ability to adjust the rank of the decomposition via the dimension $r$. The decomposition is simply:

$$\mathbf{X} \approx \mathbf{WH} \tag{10.12}$$

Where the error $||\mathbf{X} - \mathbf{WH}||_\alpha$ is minimized with respect to some norm (normally, the Frobenius norm with $\alpha = 2$). The decomposition allows the interpretation of $\mathbf{W} \in \mathbb{R}^{p\times r}$ as a matrix with a number of $r$ $p$-dimensional strictly positive basis vectors and of $\mathbf{H} \in \mathbb{R}^{r\times n}$ as a matrix of coefficients that express the $n$ data points in terms of linear sums of the $r$ basis vectors with strictly positive coefficients.

One might ask: why NNMF when I can create linear decompositions of my data with SVD and PCA? After all, these techniques are well established and are known to give the optimal approximation under the Frobenius nurm. In my perception, the key difference is that NNMF solves the constrained problem of forcing both coefficients and basis vectors to be strictly positive. For many data types, negative values (word frequencies, pixel values) are unnatural, and so the results of NNMF are immediately more interpretable. NNMF is promising in particular in situations where a signal arises from the addition of a number of positive signals. A canonical example is hyperspectral imaging, where, assuming incoherent light, the measured spectrum is the linear sum of the spectra of the individual light sources. If all goes well, the basis vectors from NNMF of a hyperspectral image will be the source spectra. In the context of natural language processing, NNMF on a collection of documents can be interpreted as *topics*. Figure 10.5 shows the NNMF-based decomposition and approximation of the "labeled faces in the wild" dataset, analogous to the SVD of the dataset shown in Figure 10.1. The two are quite different. While for SVD, the faces look as if they are gradually coming into focus, for NNMF the faces are at first not recognizable. The portraits also become brighter in the plot as more basis components are added, because, in contrast to the SVD case, the composition is purely additive.

The drawback is that NNMF is computationally more difficult (possibly NP hard), nor is the matrix approximation optimal, nor is the decomposition unique.

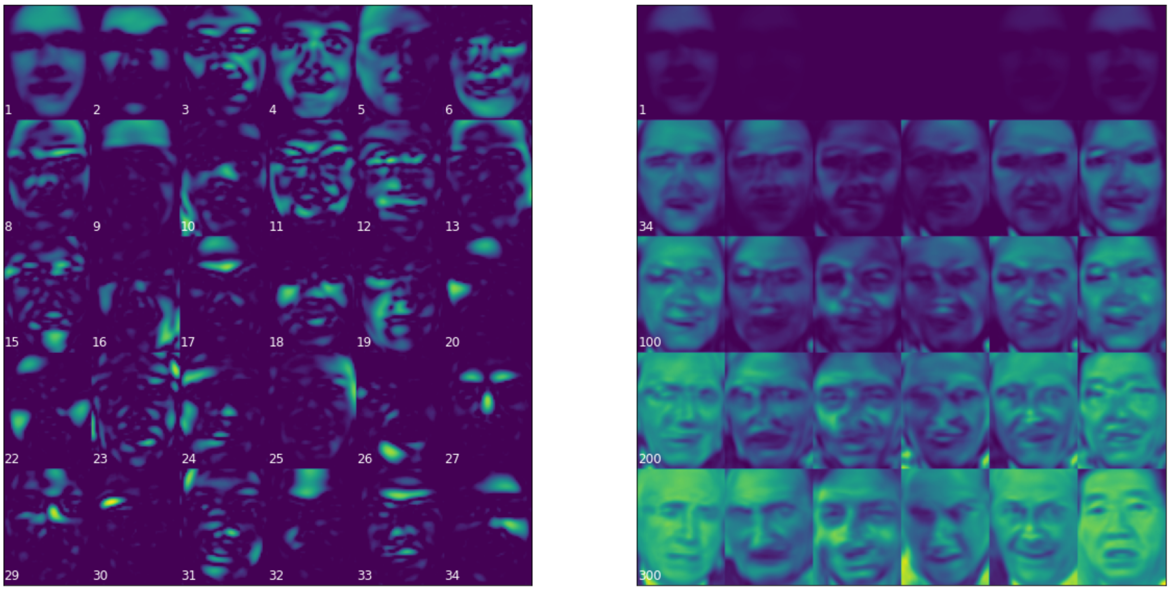


Figure 10.5: Left: The first 34 basis faces stored in the $\mathbf{W}$ matrix extracted using NNMF under Frobenius norm from the "Labeled Faces in the Wild" Dataset. The basis images are normalized and shown on a log scale because they have very different contrast and brightness. The normalized images are then all shown on the same color scale. Right: Five sample portraits from the dataset approximated using different numbers of basis faces. The NNMF was done for a value of $r = 300$. The original portraits had $2914$ pixels. In contrast to SVD, the basis vectors obtained through NNMF do not necessarily have an internal ordering in terms of how much of the variance in the dataset they explain.

## 10.8 Optimal Truncation

In any decomposition-based approximation the question emerges of how many terms to keep. The optimal value depends both on the true rank of the underlying signal but also the signal to noise ratio: truncating the decomposition of a rank $r$ data matrix at $r$ does not necessarily give the result that minimizes error. Almost always, the suggestion is to use a heuristic method, but it turns out that one can do better. The discussion below uses the context of singular value decomposition, but the problem of optimal truncation is more general and is an example tuning the hyperparameter that determines the capacity of an estimator to mold itself to a dataset.

### 10.8.1 Scree Plots, Heuristic Methods

The canonical heuristic approaches focus on *scree plots*, which simply show the singular values in decreasing order, or cumulative plots, which show the fraction of the total sum of singular values that the first $n$ singular values add up to. In case of PCA, the cumulative plot is directly interpretable as the fraction of the explained variance by the first $n$ principal components, where the total variance is given by $\sigma_2 = \sum_i \sigma_i^2$.

A scree plot is sometimes referred to as an *elbow plot*, and a common approach is to truncate at the elbow, which is known as *elbow truncation*. Scree refers to the rock fragments that gather at the foot of a steep hill due to erosion, but I wouldn't really know, because I haven't gone outside in a while.

Another common way to go about this would be to, for example, keep only the singular components that account to 90% of the decomposition.

Either method is not limited to SVD or PCA, but also show up whenever, for example, someone wants to decide on the number of clusters in cluster analysis, the number of regressors in mutlivariate regression (only sometimes), or independent component analysis (ICA).

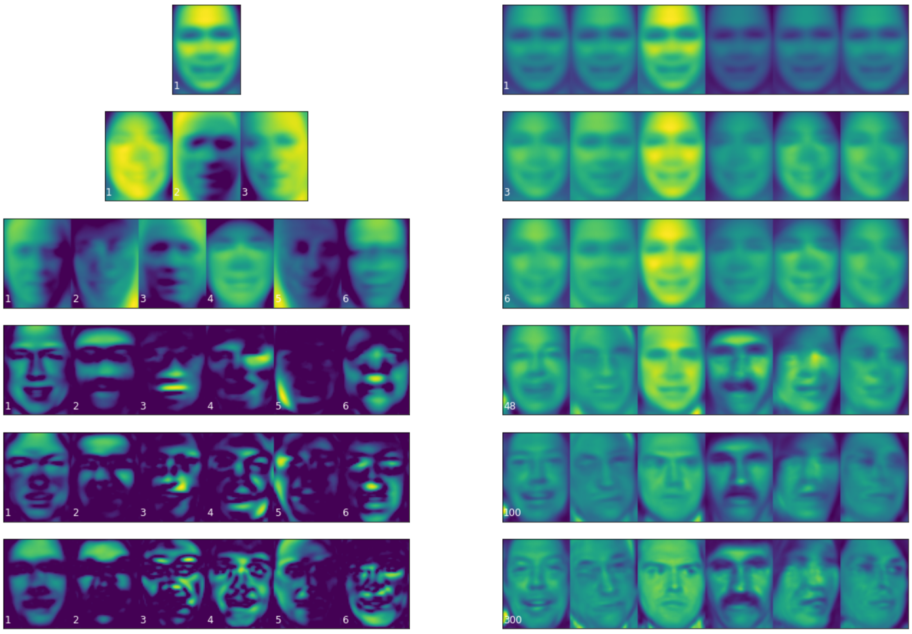


Figure 10.6: Left: The first 6 basis faces stored in the $\mathbf{W}$ matrix extracted using NNMF under Frobenius norm from the "Labeled Faces in the Wild" Dataset. Right: Sample images expressed in the reduced basis.

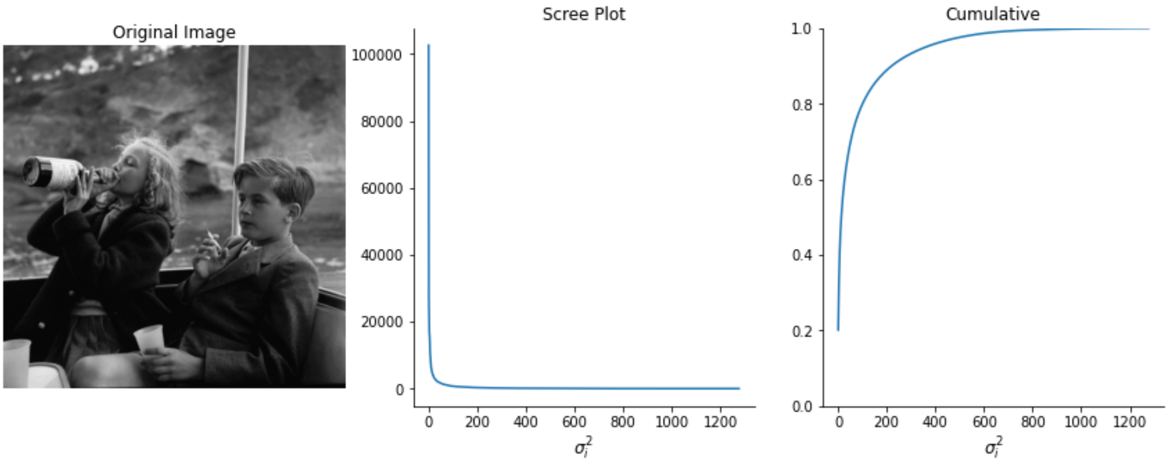


Figure 10.7: Left: Original image of Princess Yvonne und Prince Alexander zu Sayn-Wittgenstein, photographed by their mother, Princess Marianne, in 1955 (1280 by 1277 pixels). Middle: Scree plot, which shows that the first few components are dominant in the image. Right: Cumulative plot that shows that the first singular component accounts for more than 20% of the decomposition.

### 10.8.2 Gavish & Donoho's Optimal Threshold for SVD

I came across this through Gavish & Donoho (2013) and Steve Brunton's youtube channel (Brunton n/a). The idea is the following. Assume that some data matrix $\mathbf{X} \in \mathbb{R}^{N \times p}$ consists of a signal $\mathbf{X}_r$ with a rank $r$ substructure, and normally distributed noise with mean zero $\mathbf{X}_\epsilon \sim \mathcal{N}(0, \sigma_\epsilon^2)$. The method assumes that the rank of the signal $r$ is small compared to the rank of $\mathbf{X}$.

$$\mathbf{X} = \mathbf{X}_r + \mathbf{X}_\epsilon \tag{10.13}$$

The distribution of singular values of a matrix $\mathbf{X}_\epsilon \sim \mathcal{N}(0, \sigma_\epsilon^2)$ is known up to the variance. Gavish & Donoho's (2013) method is to truncate the decomposition of $\mathbf{X}$ at the singular value that fall below the largest singular value of $\mathbf{X}_\epsilon$. It turns out that, with respect to average mean square error, this truncation is asymptotically always better than any other truncation and even always better than the true thresholding at rank $r$ when the rank of the underlying signal happens to be known. When the variance of the noise $\sigma_\epsilon^2$ is not known, then Gavish & Donoho (2013) instead use the median singular value of $\mathbf{X}$.

Let $y_{\text{med}} = \text{med}\{\sigma_i : 1 \leq 0 \leq n\}$ be the median singular value. Then the optimal hard threshold for a matrix $\mathbf{X} \in \mathbb{R}^{m \times n}$ is given by:

$$\tau = \omega(\beta) y_{\text{med}} \tag{10.14}$$

Where $\beta = \frac{m}{n}$ is the aspect ratio of the matrix $\mathbf{X}$ so that $0 \leq \beta \leq 1$, and $\omega(\beta)$ is a constant that needs to be calculated numerically:

$$\omega(\beta) = \frac{\lambda_*(\beta)}{\sqrt{\mu_\beta}} \tag{10.15}$$

Where $\lambda_*(\beta)$:

$$\lambda_*(\beta) = \sqrt{2(\beta+1) + \frac{8\beta}{(\beta+1) + \sqrt{\beta^2 + 14\beta + 1}}} \tag{10.16}$$

And $\mu_\beta$ is unfortunately the median of the Marčenko-Pastur distribution, which is the unique solution to:

$$\int_{\beta_-^x} \frac{\sqrt{(\beta_+ - t)(t - \beta_-)}}{2\pi t} \mathrm{d}t = \frac{1}{2} \tag{10.17}$$

With $\beta_\pm = (1 \pm \beta)^2$. Cautiously, Gavish & Donoho (2013) provide the approximation:

$$\omega(\beta) \approx 0.56\beta^3 - 0.95\beta^2 + 1.82\beta + 1.43 \tag{10.18}$$

Which has error bounded by $\leq 0.02$ on the interval $0.001 \leq \beta \leq 1$.

## 10.9 Matrix Completion, Imputation

Imputation or Matrix completion problems deal with missing observations. The most famous application is probably the Netflix prize, in which a very large, very sparse data matrix encodes the preferences of Netflix users. Missing entries correspond to the case where it is not known what a certain user might think of a movie.

### 10.9.1 Nuclear Norm Regularization

Given a matrix $\mathbf{X} \in \mathbb{R}^{m \times n}$, in which observed entries are indexed by the set $\Omega = \{(i,j) : X_{i,j} \text{ is observed.}\}$(Hastie et al. 2015). Define the projection $P_\Omega(\mathbf{X}) \in \mathbb{R}^{m \times n}$ to be the matrix so that $P_{i,j} = X_{i,j} \; \forall \; (i,j) \in \Omega$ and $P_{i,j} = 0$ if $(i,j) \notin \Omega$, which is to say: take the matrix $\mathbf{X}$ and set all the unobserved entries to $0$.

Nuclear norm regularization corresponds to expressing completing $\mathbf{X}$ in terms of the convex optimization problem:

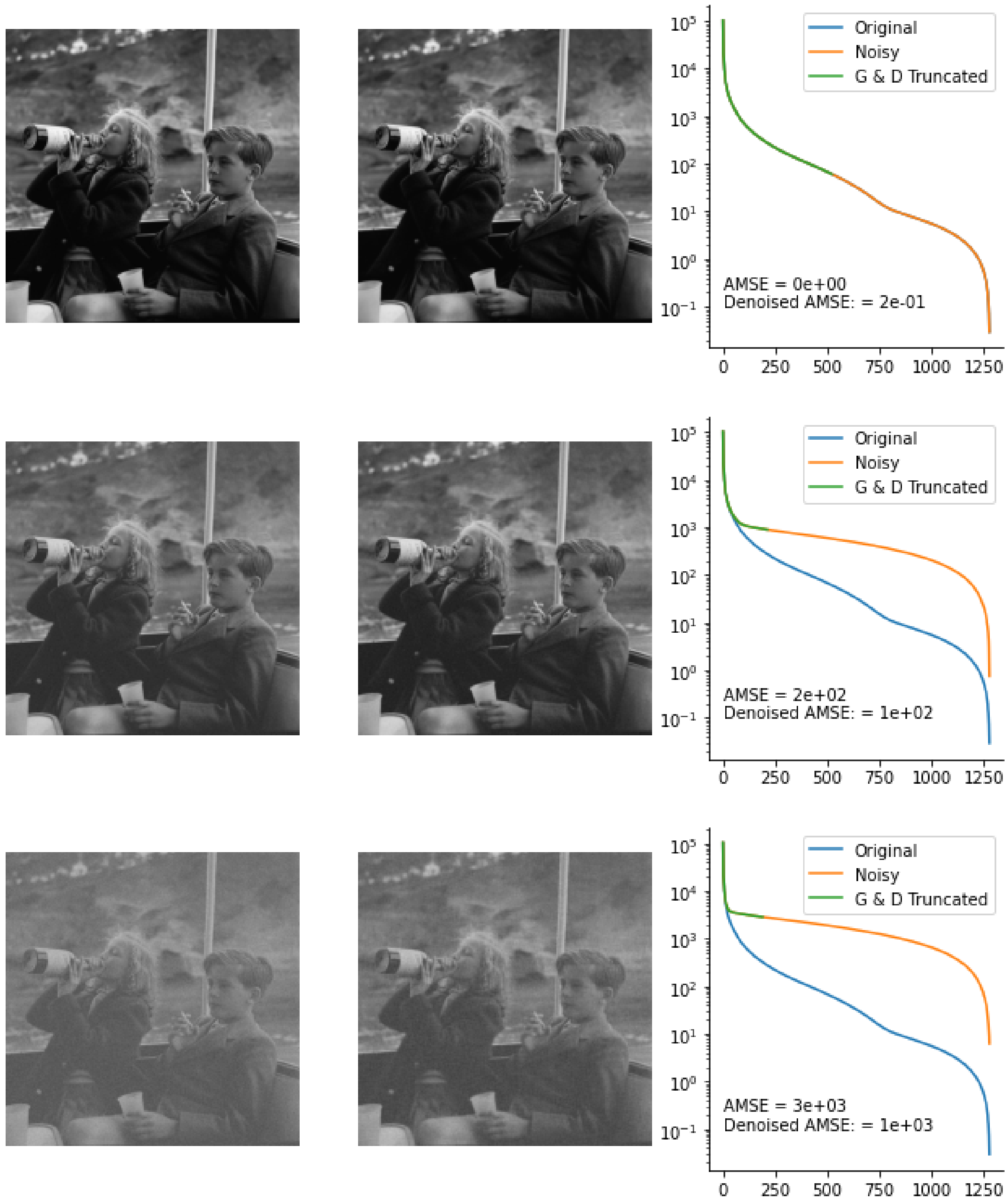


Figure 10.8: Left: Princess and Prince zu Sayn-Wittgenstein with different amounts of Gaussian noise added. Center: The same image estimated using TSVD with Gavish & Donoho's approximate rank threshold. Right: Singular value spectrum of original, noise and truncated images. TSVD reduces the average mean square error (AMSE) with respect to the original image by over 50%. Visually, the difference is not too perceptible.

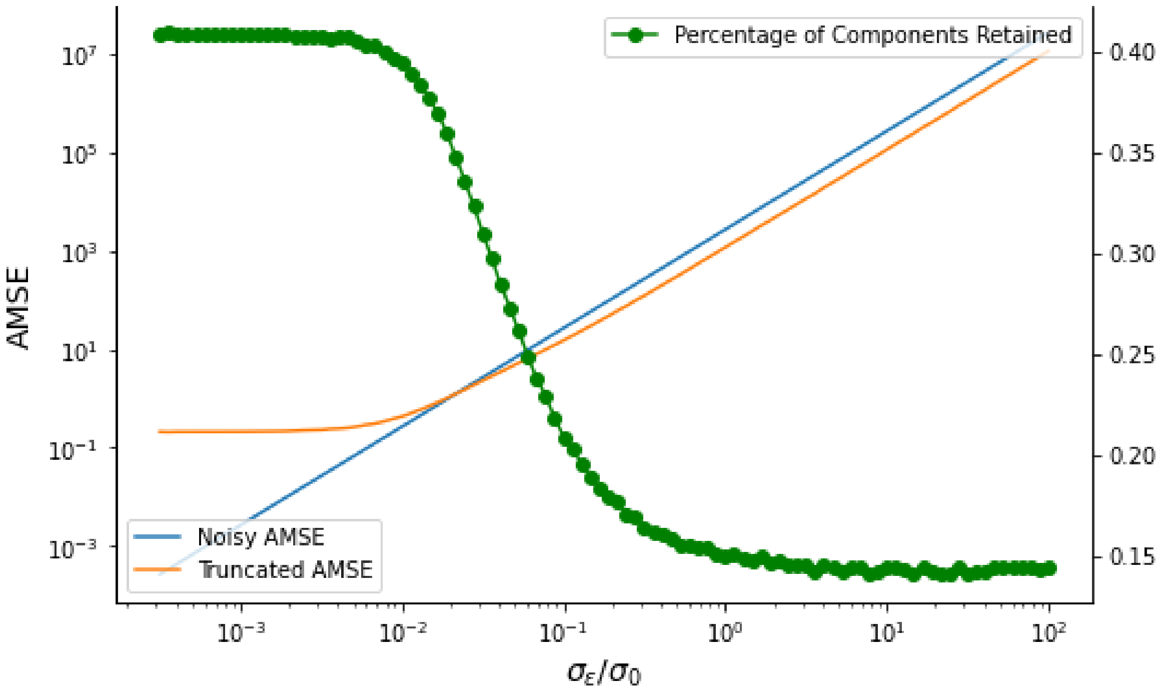


Figure 10.9: Average means square error relative to the original image for the photo with added noise (blue) and after TSVD denoising (organge). The percentage of components retained via Gavish & Donoho's approximate rank threshold shown in green. $\sigma_\epsilon$ is the standard deviation of the added (Gaussian) noise and $\sigma_0$ is the standard deviation of the pixel values of the original image. When $\sigma_\epsilon$ is small, G & D automatically truncates the TSVD at about 40% of the singular values, which is why, at first, the "denoised" image has greater error than the noisy issues.

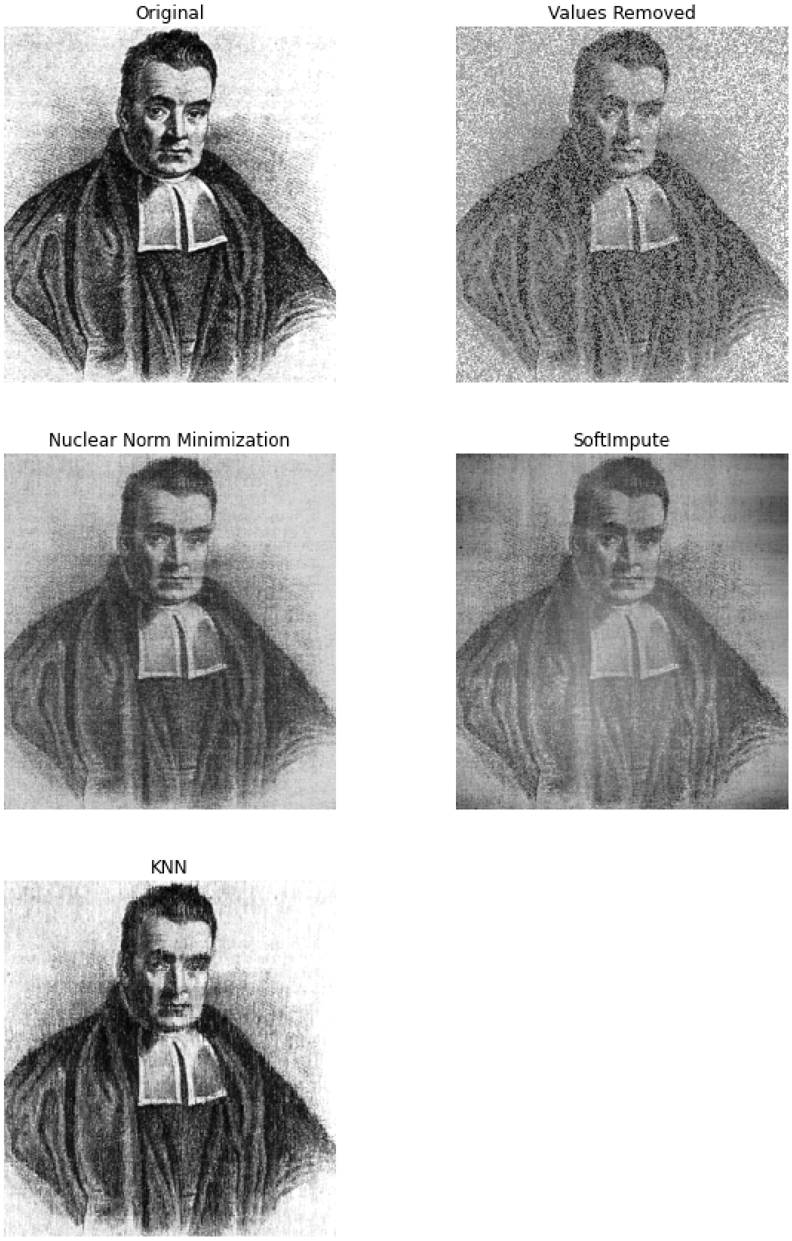


Figure 10.10: Results for different imputation methods reconstructing the image of the man rumored to be Bayes himself after 50% of the pixels are removed.

$$\arg\min_{\mathbf{M}} H(\mathbf{M}) = \frac{1}{2}||P_\Omega(\mathbf{X} - \mathbf{M})||_F^2 + \lambda||\mathbf{M}||_* \tag{10.19}$$

Where $||\cdot||_*$ is the nuclear norm of $\mathbf{M}$ (cf. section 3.13.7). The loss function is a tradeoff between accurately reproducing $\mathbf{X}$ and doing so with as low a rank as possible. Except, using the rank of $\mathbf{M}$ would make the optimization non-convex, so instead the nuclear norm is used, which is the sum of the singular values. Solving this problem is computationally still quite expensive, but is a lot of work on solving this problem for large datasets, for example `softImpute`, texttsoftImpute+, which use soft-singular value thresholding that amounts to leveraging lower-rank representations of the data (Mazumder et al. 2010).

Autoencoders are a dimensionality reduction technique that use neural networks to compress information in a lower-dimensional latent space. This is done by training a network that has a bottleneck, i.e. a layer with lower dimensionality, with the data used both as input and as output. The network up to the bottleneck ("code") is the encoder. The part of the network that takes the code as input is the decoder.

In practice, classical autoencoders apparently don't generate particularly useful or nicely structured latent spaces, nor compress particularly well.

I have frequently seen them used for anomaly detection: you train an autoencoder on your data. Datapoints that have particularly high error after passing through the autoencoder are considered outliers, i.e. "anomalous". Call it part of the "reconstruction error" type of approach to anomaly detection.

In principle, autoencoders can be used as generational models by feeding values into the decoder. However, there is no guarantee that any particular region of the latent space encodes a meaningful representation. Variational Autoencoders address this to some extent.

Autoencoders are effectively functions that map a high-dimensional input vector to a fixed, low-dimensional code. The hope is that this possibly highly non-linear function correctly projects out some lower-rank structure in the input data. The range of this function is unknown, and picking a point in the co-domain, even if it is close to a point known to be in the range, will not necessarily decode to a meaningful datapoint. The connection between code and data is deterministic.

Variational Autoencoders address this by pretending that the data are manifestations of underlying stochastic processes. The "code" learned by the variational autoencoders are the latent parameters of the process that gave rise to the input. The "decoding" consists of plugging particular values for the latent parameters and sampling from the resulting distribution. The result of the sampling is compared to the input. The variational autoencoder learns latent parameters that improve the likelihood of the resulting distribution producing a sample similar to the input. Meaningful outputs that are similar to a particular input can be generated by sampling the distribution that uses as latent parameters the code of the input, or codes that are close to the code of the input. The randomness causes the latent space to be continuously meaningful.

- An encoder takes input sample to two parameters in the latent space of representations: mean, variance
- You sample a point from the latent normal distribution that's assumed to have generated the input via $z = \mu + \sigma\epsilon$ where $\epsilon \sim \mathcal{N}$.
- The decoder maps this point $z$ back to the input image.

Concept Vectors are directions in representation spaces that encode meaningful variations in the data. For example, in images of faces, there might be a "smile" vector, so that if $z$ is the latent-space coordinate of a face then $z + s$ migth be the latent-space coordinate of the same face smiling.

The loss function for training VAEs has two parts: *reconstruction loss* forces the output to be clsoe to the input, and *regularization loss* forces structure upon the latent space (which also helps overfitting). In practice, *regularization loss* might be the KL-divergence of the code distribution with respect to a standard Normal.

# 11 Dynamical Systems

**Contents of this chapter**

This is dangerously familiar ground for physicists. A dynamical system is some system:

$$\partial_t \mathbf{x} = \mathbf{f}(\mathbf{x}, t, \mathbf{u}; \beta) \tag{11.1}$$

Where $\mathbf{x}$ are the state space coordinates of the system at some time $t$, and $\mathbf{f}$ is the *dynamics* of the system. $\mathbf{u}$ is some control input and $\beta$ are parameters. A system where $\mathbf{f}$ depends on time is called *non-autonomous* and a system that has an $\mathbf{f}$ that does not depend on time is called *autonomous*.

Conventionally, the dynamics $\mathbf{f}$ are derived from first principles. Increasingly, it is possible to infer them from data. Challenges arise from:

- Nonlinear $\mathbf{f}$, that is, the system cannot be described in the form $\partial_t \mathbf{x} = \mathbf{A}\mathbf{x}$
- Unknown $\mathbf{f}$
- High dimensional state vector $\mathbf{x}$
- Chaos, Transients
- Noise, Stochastic forcing functions
- Multiscale dynamics
- Uncertainty (in parameters etc.)

## 11.1 Mode Decompositions

Modal decompositions express the evolving state of a system in terms of a superposition of basis vectors that are the eigenfunctions of a time-translation operator. In general, the rank of the basis space is the same as the state space dimension, but in practice only a few of those modes have "energy". That means that the description of a system in terms of a modal decomposition often requires orders of magnitude fewer parameters than the description of the state space. Extracting the dominant modes also amounts to extracting the dominant dynamics of the system, with the assumption possibly being that the discarded low-energy modes are noise. In so far, modal decomposition runs closely parallel to decomposition-based dimensionality reduction techniques such as SVD, PCA and ICA, though explicitly introducing the notion of a trajectory through time.

The modes are properties of the time translation operator and the state space. Therefore, the description in terms of modes is compatible with an arbitrary, unpredictable forcing function acting on the system, so long as it does not affect the time translation operator or the state space. For example, the description of the oscillation of a guitar string in terms of modes remains valid regardless of the song that is being played. However, changing the length of the string increases the size of the state space, and increasing the tension in the string accelerates its reversion from being struck away from its equilibrium, and therefore it's time translation operator. In those cases the modes of the system are changed.

## 11.2 Koopman and Frobenius-Perron Operators

The Koopman Operator, or composition operator, is the classical analogue of the time translation operator $e^{-i\hat{\mathbf{H}}t/\hbar}$ in quantum mechanics. It describes dynamics in terms of the *Heisenberg Picture*, in which operators, rather than states, have time-dependence. It describes autonomous systems (in quantum mechanics, that would be a system where the Hamiltonian $\hat{\mathbf{H}}$ nor the size of the state space depend on time).

Consider a dynamical system evolving on a manifold $\mathscr{M}$, so that, in discrete time:

$$\mathbf{x}_{k+1} = \mathbf{f}(\mathbf{x}_k) \tag{11.2}$$

Where $\mathbf{x} \in \mathscr{M}$ are the state space coordinates of the system at a given time. Given a scalar valued function on the state space $g : \mathscr{M} \to \mathbb{R}$, the Koopman Operator $\hat{\mathbf{K}}$ acts as:

$$\hat{\mathbf{K}}g(\mathbf{x}) = g(\mathbf{f}(\mathbf{x})) \tag{11.3}$$

Which acts as time translation by one step.

The Koopman Operator is infinite dimensional and linear, even though the dynamical system might be nonlinear with a finite state space. The Frobenius-Perron Operator (sometimes also Ruelle Operator, Ruelle-Frobenius-Perron Operator, or Transfer Operator) is the right adjoint of the Koopman operator, so that an approximation of one provides an approximation of the other. Rather than a function in state space, it translates the state space density in time. Let $\hat{\mathbf{P}}$ be the Frobenius-Perron Operator. Let the state space density be $\rho(\mathbf{x}$, and define some operator $\hat{\mathbf{A}}$, so that $\hat{\mathbf{A}}\rho(\mathbf{x}) = g(\mathbf{x})$. Then, in terms of either the Koopman or the Frobenius-Perron approach, the average value of the quantity $\langle g(\mathbf{x}(t))\rangle$ at time $t$ in the Koopman picture or in the Frobenius-Perron picture is (Cvitanovic et al. 2016, Salova et al. 2019):

$$\begin{aligned} \langle g(\mathbf{x}(t))\rangle = \quad & \int_{\mathscr{M}} \mathrm{d}x \hat{\mathbf{K}}(t)\hat{\mathbf{A}}\rho(\mathbf{x}) \\ & \int_{\mathscr{M}} \mathrm{d}x \hat{\mathbf{A}}\hat{\mathbf{P}}(t)\rho(\mathbf{x}) \end{aligned} \tag{11.4}$$

Which implies that $\hat{\mathbf{K}}(t)\hat{\mathbf{A}}\left[\hat{\mathbf{P}}(t)\right]^{-1} = \left[\hat{\mathbf{K}}(t)\right]^{-1}\hat{\mathbf{A}}\hat{\mathbf{P}}(t) = \hat{\mathbf{A}}$.

The function $g(\mathbf{x}(t))$ is an *observable* of the system, which might be a pixel value, the value of a stock portfolio, the energy of a particle, or something like that. The state space coordinates $\mathbf{x}(t)$ may or may not correspond to quantities that are actually observable. Their role is to uniquely parametrize the possible states of the system. The Koopman Operator acts on observables, which is probably why it seems to come up more often in data-driven work.

## 11.3 Proper Orthogonal Decomposition (POD)

Proper Orthogonal Decomposition (POD) seeks a lower-rank representation of a dynamical system that, as far as I can see, is entirely analogous to SVD or PCA (Megretski 2004). Given a system $\mathbf{x}(t) \in \mathbb{R}^n$, one looks for a low-rank projection $\Pi_r$ that minimizes the expected value of the error:

$$\arg\min \mathbb{E}_t ||\mathbf{x}(t) - \mathbf{\Pi}\mathbf{x}(t)||_2 \tag{11.5}$$

This is the same loss function as for PCA, and so the projection matrix $\mathbf{\Pi}_r = \sum_{i=1}^{r} \mathbf{v}_i \mathbf{v}_i^T$ where $\mathbf{v}_i$ is the $i$th principal component vector.

The POD does not have anything to say about the dynamics of a system, indeed, the time-ordering of the snapshots has no effect on the result. As such, performing POD on a system across different time windows should yield different decompositions unless the system is completely at rest, even when the dynamics of the system are steady. The first orthogonal component is simply the point in state space that the system is most correlated with across the observed time-frame.

## 11.4 Dynamic Mode Decomposition (DMD)

Dynamic Mode Decomposition (DMD) fits a reduced-rank, linear dynamical system to multi-dimensional time-series data. The rank-reduction is useful when the state space is very high-dimensional, for example when each time step is an image with many pixels. Canonically, fitting and rank reduction relies on SVD, so that the method is based on $L^2$ loss. It originated in the fluid dynamics community, where very high-dimensional state spaces are common because dynamics must be resolved across many length scales.

In case of plain-vanilla DMD, the dataset consists of pairs of snapshots of the system, which show the system at two subsequent time steps, i.e. $\{(\mathbf{x}_t, \mathbf{y}_t) : \mathbf{y}_t = \mathbf{x}_{t+\delta t}\}$. Often, the different pairs show the system on different state space trajectories. That is, they may originate with many different "incarnations" of the system. To me it looks basically like an VAR(1) model, except that it can handle a very high dimensional state vector by extracting the leading eigendecomposition of the coefficient matrix without having to calculate it. It also looks a lot like fitting the transition matrix of a Markov Chain.

Given a set of $m$ snapshots of the system $\{\mathbf{x}_t : t \in [1, m], \mathbf{x}_t \in \mathbb{R}^n\}$, let $\mathbf{X}_{1,m-1} = [\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_{t-1}, \mathbf{x}_{m-1}]$ be the matrix of state vectors from $t \in [1, m-1]$ and $\mathbf{X}_{2,m}$ be the matrix of state vectors advanced by one time step from $t \in [2, m]$. Then dynamic mode decomposition essentially looks for linear dynamics using linear regression:

$$\mathbf{X}_{2,m} = \mathbf{A}\mathbf{X}_{1,m-1} \tag{11.6}$$

Where $\mathbf{A} \in \mathbb{R}^{n \times n}$ is square and therefore diagonalizable, but may be very large.

Let $a_{i,j}$ be the entry in the $i$th row and $j$th column of $\mathbf{A}$. Elementwise, the equation for the $i$th state variable at time $t$, $x_{i,t} = [\mathbf{x}_t]_i$ is written:

$$x_{i,t} = \sum_j a_{i,j} x_{i,t-1} \tag{11.7}$$

The eigenvectors of $\mathbf{A}$ correspond to normal modes of the system. The corresponding eigenvectors, which may be real or complex, predict the evolution of the mode over time.

The size of $\mathbf{A} \in \mathbb{R}^{n \times n}$ may be so large that extracting its eigenvectors and eigenvalues may be computationally intractable. DMD instead derives a smaller matrix $\mathbf{A}'$ that has the same eigenvalues as $\mathbf{A}$ and uses those to also finds the eigenvectors.

$$\begin{aligned} &1. \quad \mathbf{X}_{1,m-1} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T, \quad \mathbf{X}_{2,m} = \mathbf{A}\mathbf{U}\mathbf{\Sigma}\mathbf{V}^T \\ &2. \quad \mathbf{U}^T\mathbf{X}_{2,m}\mathbf{V}\mathbf{\Sigma}^{-1} = \mathbf{U}^T\mathbf{A}\mathbf{U} = \mathbf{A}' \\ &3. \quad \mathbf{A}'\mathbf{W} = \mathbf{W}\mathbf{\Lambda} \\ &4. \quad \mathbf{\Phi} = \mathbf{X}^T\mathbf{V}\mathbf{\Sigma}^{-1}\mathbf{W} \end{aligned} \tag{11.8}$$

If the system has low dimensional structure, the matrix $\mathbf{A}'$ can be further reduced by only keeping the first $r$ vectors in $\mathbf{U}$. The final transformation$\mathbf{\Phi}$ gives high-dimensional eigenvectors of the full matrix $\mathbf{A}$.

Brunton (n/a) suggests that DMD is, in spirit, a combination of principal component analysis and the Fourier transform in time. The fact that DMD exists in a regression framework there are a huge number of extensions that can be leveraged. In particular, Brunton (n/a) delves into combining DMD with compressed sensing. What is surprising is that the dynamics that are modeled by DMD are not necessarily linear. The matrix $\mathbf{A}$ is large enough to approximate nonlinear dynamics.

## 11.5 Extended Dynamical Mode Decomposition (EDMD)

## 11.6 Sparse Identification of Nonlinear Dynamics (SINDy)

## 11.7 DMD with Irregularly Sampled Timesteps

# 12 Exploratory Data Analysis

**Contents of this chapter**

## 12.1 Basic Steps

### 12.1.1 Import, Clean, Data Census

### 12.1.2 Single Variable Explorations

### 12.1.3 Pair-wise Explorations

### 12.1.4 Multivariate Analysis

### 12.1.5 Estimation and Hypothesis Testing

### 12.1.6 Visualization

# 13 Causality

**Contents of this chapter**

## 13.1 Generalized Random Forests

Notes on **?**. Like most regression techniques, random forests are normally used to estimate the conditional mean of some data generating distribution, i.e. $\mu(x) = \mathbb{E}\left[Y|X = x\right]$. Athey, Tibshirani and Wager proposed generalized random forests, which estimate, more generally, some quantity $\theta(x)$. The main aim, and the reason why I am grouping this method under “causality", is the estimation of heterogenous treatment effects.

# 14 Machine Learning

**Contents of this chapter**

# 15 Algorithms

**Contents of this chapter**

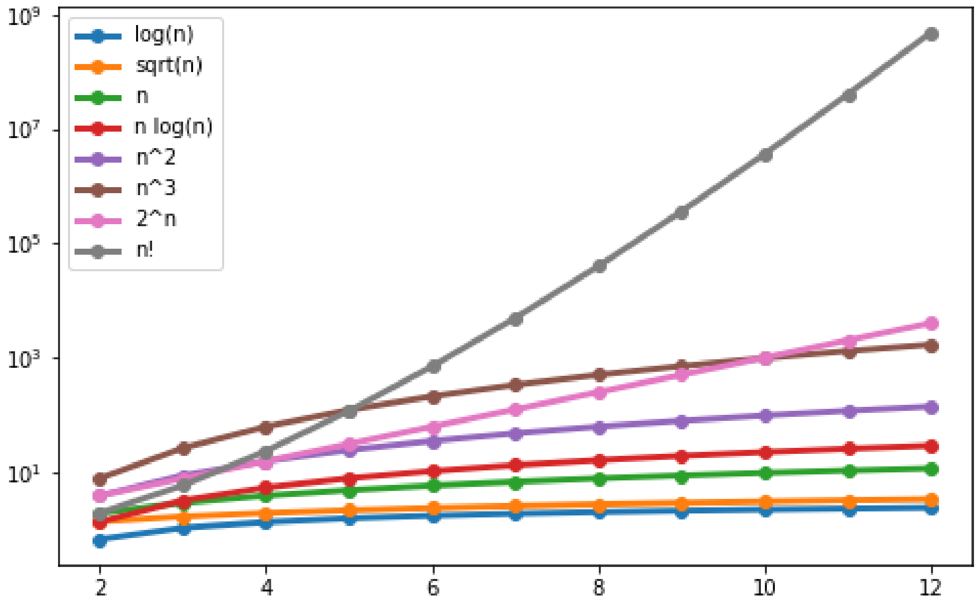


Figure 15.1: Common time complexities for algorithms.

# Bibliography

amoeba (2015), 'Relationship between svd and pca.'.
**URL:** *https://stats.stackexchange.com/questions/134282/relationship-between-svd-and-pca-how-to-use-svd-to-perform-pca*

Barnes, R. J. (n/a), 'Matrix differentiation (and some other stuff)'.
**URL:** *https://atmos.washington.edu/~dennis/MatrixCalculus.pdf*

Bell, J. (2015), 'The axiom of choice'.
**URL:** *https://plato.stanford.edu/entries/axiom-choice/*

Blitzstein, J. K. & Hwang, J. (2019), *Introduction to probability*, Crc Press.

Bogart, K. P. (2004), *Combinatorics through guided discovery*, Kenneth P. Bogart.
**URL:** *https://bogart.openmathbooks.org/*

Bradley, A. R. (n/a), 'Notes on sets and multisets'.
**URL:** *http://theory.stanford.edu/~arbrad/pivc/sets.pdf*

Brooks, J. & Jot, S. (n.d.), 'pcal1: An implementation in r of three methods for l1-norm principal component analysis'.

Brunton, S. (n/a), 'Steve brunton's youtube channel'.
**URL:** *https://www.youtube.com/channel/UCm5mt-A4w61lknZ9lCsZtBw*

Colyer, A. (2019), 'The why and how of nonnegative matrix factorization'.
**URL:** *https://blog.acolyer.org/2019/02/18/the-why-and-how-of-nonnegative-matrix-factorization/*

Cvitanovic, P., Artuso, R., Mainieri, R., Tanner, G. & Vattay, G. (2016), 'Chaos: classical and quantum, chaosbook. org', *Niels Bohr Institute, Copenhagen, Denmark* .

Damask, J. (2019), 'A consistently oriented basis for eigenanalysis', *arXiv preprint arXiv:1912.12983* .
**URL:** *https://arxiv.org/abs/1912.12983*

Friedman, J., Hastie, T. & Tibshirani, R. (2001), *The elements of statistical learning*, Vol. 1, Springer series in statistics New York.

Gavish, M. & Donoho, D. L. (2013), 'The optimal hard threshold for singular values is 4/sqrt(3)'.

Gera, R. (2009), 'Numerical linear algebra lecture notes, chapter 7'.
**URL:** *http://faculty.nps.edu/rgera/MA3042/2009/ch7.4.pdf*

Gillis, N. (2014), 'The why and how of nonnegative matrix factorization'.
**URL:** *https://arxiv.org/abs/1401.5226*

Hastie, T., Mazumder, R., Lee, J. D. & Zadeh, R. (2015), 'Matrix completion and low-rank svd via fast alternating least squares', *The Journal of Machine Learning Research* **16**(1), 3367–3402.

Kronenburg, M. (2011), 'The binomial coefficient for negative arguments', *arXiv preprint arXiv:1105.3689* .
**URL:** *https://arxiv.org/pdf/1105.3689.pdf*

Lindgren, G. (2006), 'Lectures on stationary stochastic processes', *PhD course of Lund's University* .

Mathworks (n/a), 'Eigenvalue and singular value decompositions'.
**URL:** *https://www.mathworks.com/content/dam/mathworks/mathworks-dot-com/moler/eigs.pdf*

Mazumder, R., Hastie, T. & Tibshirani, R. (2010), ‘Spectral regularization algorithms for learning large incomplete matrices’, *Journal of machine learning research* **11**(Aug), 2287–2322.

Megretski, A. (2004), ‘Proper orthogonal decompostion’.
**URL:** *http://web.mit.edu/6.242/www/images/lec6_6242_2004.pdf*

Salova, A., Emenheiser, J., Rupe, A., Crutchfield, J. P. & D’Souza, R. M. (2019), ‘Koopman operator and its approximations for systems with symmetries’, *Chaos: An Interdisciplinary Journal of Nonlinear Science* **29**(9), 093128.

Software, N. S. (n/a), ‘Ridge regression’.
**URL:** *https://ncss-wpengine.netdna-ssl.com/wp-content/themes/ncss/pdf/Procedures/NCSS/Ridge_Regression.pdf*

The Bright Side of Mathematics, n. (2019), ‘Lecture series on measure theory’.
**URL:** *https://www.youtube.com/watch?v=12kFDeN6xuI&list=PLBh2i93oe2qvMVqAzsX1Kuv6-4fjazZ8j&index=14*

Wasserman, L. (2013), *All of statistics: a concise course in statistical inference*, Springer Science & Business Media.

Wilf, H. S. (2013), *Generating functionology*, Elsevier.
**URL:** *https://www.math.upenn.edu/~wilf/gfology2.pdf*