Far from complete and far from correct. These notes exist because I learn better when I express things in my own words. In some cases I might be assembling material for teaching purposes. Might delete later.

balthasar@rhizomeworks.com

Contents

4

1 Combinatorics

Contents of this chapter

1.1 Useful Expressions

1.1.1 Binomial Coefficients and Binomial Expansions

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

$$\binom{n}{k} = \begin{cases} \frac{n!}{k!(n-k)!} & \text{for } n \geq k \\ \\ 0 & \text{otherwise} \end{cases} \tag{1.1}$$

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

$$\binom{-n}{k} = \begin{cases} (-1)^k \binom{n+k-1}{k} & \text{for } n \geq k \\ \\ 0 & \text{otherwise} \end{cases} \tag{1.2}$$

$$\binom{-n}{-k} = \begin{cases} (-1)^{k-n} \binom{k-1}{k-n} & \text{for } n \geq k \\ \\ 0 & \text{otherwise} \end{cases} \tag{1.3}$$

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.

$$(x + y)^n = \sum_{k=0}^{n} \binom{n}{k} x^k y^{n-k} \tag{1.4}$$

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

$$\frac{1}{(y + x)^n} = \sum_{k=0}^{\infty} \binom{-n}{k} x^k y^{-n-k} = (-1)^k \binom{n+k-1}{k} x^k y^{-(n+k)} \tag{1.5}$$

1.1.1.1 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:

$$f(x) = \sum_{k=0}^{\infty} \frac{f^{(k)}(x=0)}{k!} x^k \tag{1.6}$$

The derivatives of f are:

$$
\begin{aligned}
f^{(0)}(x) &= (y + x)^{-n} \\
f^{(1)}(x) &= (y + x)^{-(n+1)}(-1)n \\
f^{(2)}(x) &= (y + x)^{-(n+2)}(-1)^2 n(n + 1) \\
&\vdots \\
f^{(k)}(x) &= (y + x)^{-(n+k)}(-1)^k n(n + 1) \dots (n + k - 1) \\
&= (y + x)^{-(n+k)}(-1)^k \frac{(n+k-1)!}{(n-1)!}
\end{aligned}
\tag{1.7}
$$

Combining Eqns. 1.6 and 1.7 gives:

$$f(x) = \sum_{k=0}^{\infty} (-1)^k \binom{n+k-1}{k} x^k y^{-(n+k)} \tag{1.8}$$

1.1.1.2 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:

$$
\begin{aligned}
f^{(0)}(x) &= (y + x)^m \\
f^{(1)}(x) &= (y + x)^{m-1} m \\
f^{(2)}(x) &= (y + x)^{m-2} m(m - 1) \\
&\vdots \\
f^{(k)}(x) &= (y + x)^{n-k} m(m - 1) \dots (m - k + 1)
\end{aligned}
\tag{1.9}
$$

Combining Eqns. 1.6 and 1.9 gives:

$$f(x) = \sum_{k=0}^{\infty} \binom{m}{k} x^k y^{m-k} \tag{1.10}$$

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

$$\binom{m}{k} = \frac{m(m - 1)(m - 2) \dots (m - k + 2)(m - k + 1)}{k!} \tag{1.11}$$

1.1.2   Multinomial Expansion

$$(x_1 + x_2 + x_3 + \ldots + x_k)^n = \sum_{\substack{i_1, i_2, i_3, \ldots, i_k \\ \sum_j i_j = n}} \binom{n}{i_1, i_2, \ldots i_k} x^{i_1} x^{i_2} x^{i_3} \ldots x^{i_k} \tag{1.12}$$

with:

$$\binom{n}{i_1, i_2, \ldots i_k} = \frac{n!}{i_1! i_2!, i_3! \ldots i_k!} \tag{1.13}$$

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

1.1.3   Unnamed Polynomial Identity

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

$$\prod_i^n (1 - x_i) = \sum_{s=0}^n (-1)^s \sum_{0 \leq \underbrace{i_1, i_2, \ldots, i_s}_{\{i\}_s} \leq n} \prod_{i \in \{i\}_s} x_i \tag{1.14}$$

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 $\binom{n}{s}$.

1.1.4   Factorial Expansion

$$x^{\underline{n}} = \frac{x!}{(x-n)!} = \sum_{k=0}^n s(n, k) x^k \tag{1.15}$$

where

$$s(n, k) = (-1)^{n-k} \begin{bmatrix} n \\ k \end{bmatrix} \tag{1.16}$$

are the stirling numbers of the first kind.

1.1.5   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 1.2

They are given by an explicit formula:

$$S(k, n) = \frac{1}{n!} \sum_{i=1}^n (-1)^{n-j} \binom{n}{j} j^k \tag{1.17}$$

They can also be generated via the recurrence relation:

$$\left\{ \begin{matrix} k+1 \\ n \end{matrix} \right\} = n \left\{ \begin{matrix} k \\ n \end{matrix} \right\} + \left\{ \begin{matrix} k \\ n-1 \end{matrix} \right\} \tag{1.18}$$

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:

$$n \left\{ \begin{array}{c} k \\ n \end{array} \right\} = 1 \tag{1.19}$$

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:

$$\left\{ \begin{array}{c} k \\ n - 1 \end{array} \right\} = 1 \tag{1.20}$$

Furthermore, the following holds:

$$\left\{ \begin{array}{c} 0 \\ 0 \end{array} \right\} = 1 \tag{1.21}$$

$$\left\{ \begin{array}{c} k \\ 0 \end{array} \right\} = \left\{ \begin{array}{c} 0 \\ n \end{array} \right\} = 0 \tag{1.22}$$

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

1.2 Taxonomy of Distribution Problems: The Twentyfold Way

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 1.1: 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 no conditions	n^k functions	$\sum_{i=1}^k S(n,i)$ set partitions ($\le n$ parts)
2. Distinct Each gets at most one	$n^{\underline{k}}$ k-element permutations	1 if $k \le n$; 0 otherwise
3. Distinct Each gets at least one	$S(k,n)n!$ onto functions	$S(k,n)$ set partitions (n parts)
4. Distinct Each gets exactly one	$k! = n!$ permutations	1 if $k = n$; 0 otherwise
5. Distinct, order matters	$(k+n-1)^{\underline{k}}$ ordered functions	$\sum_{i=1}^n L(k,i)$ broken permutations ($\le n$ parts)
6. Distinct, order matters Each gets at least one	$(k)^{\underline{n}}(k-1)^{\underline{k-n}}$ ordered onto functions	$L(k,n) = \binom{k}{n}(k-1)^{\underline{k-n}}$ broken permutations (n parts)
7. Identical no conditions	$\binom{n+k-1}{k}$ multisets	$\sum_{i=1}^n P(k,i)$ number partitions ($\le n$ parts)
8. Identical Each gets at most one	$\binom{n}{k}$ subsets	1 if $k \le n$; 0 otherwise
9. Identical Each gets at least one	$\binom{k-1}{n-1}$ compositions (n parts)	$P(k,n)$ number partitions (n parts)
10. Identical Each gets exactly one	1 if $k = n$; 0 otherwise	1 if $k = n$; 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.

1.2.1 Distinct Objects

1.2.1.1 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 <u>with</u> replacement.

$$S = \{(a_1, a_2, ..., a_k) | a_i \in A, |A| = n\}$$

$$|S| = n^k$$

(1.23)

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.

1.2.1.2 Indistinct Recipients

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

$$|A| = \sum_{i=1}^{k} S(n, i) \tag{1.24}$$

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 1.1.5):

$$S(k, n) = \left\{ \begin{array}{c} k \\ n \end{array} \right\} = \frac{1}{n!} \sum_{j=0}^{n} (-1)^{n-j} \binom{n}{j} j^k \tag{1.25}$$

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.

1.2.2 Distinct Objects, Every Recipient Receives At Most One

1.2.2.1 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.

$$S = \{(a_1, a_2, ..., a_k) | a_i \in A, |A| = n, a_i \neq a_j\}$$

$$|S| = \frac{n!}{(n-k)!} = n^{\underline{k}} \text{ if } k \leq n, \ 0 \text{ otherwise.} \tag{1.26}$$

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.

1.2.2.2 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 <u>without</u> 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 $n < k$.

$$S = \{(a_1, a_2, ..., a_k)|a_i \in A, |A| = n, a_i = a_j\}$$

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

(1.27)

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.

1.2.3 Distinct Objects, Every Recipient Receives at Least One

1.2.3.1 Distinct Recipients

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

(1.28)

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.

1.2.3.2 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 1.1.5):

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

(1.29)

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.

14

1.2.4 Distinct Objects, Every Recipient Receives Exactly One

1.2.4.1 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 <u>without</u> replacement and with the requirement that all of the n are selected.

$$S = \{(a_1, a_2, ..., a_k) | a_i \in A, |A| = k = n, a_i \neq a_j\}$$

$$|S| = n! = k! \text{ if } k = n, \ 0 \text{ otherwise.}$$

(1.30)

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.

1.2.4.2 Indistinct Recipients

$$S = \{(a_1, a_2, ..., a_k) | a_i \in A, |A| = k, a_i = a_j\}$$

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

(1.31)

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?

1.2.5 Distinct Objects, Distributed in Ordered Groups

1.2.5.1 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 <u>without replacement</u>.

$$S = \{(\mathbf{a}_{\{i_1\}} = (a_{i_{1,1}}, a_{i_{1,1}}, ..., a_{i_{1,l}}), \mathbf{a}_{\{i_2\}}, ..., \mathbf{a}_{\{i_n\}}) | \sum_j^n |\mathbf{a}_{i_j}| = k, a_{i_{j,i}} \in A, |A| = n\}$$

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

(1.32)

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.

1.2.5.2   Indistinct Recipients

k objects are broken up into n

$$|S| = \sum_{i=1}^{n} L(k, i) \tag{1.33}$$

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:

$$L(k, n) = \binom{k}{n} (k-1)^{\underline{k-n}} = \frac{k!}{n!} \binom{k-1}{n-1} \tag{1.34}$$

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?

1.2.6   Distinct Objects, Distributed in Ordered Groups of At Least One

1.2.6.1   Distinct Recipients

$$|S| = \frac{k!}{(k-n)!} \frac{(k-1)!}{(n-1)!} = k^{\underline{n}} (k-1)^{\underline{k-n}} \tag{1.35}$$

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.

1.2.6.2   Indistinct Recipients

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

$$L(k,n) = \binom{k}{n}(k-1)^{\underline{k-n}} = \frac{k!}{n!}\binom{k-1}{n-1} \tag{1.36}$$

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?

1.2.7   Identical Objects

1.2.7.1   Distinct Recipients

$$|S| = \binom{k+n-1}{k} \tag{1.37}$$

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?

1.2.7.2   Indistinct Recipients

$$|S| = \sum_{i=1}^{n} P(k, i) \tag{1.38}$$

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?

1.2.8   Identical Objects, Each Receives At Most One

1.2.8.1   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.

$$|S| = \binom{n}{k} \tag{1.39}$$

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.

1.2.8.2   Indistinct Recipients

$$|S| = 1 \text{ if } k \leq n, \ 0 \text{ otherwise} \tag{1.40}$$

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.

1.2.9   Identical Objects, Each Receives At Least One

1.2.9.1   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.

$$|S| = \binom{k + n - 1 - k}{k - n} = \binom{n - 1}{k - 1} \tag{1.41}$$

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?

1.2.9.2   Indistinct Recipients

$$|S| = P(k, n) \tag{1.42}$$

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.

1.2.10   Identical Objects, Each Receives Exactly One

1.2.10.1   Distinct Recipients

$$|S| = 1 \text{ if } k = n, \ 0 \text{ otherwise} \tag{1.43}$$

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.

1.2.10.2   Indistinct Recipients

$$|S| = 1 \text{ if } k = n, \ 0 \text{ otherwise} \tag{1.44}$$

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.

1.3 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 $P(1)^1 P(2)$. 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:

$$
\begin{aligned}
E_P(s) \quad &= P(a) + P(a)^2 + P(a)P(b) + P(a)^2 P(b) + + P(a)P(b)^2 + P(a)^2 P(b)^2 \\
&\left(P(a) + P(a)^2 \right) \left(1 + P(b) + P(b)^2 \right)
\end{aligned}
\tag{1.45}
$$

1.3.1 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:

$$
E_P(s) = \prod_i^n (x^0 + x^1) = (1 + x)^n = \sum_i^n \binom{n}{i} x^i
\tag{1.46}
$$

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 $\binom{n}{i}$.

1.3.2 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 $P(banana) = x^{30}$.

$$
E_P(s) = \left(\sum_{i=0}^{5} x^{20} \right) \left(\sum_{i=0}^{4} x^{25} \right) \left(\sum_{i=0}^{3} x^{30} \right)
\tag{1.47}
$$

Which results in some power series of the form:

$$
E_P(s) = 1x^0 + 1x^{20} + 1x^{25} + 1x^{30} + 1x^{40} + ... + 2x^{60} + ... + 1x^{290}
\tag{1.48}
$$

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:

$$
\sum_{l=0}^{n+m+h} d_l x^l = \left(\sum_{i=0}^{n} a_i x^i \right) \left(\sum_{j=0}^{m} b_j x^j \right) \left(\sum_{k=0}^{h} c_k x^k \right)
\tag{1.49}
$$

$$
d_l = \sum_{\substack{i,\, j,\, k \\ i + j + k = l}} a_i b_j c_k
\tag{1.50}
$$

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

$$d_l = \sum_{\substack{i,j,k \\ i+j+k=l}} 1 = \binom{l+3-1}{l} \tag{1.51}$$

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:

$$d_l = \sum_{i=0}^{l} \sum_{j=0}^{l-i} a_i b_j c_{l-i-j} \tag{1.52}$$

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.

1.3.3 Example: Dice

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

$$E_P = \left(\sum_{i=0}^{\infty} a_i x^i \right)^n = \sum_{i=0}^{\infty} d_i x^i \tag{1.53}$$

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

$$E_P = \left(\sum_{i=1}^{k} x^i \right)^n = \left(x(1-x^k) \sum_{i=0}^{\infty} x^i \right)^n = x^n \left(\frac{1-x^k}{1-x} \right)^n = x^n \left(\sum_{i=0}^{n} (-1)^i \binom{n}{i} x^{ik} \right) \left(\sum_{j=0}^{\infty} (-1)^j \binom{-n}{j} x^j \right) \tag{1.54}$$

The coefficient for x^s is given the sum:

$$d_s = \sum_{ki+j=s-n} (-1)^{i+j} \binom{n}{i} \binom{-n}{j} i \in [0,n] j \in [0,\infty] \tag{1.55}$$

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

$$6i + j = 7 - 2 = 5s \tag{1.56}$$

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

$$d_s = (-1)^5 \binom{2}{0} \binom{-2}{5} = (-1)^{10} 1 \binom{2+5-1}{5} = \binom{6}{5} = \frac{6!}{5!1!} = 6 \tag{1.57}$$

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

$$[6,1], [5,2], [4,3], [3,4], [2,5], [1,6] \tag{1.58}$$
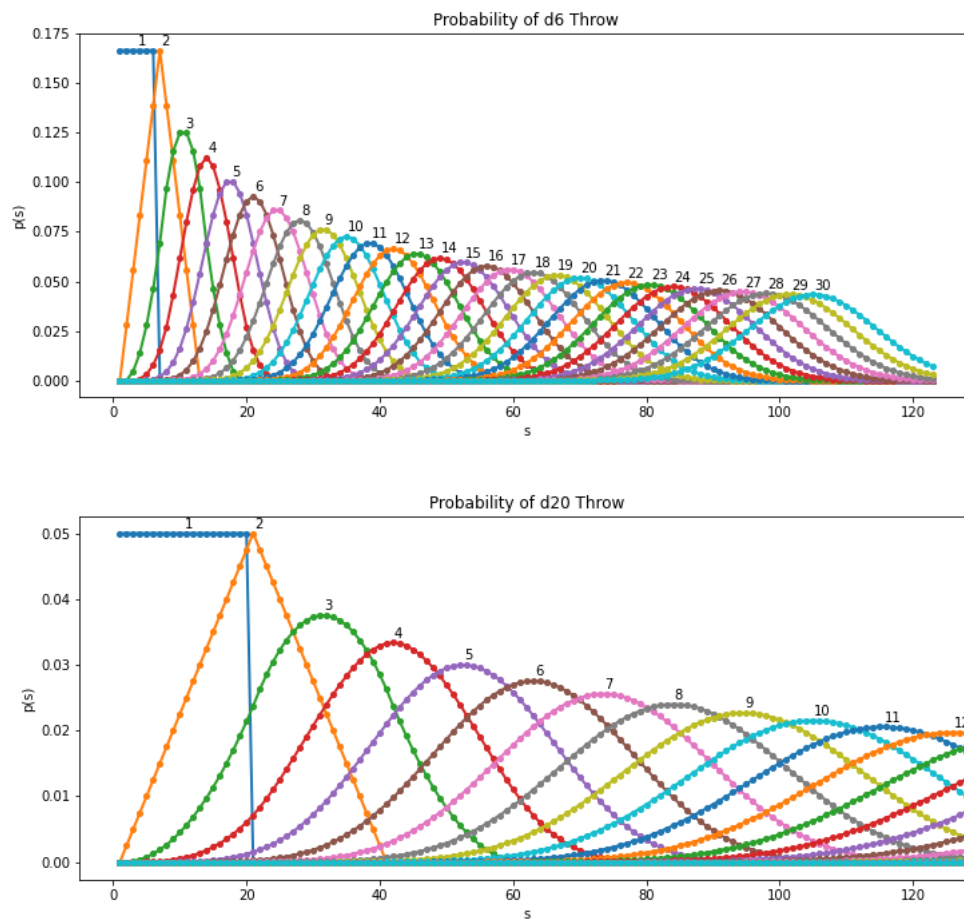
Figure 1.2: Probability of the sum of 6-sided and 20-sided dice, calculated with Eqn. 1.57. Note that these are simply the convolutions of $n = 1, 2, 3, ...$ square waves, which rapidly adopts the shape of a Bell curve.

2 Probability

Contents of this chapter

2.1 Measure Theory

2.2 Change of Variable

2.3 Copulas

2.4 Relationship Between Distributions

3 Stochastic Processes and Time Series Analysis

Contents of this chapter

3.1   Markov Chains

3.2   Martingales

3.3   Hidden Markov Models

4 Linear Algebra and Multivariable Calculus

Contents of this chapter

4.1 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 $\mathbf{x}^\alpha = \prod_i x_i^{\alpha_i}$.

4.1.1 Example: Multinomial Coefficients

Instead of:

$$\sum_{0 \leq i_1, i_2, i_3, ..., i_k \leq n} \binom{n}{i_1, i_2, i_3, ..., i_k} \tag{4.1}$$

Write:

$$\sum_{0 \le |\alpha| \le n} \binom{n}{\alpha} \tag{4.2}$$

4.1.2 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 $\mathbf{a}$:

$$f(\mathbf{x}) = \sum_{|\alpha| \ge 0} \frac{(\mathbf{x} - \mathbf{a})^\alpha}{\alpha!} (\partial^\alpha) f \tag{4.3}$$

4.2 $\mathbf{A} = \mathbf{LL}^\dagger$ Cholesky Decomposition

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

$$\mathbf{A} = \mathbf{LL}^\dagger \tag{4.4}$$

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.

4.2.1 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 $\mathbf{\Sigma} = \mathbf{LL}^\dagger$, then the vector $\mathbf{v}_t = \mathbf{Lu}_t$ has the desired covariance.

4.3 Generalized Eigenvectors

4.4 $\mathbf{A} = \mathbf{V \Lambda V}^{-1}$ Spectral Theorems, Diagonalization

Spectral theorems deal with diagonalizable linear operators.

A diagonalization of a matrix 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

$$\mathbf{A} = \mathbf{V \Lambda V}^{-1} \tag{4.5}$$

4.4.1 $\mathbf{A} = \mathbf{V \Lambda V}^T$ Eigendecomposition of Symmetric Matrices

A symmetric matrix $\mathbf{A}$ has orthogonal eigenvectors so that $\mathbf{V}^{-1} = \mathbf{V}^T$ and real eigenvalues. In that case, the diagonalization is:

$$\mathbf{A} = \mathbf{V \Lambda V}^T = \begin{bmatrix} | & | & \cdots & | \\ v_1 & v_2 & \ddots & v_n \\ | & | & \cdots & | \end{bmatrix} \begin{bmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & \cdots & \lambda_n \end{bmatrix} \begin{bmatrix} \rule{1cm}{0.4pt} & v_1 & \rule{1cm}{0.4pt} \\ \rule{1cm}{0.4pt} & v_2 & \rule{1cm}{0.4pt} \\ \vdots & \ddots & \vdots \\ \rule{1cm}{0.4pt} & v_n & \rule{1cm}{0.4pt} \end{bmatrix} \tag{4.6}$$
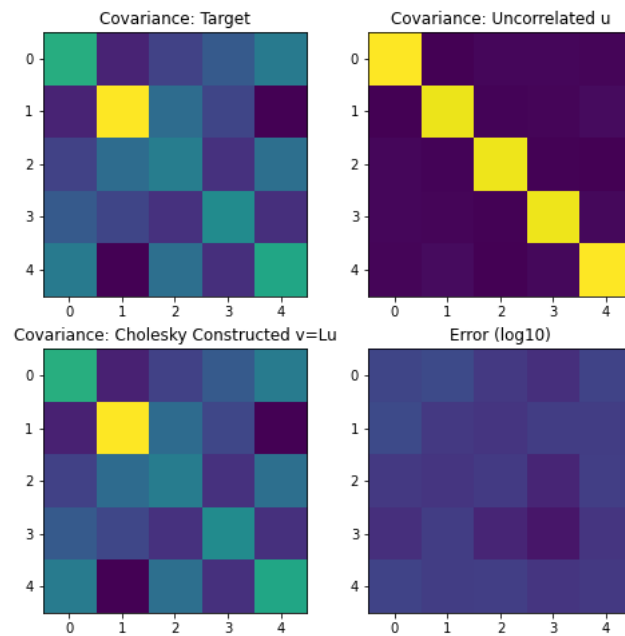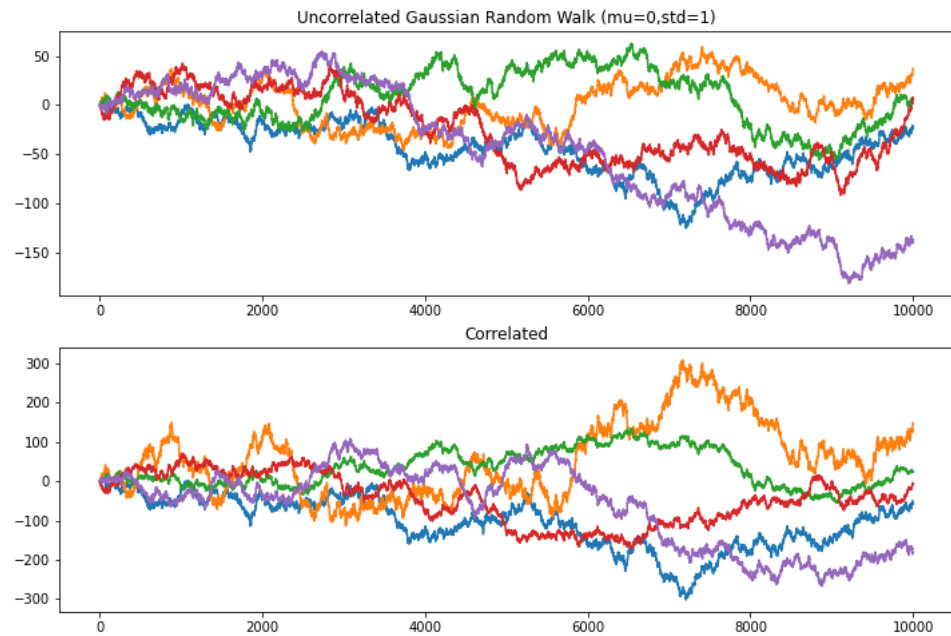
Figure 4.1: Creating correlated random variables from uncorrelated random variables using the Cholesky decomposition of the covariance matrix. The 5 uncorrelated random variables are sampled from a standard normal distribution. It is difficult to see a difference between the correlated and uncorrelated random walks.

This means that enables the expression of $\mathbf{A}$ in terms of projections on the eigenvectors:

$$\mathbf{A} = \sum_i \lambda_i (v_i \otimes v_i) \tag{4.7}$$

Where $\otimes$ is the outer product. Since the eigenvectors are an orthonormal basis, $\sum_i v_i \otimes v_i = \mathbb{I}$.

4.4.2 $\mathbf{H} = \mathbf{U\Lambda U}^T$ 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 $\mathbf{H} = \mathbf{U\Lambda U}^T$.

4.4.3 Eigenvalue Sensitivity and Accuracy

4.4.3.1 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 (**?**). Let $|| \cdot ||$ denote a submultiplicative matrix norm, then:

$$\begin{aligned} \Lambda + \delta\Lambda &= \mathbf{X}^{-1}\left(\mathbf{A} + \delta\mathbf{A}\right)\mathbf{X} \\ \delta\Lambda &= \mathbf{X}^{-1}\delta\mathbf{A}\mathbf{X} \\ ||\delta\Lambda|| &= ||\mathbf{X}^{-1}\delta\mathbf{A}\mathbf{X}|| \leq ||\mathbf{X}^{-1}||||\mathbf{X}||||\delta\mathbf{A}|| \end{aligned} \tag{4.8}$$

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 σ_1 and σ_2 are the square roots of the largest and the smallest eigenvalue of $\mathbf{X}^\dagger\mathbf{X}$ respectively (cf. section 4.10 on matrix norms). In that case, the sensitivity of the eigenvalues to a change in $\mathbf{A}$ is:

$$||\delta\mathbf{\Lambda}||_{(2)} \leq \frac{\sigma_1}{\sigma_n}||\delta\mathbf{A}||_{(2)} = \kappa(\mathbf{X})||\delta\mathbf{A}||_{(2)} \tag{4.9}$$

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.

4.4.3.2 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 $||\lambda_i||_{(2)} \leq \kappa(\lambda_i, \mathbf{H})||\mathbf{H}||_{(2)} = 1 \times ||\mathbf{H}||_{(2)}$.

4.5 $\mathbf{A} = \mathbf{U\Sigma V}^\dagger$ Singular Value Decomposition

The Singular Value Decomposition (SVD) exists for any matrix $\mathbf{A} \in \mathbb{C}^{m \times n}$, and is a closely related alternative to the eigendecomposition (cf. section 4.4) that works for non-square matrices. The decomposition comes up incessantly in the context of data analysis. In general, the decomposition has the form:

$$\mathbf{A} = \mathbf{U\Sigma V}^\dagger \tag{4.10}$$

Where $\mathbf{U}^\dagger\mathbf{U} = \mathbf{I}$, $\mathbf{V}$ is unitary, and Σ is a diagonal matrix with real and positive entries σ_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 $\mathbf{AA}^\dagger$. $\mathbf{V}$ 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, $\mathbf{AA}^\dagger$. If the $\mathbf{A}$ is square and symmetric ($\mathbf{A} = \mathbf{A}^\mathbf{T}$, cf. section 4.7.5), then the singular values are simply the absolute values of the eigenvalues of $\mathbf{A}$.

The singular values are always real and positive or zero, because $\mathbf{A}^\dagger\mathbf{A}$ and $\mathbf{AA}^\dagger$ are hermitian and positive semi-definite. When the matrix $\mathbf{A}$has purely real entries, then the left and right singular vectors are also
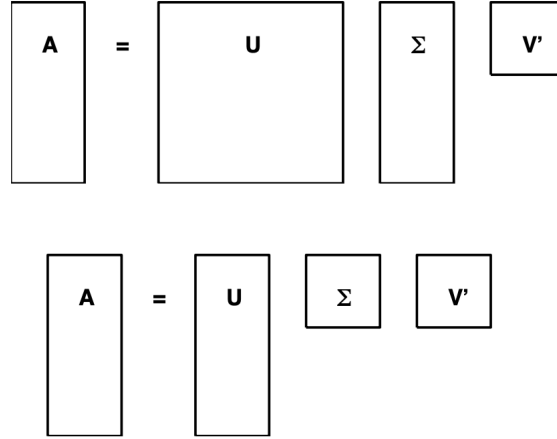
Figure 4.2: Dimensions for Full vs. Economy SVDs

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 $\mathbf{u}_i, \mathbf{v}_i \to -\mathbf{u}_i, -\mathbf{v}_i$ if $\mathbf{u}_i$ and $\mathbf{v}_i$ are the ith 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 7.2 on performing SVD specifically on data matrices.

4.5.1 Full and Economy SVDs

While these properties are always true, unfortunately people use a range of conventions when it comes to the size of $\mathbf{U}$, $\boldsymbol{\Sigma}$ 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 Σ 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 Σ is square, i.e. $\mathbf{U} : m \times n$, $\boldsymbol{\Sigma} : n \times n$ 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 Σ be a square matrix significantly simplifies calculations. Friedman et al.'s (2001) is known as the *economy SVD*.

The two different layouts are illustrated in Figure 4.2, which I brazenly copied from Mathworks (n/a).

4.5.2 Matrix Approximation

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

$$\mathbf{A} = \sum_{i=1}^{n} \sigma_i(\mathbf{u}_i \otimes \mathbf{v}_i) \tag{4.11}$$

Where $\mathbf{u}_i$ is the ith left singular vector and $\mathbf{v}_i$ is the ith right singular vector. 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.

$\arg\min_{\mathrm{rank}(\mathbf{A}^{(r)})=r\leq n} ||\mathbf{A^r} - \mathbf{A}||_F$. (The Frobenius norm essentially measures the elementwise least-squares error, cf. section 4.10.3). 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 (the hope is that these correspond to noise). For a more elaborate discussion of SVD in data analysis, see section 7.2

4.6 Types of Transformations

4.6.1 Similarity Transformations

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

$$\mathbf{A} = \mathbf{TBT^{-1}} \tag{4.12}$$

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

4.6.2 Affine Transformations

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

$$f : V \to W \tag{4.13}$$

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:

$$\mathbf{Ax} + \mathbf{b} = \underbrace{\left[\ \mathbf{A} \mid \mathbf{b}\ \right]}_{\mathbf{M}} \underbrace{\begin{bmatrix} \mathbf{x} \\ 1 \end{bmatrix}}_{\mathbf{c}} \tag{4.14}$$

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

4.6.3 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 **??**). Unitary transformations include translations, reflections and rotations.

4.6.4 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.

$$f : V_1 \times V_2 \times ... \times V_n \to W \tag{4.15}$$

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.

4.6.5 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.

$$f : V_1 \times V_2 \times ... \times V_n \to K \tag{4.16}$$

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

4.7 Types of Matrices

4.7.1 $\mathrm{sgn}\left(\mathbf{x}^\dagger \mathbf{H} \mathbf{x}\right)$ 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 4.11.5). 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.

4.7.2 Triangular

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

$$\mathbf{L} = \begin{bmatrix} l_{1,1} & & & & & 0 \\ l_{2,1} & l_{2,2} & & & & \\ l_{3,1} & l_{3,2} & \ddots & & & \\ \vdots & \vdots & \ddots & \ddots & & \\ \vdots & \vdots & & \ddots & \ddots & \\ l_{n,1} & l_{n,2} & \dots & \dots & l_{n,n-1} & l_{n,n} \end{bmatrix} \tag{4.17}$$

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

4.7.3 $\mathbf{AB} - \mathbf{BA} = 0$ 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:

$$\mathbf{A} = \mathbf{V}\mathbf{\Lambda_A}\mathbf{V}^{-1} \mathbf{B} = \mathbf{V}\mathbf{\Lambda_B}\mathbf{V}^{-1} \tag{4.18}$$

They commute, because:

$$\mathbf{AB} = \mathbf{V}\mathbf{\Lambda_A}\mathbf{V}^{-1}\mathbf{V}\mathbf{\Lambda_B}\mathbf{V}^{-1} = \mathbf{V}\mathbf{\Lambda_B}\mathbf{\Lambda_A}\mathbf{V}^{-1} = \mathbf{V}\mathbf{\Lambda_B}\mathbf{V}^{-1}\mathbf{V}\mathbf{\Lambda_A}\mathbf{V}^{-1} = \mathbf{BA} \tag{4.19}$$

4.7.4 $\mathbf{AB} + \mathbf{BA} = 0$ Anticommuting

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

4.7.5 $\mathbf{A}^\dagger = \mathbf{A}$ Hermitian, Symmetric

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

$$\mathbf{A}^{*T} = \mathbf{A}^\dagger = \mathbf{A} \tag{4.20}$$

4.7.5.1 Properties

There are many properties of Hermitian matrices.

- 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 $\mathbf{A}^T = \mathbf{A}$.

4.7.6 $\mathbf{A}^\dagger = -\mathbf{A}$ Skew Hermitian, Skew Symmetric

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

$$\mathbf{A}^{*T} = \mathbf{A}^\dagger = -\mathbf{A} \tag{4.21}$$

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

$$\mathbf{A}^T = -\mathbf{A} \tag{4.22}$$

4.7.7 $\mathbf{U}^\dagger = \mathbf{U}^{-1}$ 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 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 $SU(n)$.

4.7.8 $\mathbf{A} = \mathbf{TBT}^{-1}$ Similarity

Two matrices are said to be similar if they can be related through a similarity transformation $\mathbf{A} = \mathbf{TBT}^{-1}$ where $\mathbf{T}$ is some nonsingular matrix (cf. section 4.6.1). An important example is that square matrices are similar to diagonal matrices, see section 4.4.

4.7.9 $\mathbf{AA} = \mathbf{I}$ Involutory

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

$$\mathbf{AA} = \mathbf{I} \tag{4.23}$$

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

4.8 Properties of Norms

4.9 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:

$$||\mathbf{x}||^p = \left[\sum_i^n |x_i|^p \right]^{1/p} \tag{4.24}$$

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.

4.9.1 L^1 Taxicab / Manhattan Norm

$$||\mathbf{x}||^1 = \sum_i |x_i| \tag{4.25}$$

4.9.2 L^2 Euclidian Norm

$$||\mathbf{x}||^2 = \sum_i |x_i|^2 \tag{4.26}$$

4.9.3 L^∞ Maximum Norm

$$||\mathbf{x}||^\infty = \max(x_1, x_2, ..., x_n) \tag{4.27}$$

4.9.4 $L^{-\infty}$ Minimum Norm

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

$$||\mathbf{x}||^{-\infty} = \min(x_1, x_2, ..., x_n) \tag{4.28}$$

4.10 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.

4.10.1 $||\mathbf{A}||_{(\alpha)}$ 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 $|| \cdot ||_{\alpha}$:

$$||\mathbf{A}||_{(\alpha)} = \sup \left\{ \frac{||\mathbf{Av}||_{\alpha}}{||\mathbf{x}||_{\alpha}} : \mathbf{v} \in V \right\} \tag{4.29}$$

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 4.10.6).

Note that $||A||_{(q)}$ 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.

4.10.2 $||\mathbf{A}||_q$ 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:

$$||\mathbf{A}||_q = \left(\sum_i \sum_j a_{i,j}^q \right)^{1/q} \tag{4.30}$$

For $q = 2$, this becomes the Frobenius norm (section 4.10.3). For vectors $\mathbf{v} \in \mathbb{R}^n$, the q-norm is more known as p-norm or L^p norm (cf. section 4.9).

4.10.3 $||\mathbf{A}||_F$ 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:

$$||\mathbf{A}||_F = \sqrt{\sum_i \sum_j a_{i,j}^2} \tag{4.31}$$

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

4.10.4 $||\mathbf{A}||_{(1)}$ (1)-Norm

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

$$||\mathbf{A}||_{(1)} = \max_{1 \leq j \leq n} \sum_{i=1}^{m} |a_{i,j}| \tag{4.32}$$

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

4.10.5 $||\mathbf{A}||_{(\infty)}$ (∞)-Norm

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

$$||\mathbf{A}||_{(\infty)} = \max_{1 \le i \le m} \sum_{j=1}^{n} |a_{i,j}| \tag{4.33}$$

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

4.10.6 $||\mathbf{A}||_{(2)}$ (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:

$$||A||_{(2)} = \max_{\mathbf{x} \neq 0} \frac{||\mathbf{A}\mathbf{x}||_2}{||\mathbf{x}||_2} \tag{4.34}$$

Which is the square root of the largest eigenvalue of $A^T A$. Or, equivalently, $||A||_{(2)} = \sigma_1$, where σ_1 is the largest singular value of the SVD of $\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T$. That means that $||A^{-1}|| = \frac{1}{\sigma_n}$, where σ_n is the smallest singular value of the SVD of $\mathbf{A}$.

4.11 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).

4.11.1   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:

$$\frac{\mathrm{d}\mathbf{f}(\mathbf{x})}{\mathrm{d}\mathbf{x}} = \left[\frac{\partial \mathbf{f}}{\partial x_1}, \dots, \frac{\partial \mathbf{f}}{\partial x_n} \right] = \begin{bmatrix} \frac{\partial f_1}{\partial x_1} & \cdots & \frac{\partial f_1}{\partial x_n} \\ \vdots & \vdots & \vdots \\ \frac{\partial f_m}{\partial x_1} & \cdots & \frac{\partial f_m}{\partial x_n} \end{bmatrix} \tag{4.35}$$

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.

$$\nabla_{\mathbf{x}} \left(\mathbf{u}^T \mathbf{x} \right) = \left[\frac{\partial}{\partial x_1} \left(\sum_i u_i x_i \right), \dots, \frac{\partial}{\partial x_n} \left(\sum_i u_i x_i \right) \right] = \mathbf{u}^T$$

$$\nabla_{\mathbf{x}} \left(\mathbf{x}^T \mathbf{u} \right) = \left[\frac{\partial}{\partial x_1} \left(\sum_i u_i x_i \right), \dots, \frac{\partial}{\partial x_n} \left(\sum_i u_i x_i \right) \right] = \mathbf{u}^T$$

$$\nabla_{\mathbf{x}} \left(\mathbf{x}^T \mathbf{x} \right) = \left[\frac{\partial}{\partial x_1} \left(\sum_i x_i^2 \right), \dots, \frac{\partial}{\partial x_n} \left(\sum_i x_i^2 \right) \right] = 2\mathbf{x}^T$$

$$\nabla_{\mathbf{x}} \left(\mathbf{A}\mathbf{x} \right) = \begin{bmatrix} \underbrace{\frac{\partial}{\partial x_1} \left(\sum_i A_{1i} x_i \right)}_{A_{11}} & \cdots & \underbrace{\frac{\partial}{\partial x_n} \left(\sum_i A_{1i} x_i \right)}_{A_{1n}} \\ \vdots & \vdots & \vdots \\ \underbrace{\frac{\partial}{\partial x_1} \left(\sum_i A_{ni} x_i \right)}_{A_{n1}} & \cdots & \underbrace{\frac{\partial}{\partial x_n} \left(\sum_i A_{ni} x_i \right)}_{A_{nn}} \end{bmatrix} = \mathbf{A} \tag{4.36}$$

4.11.2 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:

$$\frac{d\left[f^{-1}(b)\right]}{dx} = \frac{1}{\frac{df(a)}{dx}} \tag{4.37}$$

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 $\mathbf{f} : \mathbf{x} \to \mathbf{y}$:

$$\nabla_{\mathbf{y}}\left[\mathbf{f}^{-1}\right] = [\nabla_{\mathbf{x}}\mathbf{f}]^{-1} \tag{4.38}$$

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 $\mathbf{a}$.

4.11.3 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.

4.11.4 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:

$$\mathrm{d}^n x = |\nabla_{\mathbf{y}}\mathbf{H}|\,\mathrm{d}^n y \tag{4.39}$$

And:

$$\int_{\mathbf{x}} \mathrm{d}^n\mathbf{x}\, f(\mathbf{x}) = \int_{\mathbf{y}} \mathrm{d}^n\mathbf{y}\, |\nabla_{\mathbf{y}}\mathbf{H}|\, f(\mathbf{H}(\mathbf{y})) \tag{4.40}$$

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

$$
\begin{aligned}
\mathrm{d}^n y &= \left|\nabla_{\mathbf{x}}\mathbf{H}^{-1}(\mathbf{x})\right|\mathrm{d}^n x \\
&= \left|[\nabla_{\mathbf{y}}\mathbf{H}(\mathbf{y})]^{-1}\right|\mathrm{d}^n x \\
\mathrm{d}^n x &= \frac{1}{\left|[\nabla_{\mathbf{y}}\mathbf{H}(\mathbf{y})]^{-1}\right|}\mathrm{d}^n y
\end{aligned}
\tag{4.41}$$

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).

4.11.5 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 $\mathbf{H}(f)_{i,j} = \frac{\partial^2 f}{\partial x_i \partial x_j}$.

$$\mathbf{H}(f) = \begin{bmatrix} \frac{\partial^2 f}{\partial x_1^2} & \frac{\partial^2 f}{\partial x_1 \partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_1 \partial x_n} \\ \frac{\partial^2 f}{\partial x_2 \partial x_1} & \frac{\partial^2 f}{\partial x_2^2} & \cdots & \frac{\partial^2 f}{\partial x_1 \partial x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 f}{\partial x_n \partial x_1} & \frac{\partial^2 f}{\partial x_n \partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_n^2} \end{bmatrix} \tag{4.42}$$

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

4.11.5.1 Testing Convexity

The definiteness (cf. section 4.7.1) 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.

5 Information Theory

Contents of this chapter

5.1 Entropy

5.2 Mutual Information

5.3 Kullback-Leibler Divergence

6 Linear Regression

Contents of this chapter

6.1 Ordinary Linear Regression

6.2 R^2 Value

6.3 Regression Diagnostics

6.4 t-Statistics

6.5 AIC and BIC

6.6 Factor Regression

6.7 Factor Models

7 Unsupervised Learning

Contents of this chapter

7.1 Independent Component Analysis

Independent Component Analysis (ICA) is a decomposition of data into components that have vanishing mutual information.

7.2 Eigenanalysis, Singular Value Decomposition of Data Matrices

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 4.5). 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 the same as for PCA (cf. section **??**) except that it is arrived at slightly differently, because PCA involves diagonalizing the covariance matrix. The decomposition of the dataset can be understood as follows. In the SVD of the data matrix:

$$\mathbf{X} = \mathbf{U}\Sigma\mathbf{V}^T \tag{7.1}$$

The individual rows of $\mathbf{X}$, i.e. the individual data points, are expressed as:

$$\mathbf{x}_i^T = \sum_{j=1}^{p} u_{i,j} \sigma_j \mathbf{v}_j^T \tag{7.2}$$

Where $\mathbf{x}_i$ is the ith 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 σ_i, which give a measure of the global strength of the corresponding basis vector.

SVD is scalable to very large datasets and finds many applications in the wild, including page rank, facial recognition, recommendation algorithms etc.
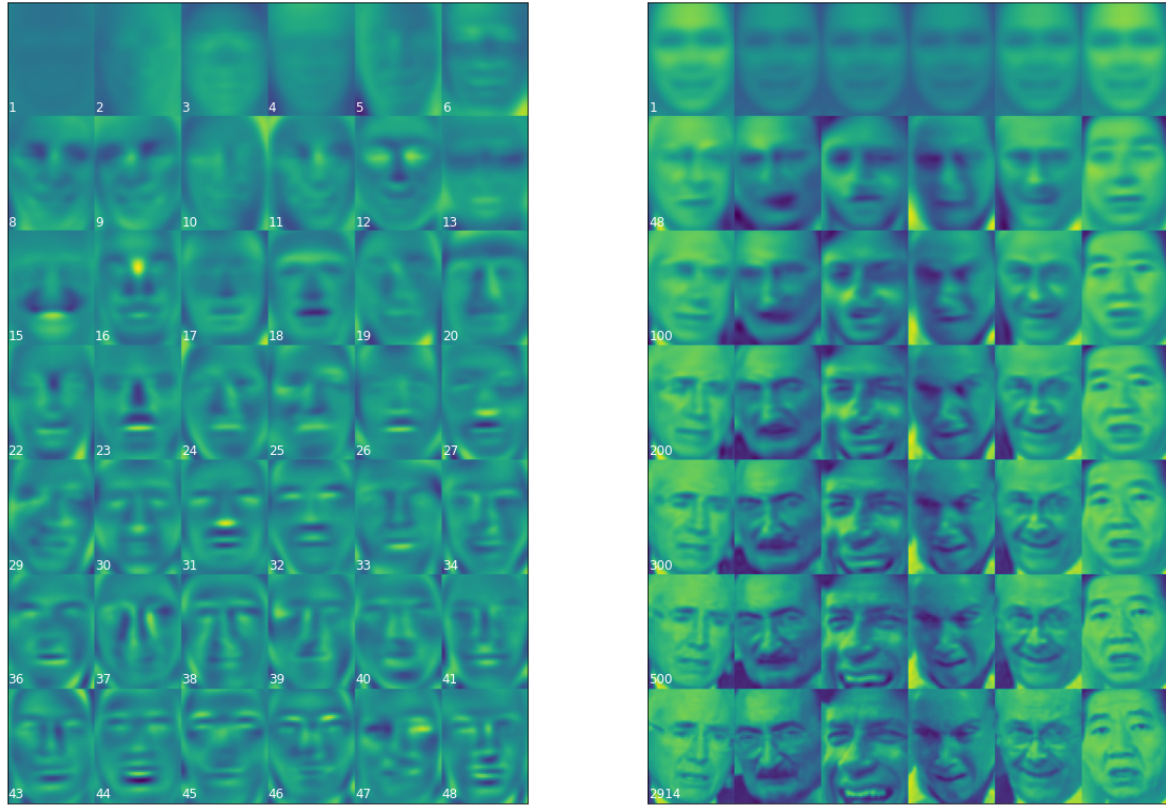
Figure 7.1: Left: The first 48 eigenfaces. The colorscale is consistent across the images. As can be expected, the eigenfaces seem to have an ordering from more general features, that are highly prevalent in the dataset, towards more specific features. Right: Five sample portraits from the dataset approximated using different numbers of eigenvectors. 2914 corresponds to the original image, which had 2914 pixels (degrees of freedom).

7.2.1 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 7.1.

Facial recognition may be performed by projecting a new face into the space of eigenfaces and some form of matching to the coefficients of a known subject. This could be simply the euclidean distance, or it could be a classifier. In case of a classifier, 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.

7.2.2 Constructing A Consistently Oriented Eigenbasis

As discussed in section 4.5, 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 repeatedly on a slowly evolving system, the derived eigenbasis of the system can abruptly change, from, for example, a left-handed to a right-handed coordinate system. Hence there is a need to find a way to consistently orient the basis of eigenvectors.

Heuristic techniques can work by taking the dot product between the singular vectors derived for two

separate SVDs, and flipping them whenever the dot product is negative. A more rigorous technique was derived by Damask (2019). Damask's (2019) technique relies on constructing the sequence of reflection and rotation operations necessary to transform a constituent basis into the derived eigenbasis. The gradual transformation begins with the singular vector with the largest singular value, because it has the greatest numerical certainty.

Once consistently oriented, directional statistics (cf. section 8) can be applied to analyze the evolution of the eigensystem over different measurements.

In general, singular vectors are only labeled by their corresponding singular value (assuming no degeneracy). Rank order changes require attaching an additional label to the eigenvectors.

7.3 Principal Component Analysis (PCA)

Principal Component Analysis (PCA) is a decomposition of data into linearly uncorrelated components. It turns out that these axes are the right-singular eigenvectors $\mathbf{V}$ that are found from SVD. SVD and PCA have close correspondence (cf. section 7.3.1).

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 4.5). Just like SVD, the lower rank approximations $\mathbf{X}$ are the best possible approximations with respect to the Frobenius Norm (cf. section 4.10.3).

The $p \times p$ covariance matrix $\mathbf{C}$ of $\mathbf{X}$ is:

$$\mathbf{C} = \frac{(\mathbf{X} - \langle \mathbf{X} \rangle)^T (\mathbf{X} - \langle \mathbf{X} \rangle)}{n - 1} \tag{7.3}$$

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 λ_i:

$$\mathbf{C} = \mathbf{V} \mathbf{\Lambda} \mathbf{V}^T \tag{7.4}$$

Where the eigenvectors in V are ordered so that the eigenvectors along the diagonal of $\mathbf{\Lambda}$ have decreasing magnitude. The eigenvectors are the principal axes or principal directions of the data.

7.3.1   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:

$$(\mathbf{X} - \langle \mathbf{X} \rangle) = \mathbf{U} \mathbf{\Sigma} \mathbf{V}^T \tag{7.5}$$

Then:

$$\mathbf{C} = \frac{\mathbf{V} \mathbf{\Sigma} \mathbf{U}^T \mathbf{U} \mathbf{\Sigma} \mathbf{V}^T}{n - 1} = \mathbf{V} \frac{\mathbf{\Sigma}^2}{n - 1} \mathbf{V}^T \tag{7.6}$$

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 $\lambda_i = \frac{\sigma_i^2}{n-1}$.

7.3.2 Bayesian PCA

7.4 $\mathrm{X} = \mathrm{WH}$ Non-Negative Matrix Factorization

Non-Negative Matrix Factorization works for positive matrices and is interesting as a method for dimensionality reduction. It works by expressing a 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 $\mathbf{H} \in \mathbb{R}_+^{r \times n}$. An excellent introduction is 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.

8 Statistics

Contents of this chapter

8.1 Directional Statistics

8.1.1 Mean Direction

The mean direction of a collection of i vectors $\{\mathbf{x}\}_i$ is (**?**):

$$\langle \mathbf{x} \rangle = \frac{\mathbf{x_s}}{||\mathbf{x_s}||_2} \tag{8.1}$$

Where

$$\mathbf{x_s} = \sum_i \mathbf{x_i} \tag{8.2}$$

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 (**?**)).

8.1.2 Dispersion

Measuring the variance of a vector is more difficult.

8.1.3 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, both common and differential modes of estimation exist. One way of meausuring directional dispersion is to look at the *mean resultant length*:

$$\mu_r = \frac{||\mathbf{x_s}||}{N} \quad 0 \leq \mu_r \leq 1 \tag{8.3}$$

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, **?** 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.

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*

Bogart, K. P. (2004), *Combinatorics through guided discovery*, Kenneth P. Bogart.
 URL: *https://bogart.openmathbooks.org/*

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/*

Damask, J. (2019), 'A consistently oriented basis for eigenanalysis', *arXiv preprint arXiv:1912.12983* .

Friedman, J., Hastie, T. & Tibshirani, R. (2001), *The elements of statistical learning*, Vol. 1, Springer series in statistics New York.

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', *Regularization, optimization, kernels, and support vector machines* **12**(257), 257–291.

Kronenburg, M. (2011), 'The binomial coefficient for negative arguments', *arXiv preprint arXiv:1105.3689* .
 URL: *https://arxiv.org/pdf/1105.3689.pdf*

Mathworks (n/a), 'Eigenvalue and singular value decompositions'.
 URL: *https://www.mathworks.com/content/dam/mathworks/mathworks-dot-com/moler/eigs.pdf*